

Learning-based image denoising methods have been bounded to situations where well-aligned noisy and clean images are given, or samples are synthesized from predetermined noise models, e.g., Gaussian. While recent generative noise modeling methods aim to simulate the unknown distribution of real-world noise, several limitations still exist. In a practical scenario, a noise generator should learn to simulate the general and complex noise distribution without using paired noisy and clean images. However, since existing methods are constructed on the unrealistic assumption of real-world noise, they tend to generate implausible patterns and cannot express complicated noise maps. Therefore, we introduce a Clean-to-Noisy image generation framework, namely C2N, to imitate complex real-world noise without using any paired examples. We construct the noise generator in C2N accordingly with each component of real-world noise characteristics to express a wide range of noise accurately. Combined with our C2N, conventional denoising CNNs can be trained to outperform existing unsupervised methods on challenging real-world benchmarks by a large margin.

基于学习的图像去噪方法已被限制在以下情况：给出对齐的噪声和干净图像，或从预定噪声模型（例如高斯）合成样本。尽管最近的生成性噪声建模方法旨在模拟真实世界噪声的未知分布，但仍存在一些局限性。在实际场景中，噪声发生器应学会模拟一般和复杂的噪声分布，而无需使用成对的噪声和干净图像。然而，由于现有的方法是建立在对真实世界噪声的不现实假设上的，因此它们往往会产生不可信的模式，并且不能表示复杂的噪声映射。因此，我们引入了一个干净到噪声的图像生成框架，即C2N，来模拟复杂的真实世界噪声，而不使用任何成对的示例。我们在C2N中构造了相应的噪声发生器，每个分量都具有真实世界的噪声特性，以准确地表达大范围的噪声。结合我们的C2N，传统的去噪CNN可以在挑战现实世界基准测试时大大优于现有的无监督方法。

Continually learning in the real world must overcome many challenges, among which noisy labels are a common and inevitable issue. In this work, we present a replay-based continual learning framework that simultaneously addresses both catastrophic forgetting and noisy labels for the first time. Our solution is based on two observations; (i) forgetting can be mitigated even with noisy labels via self-supervised learning, and (ii) the purity of the replay buffer is crucial. Building on this regard, we propose two key components of our method: (i) a self-supervised replay technique named Self-Replay, which can circumvent erroneous training signals arising from noisy labeled data, and (ii) the Self-Centered filter that maintains a purified replay buffer via centrality-based stochastic graph ensembles. The empirical results on MNIST, CIFAR-10, CIFAR-100, and WebVision with real-world noise demonstrate that our framework can maintain a highly pure replay buffer amidst noisy streamed data while greatly outperforming the combinations of the state-of-the-art continual learning and noisy label learning methods.

在现实世界中不断学习必须克服许多挑战，其中噪音标签是一个常见且不可避免的问题。在这项工作中，我们提出了一个基于回放的连续学习框架，该框架首次同时解决了灾难性遗忘和噪音标签问题。我们的解决方案基于两个观察结果；(i) 通过自监督学习，即使使用有噪声的标签，遗忘也可以减轻，(ii) 重放缓冲区的纯度至关重要。基于这一点，我们提出了该方法的两个关键组成部分：(i) 一种称为自重放的自监督重放技术，它可以避免由噪音标记数据引起的错误训练信号；(ii) 自中心滤波器，它通过基于中心性的随机图集合维持一个纯净的重放缓冲区。在MNIST、CIFAR-10、CIFAR-100和WebVision上使用真实噪声的实证结果表明，我们的框架可以在噪声流数据中保持高纯度的重放缓冲区，同时大大优于最先进的连续学习和噪音标签学习方法的组合。

We introduce a method to render Neural Radiance Fields (NeRFs) in real time using PlenOctrees, an octree-based 3D representation which supports view-dependent effects. Our method can render 800x800 images at more than 150 FPS, which is over 3000 times faster than conventional NeRFs. We do so without sacrificing quality while preserving the ability of NeRFs to perform free-viewpoint rendering of scenes with arbitrary geometry and view-dependent effects. Real-time performance is achieved by pre-tabulating the NeRF into a PlenOctree. In order to preserve view-dependent effects such as specularities, we factorize the appearance via closed-form spherical basis functions. Specifically, we show that it is possible to train NeRFs to predict a spherical harmonic representation of radiance, removing the viewing direction as an input to the neural network. Furthermore, we show that PlenOctrees can be directly optimized to further minimize the reconstruction loss, which leads to equal or better quality compared to competing methods. Moreover, this octree optimization step can be used to reduce the training time, as we no longer need to wait for the NeRF training to converge fully. Our real-time neural rendering approach may potentially enable new applications such as 6-DOF industrial and product visualizations, as well as next generation AR/VR systems. PlenOctrees are amenable to in-browser rendering as well; please visit the project page for the interactive online demo, as well as video and code: <https://alexyu.net/plenoctrees>.

我们介绍了一种使用八叉树实时渲染神经辐射场 (NeRFs) 的方法，八叉树是一种基于八叉树的三维表示，支持视图相关的效果。我们的方法可以以超过150 FPS的速度渲染800x800幅图像，比传统的NeRFs快3000倍以上。我们在不牺牲质量的情况下这样做，同时保持NeRFs对具有任意几何体和视图相关效果的场景执行自由视点渲染的能力。实时性能是通过将NeRF预制表成一个八叉树来实现的。为了保留与视图相关的效果，例如镜面反射，我们通过闭合形式的球面基函数对外观进行因子分解。具体地说，我们证明了训练NeRFs预测辐射度的球谐表示是可能的，去掉了观察方向作为神经网络的输入。此外，我们还表明，可直接优化正八叉树以进一步最小化重建损失，从而与竞争方法相比，获得相同或更好的质量。此外，此八叉树优化步骤可用于减少训练时间，因为我们不再需要等待NeRF训练完全收敛。我们的实时神经渲染方法可能实现新的应用，如6自由度工业和产品可视化，以及下一代AR/VR系统。

PlenOctrees也适用于浏览器内渲染；请访问交互式在线演示的项目页面，以及视频和代码：<https://alexyu.net/plenoctrees>。

Deep neural networks (DNNs) for the semantic segmentation of images are usually trained to operate on a predefined closed set of object classes. This is in contrast to the ""open world"" setting where DNNs are envisioned to be deployed to. From a functional safety point of view, the ability to detect so-called ""out-of-distribution"" (OOD) samples, i.e., objects outside of a DNN's semantic space, is crucial for many applications such as automated driving. A natural baseline approach to OOD detection is to threshold on the pixel-wise softmax entropy. We present a two-step procedure that significantly improves that approach. Firstly, we utilize samples from the COCO dataset as OOD proxy and introduce a second training objective to maximize the softmax entropy on these samples. Starting from pretrained semantic segmentation networks we re-train a number of DNNs on different in-distribution datasets and consistently observe improved OOD detection performance when evaluating on completely disjoint OOD datasets. Secondly, we perform a transparent post-processing step to discard false positive OOD samples by so-called ""meta classification"". To this end, we apply linear models to a set of hand-crafted metrics derived from the DNN's softmax probabilities. In our experiments we consistently observe a clear additional gain in OOD detection performance, cutting down the number of detection errors by 52% when comparing the best baseline with our results. We achieve this improvement sacrificing only marginally in original segmentation performance. Therefore, our method contributes to safer DNNs with more reliable overall system performance.

用于图像语义分割的深度神经网络（DNN）通常被训练为在预定义的闭合对象类集上进行操作。这与设想将DNN部署到的“开放世界”设置形成对比。从功能安全的角度来看，检测所谓的“分布外”（OoD）样本（即DNN语义空间之外的对象）的能力对于许多应用（如自动驾驶）至关重要。OoD检测的一种自然基线方法是在像素级softmax熵上设置阈值。我们提出了一个两步程序，大大改进了该方法。首先，我们利用COCO数据集中的样本作为OoD代理，并引入第二个训练目标来最大化这些样本上的softmax熵。从预先训练的语义分割网络开始，我们在不同的分布数据集上重新训练了许多DNN，并在评估完全不相交的OoD数据集时一致地观察到改进的OoD检测性能。其次，我们执行透明的后处理步骤，通过所谓的“元分类”丢弃假阳性的OoD样本。为此，我们将线性模型应用于一组手工制作的指标，这些指标来自DNN的softmax概率。在我们的实验中，我们始终观察到在OoD检测性能方面有明显的额外增益，在将最佳基线与我们的结果进行比较时，检测错误的数量减少了52%。我们实现了这一改进，只牺牲了少量的原始分割性能。因此，我们的方法有助于实现更安全的DNN和更可靠的总体系统性能。

RGB-D saliency detection has attracted increasing attention, due to its effectiveness and the fact that depth cues can now be conveniently captured. Existing works often focus on learning a shared representation through various fusion strategies, with few methods explicitly considering how to preserve modality-specific characteristics. In this paper, taking a new perspective, we propose a specificity-preserving network for RGB-D saliency detection, which benefits saliency detection performance by exploring both the shared information and modality-specific properties (e.g., specificity). Specifically, two modality-specific networks and a shared learning network are adopted to generate individual and shared saliency maps. A cross-enhanced integration module (CIM) is proposed to fuse cross-modal features in the shared learning network, which are then propagated to the next layer for integrating cross-level information. Besides, we propose a multi-modal feature aggregation (MFA) module to integrate the modality-specific features from each individual decoder into the shared decoder, which can provide rich complementary multi-modal information to boost the saliency detection performance. Further, a skip connection is used to combine hierarchical features between the encoder and decoder layers. Experiments on six benchmark datasets demonstrate that our SP-Net outperforms other state-of-the-art methods.

RGB-D显著性检测由于其有效性和现在可以方便地捕获深度线索的事实而受到越来越多的关注。现有的研究通常侧重于通过各种融合策略学习共享表示，很少有方法明确考虑如何保留特定于模态的特征。在本文中，我们从一个新的角度提出了一种用于RGB-D显著性检测的特异性保持网络，该网络通过探索共享信息和特定于模态的属性（例如特异性）来提高显著性检测性能。具体而言，采用两个模态特定网络和一个共享学习网络来生成个体和共享显著性图。提出了一种交叉增强集成模块（CIM）来融合共享学习网络中的交叉模态特征，然后将这些特征传播到下一层以集成跨层次信息。此外，我们还提出了一个多模态特征聚合（MFA）模块，将各个解码器的模态特征集成到共享解码器中，从而提供丰富的互补多模态信息，提高显著性检测性能。此外，跳过连接用于组合编码器层和解码器层之间的分层特征。在六个基准数据集上的实验表明，我们的SP-Net优于其他最先进的方法。

Visual grounding on 3D point clouds is an emerging vision and language task that benefits various applications in understanding the 3D visual world. By formulating this task as a grounding-by-detection problem, lots of recent works focus on how to exploit more powerful detectors and comprehensive language features, but (1) how to model complex relations for generating context-aware object proposals and (2) how to leverage proposal relations to distinguish the true target object from similar proposals are not fully studied yet. Inspired by the well-known transformer architecture, we propose a relation-aware visual grounding method on 3D point clouds, named as 3DVG-Transformer, to fully utilize the contextual clues for relationenhanced proposal generation and cross-modal proposal disambiguation, which are enabled by a newly designed coordinate-guided contextual aggregation (CCA) module in the object proposal generation stage, and a multiplex attention (MA) module in the cross-modal feature fusion stage. We validate that our 3DVG-Transformer outperforms the state-of-the-art methods by a large margin, on two point cloud-based visual grounding datasets, ScanRefer and Nr3D/Sr3D from ReferIt3D, especially for complex scenarios containing multiple objects of the same category.

三维点云的视觉基础是一项新兴的视觉和语言任务，有利于理解三维视觉世界的各种应用。通过将此任务描述为一个基于检测的问题，最近的许多工作都集中在如何开发更强大的检测器和综合的语言特性，但是（1）如何对复杂关系建模以生成上下文感知的对象建议，以及（2）如何利用建议关系来区分真实目标对象和类似建议尚未得到充分研究。受著名的transformer架构的启发，我们提出了一种基于3D点云的关系感知视觉接地方法，称为3DVG transformer，以充分利用上下文线索，实现关系增强型提案生成和跨模式提案消歧，在对象建议生成阶段，新设计的坐标引导上下文聚合（CCA）模块和跨模态特征融合阶段的多重注意（MA）模块启用了这些功能。我们验证了我们的3DVG Transformer在两个基于点云的视觉接地数据集（来自ReferIt3D的ScanRefer和Nr3D/Sr3D）上大大优于最先进的方法，尤其是在包含相同类别多个对象的复杂场景中。

The non-local self-similarity property of natural images has been exploited extensively for solving various image processing problems. When it comes to video sequences, harnessing this force is even more beneficial due to the temporal redundancy. In the context of image and video denoising, many classically-oriented algorithms employ self-similarity, splitting the data into overlapping patches, gathering groups of similar ones and processing these together somehow. With the emergence of convolutional neural networks (CNN), the patch-based framework has been abandoned. Most CNN denoisers operate on the whole image, leveraging non-local relations only implicitly by using a large receptive field. This work proposes a novel approach for leveraging self-similarity in the context of video denoising, while still relying on a regular convolutional architecture. We introduce a concept of patch-craft frames - artificial frames that are similar to the real ones, built by tiling matched patches. Our algorithm augments video sequences with patch-craft frames and feeds them to a CNN. We demonstrate the substantial boost in denoising performance obtained with the proposed approach.

自然图像的非局部自相似特性已被广泛应用于解决各种图像处理问题。当涉及到视频序列时，由于时间冗余，利用这种力量更为有利。在图像和视频去噪方面，许多面向经典的算法采用自相似性，将数据分割成重叠的块，收集相似块的组并以某种方式处理这些块。随着卷积神经网络（CNN）的出现，基于补丁的框架已经被抛弃。大多数CNN去噪器对整个图像进行操作，仅通过使用一个大的感受野隐式地利用非局部关系。这项工作提出了一种在视频去噪环境中利用自相似性的新方法，同时仍然依赖于常规卷积结构。我们介绍了补丁工艺框架的概念——与真实框架相似的人造框架，通过拼接匹配的补丁构建。我们的算法使用patch craft帧增强视频序列，并将其反馈给CNN。我们证明了所提出的方法在去噪性能上的显著提升。

Text-based image retrieval has seen considerable progress in recent years. However, the performance of existing methods suffers in real life since the user is likely to provide an incomplete description of an image, which often leads to results filled with false positives that fit the incomplete description. In this work, we introduce the partial-query problem and extensively analyze its influence on text-based image retrieval. Previous interactive methods tackle the problem by passively receiving users' feedback to supplement the incomplete query iteratively, which is time-consuming and requires heavy user effort. Instead, we propose a novel retrieval framework that conducts the interactive process in an Ask-and-Confirm fashion, where AI actively searches for discriminative details missing in the current query, and users only need to confirm AI's proposal. Specifically, we propose an object-based interaction to make the interactive retrieval more user-friendly and present a reinforcement-learning-based policy to search for discriminative objects. Furthermore, since fully-supervised training is often infeasible due to the difficulty of obtaining human-machine dialog data, we present a weakly-supervised training strategy that needs no human-annotated dialogs other than a text-image dataset. Experiments show that our framework significantly improves the performance of text-based image retrieval. Code is available at <https://github.com/CuthbertCai/Ask-Confirm>.

近年来，基于文本的图像检索取得了长足的进展。然而，现有方法的性能在现实生活中会受到影响，因为用户可能会提供图像的不完整描述，这通常会导致结果充满符合不完整描述的误报。在这项工作中，我们介绍了部分查询问题，并广泛分析了它对基于文本的图像检索的影响。以前的交互方法通过被动地接收用户的反馈来解决这个问题，以迭代的方式补充不完整的查询，这既耗时又需要用户付出巨大的努力。相反，我们提出了一种新的检索框架，该框架以询问和确认的方式进行交互过程，AI主动搜索当前查询中缺失的区别性细节，用户只需确认AI的提议。具体地说，我们提出了一种基于对象的交互方式，使得交互检索更加友好，并提出了一种基于强化学习的策略来搜索有区别的对象。此外，由于很难获得人机对话数据，因此完全监督训练通常是不可行的，因此我们提出了一种弱监督训练策略，除了文本图像数据集之外，不需要人工注释对话。实验表明，该框架显著提高了基于文本的图像检索的性能。代码可在<https://github.com/CuthbertCai/Ask-Confirm>.

The successful deployment of artificial intelligence (AI) in many domains from healthcare to hiring requires their responsible use, particularly in model explanations and privacy. Explainable artificial intelligence (XAI) provides more information to help users to understand model decisions, yet this additional knowledge exposes additional risks for privacy attacks. Hence, providing explanation harms privacy. We study this risk for image-based model inversion attacks and identified several attack architectures with increasing performance to reconstruct private image data from model explanations. We have developed several multi-modal transposed CNN architectures that achieve significantly higher inversion performance than using the target model prediction only. These XAI-aware inversion models were designed to exploit the spatial knowledge in image explanations. To understand which explanations have higher privacy risk, we analyzed how various explanation types and factors influence inversion performance. In spite of some models not providing explanations, we further demonstrate increased inversion performance even for non-explainable target models by exploiting explanations of surrogate models through attention transfer. This method first inverts an explanation from the target prediction, then reconstructs the target image. These threats highlight the urgent and significant privacy risks of explanations and calls attention for new privacy preservation techniques that balance the dual-requirement for AI explainability and privacy.

人工智能（AI）在从医疗保健到招聘等许多领域的成功部署需要其负责任的使用，特别是在模型解释和隐私方面。可解释人工智能（XAI）提供了更多信息来帮助用户理解模型决策，但这些额外的知识暴露了隐私攻击的额外风险。因此，提供解释会损害隐私。我们研究了基于图像的模型反转攻击的这种风险，并确定了几种性能提高的攻击体系结构，以从模型解释中重建私有图像数据。我们已经开发了几种多模式转置CNN结构，与仅使用目标模型预测相比，它们实现了显著更高的反演性能。这些XAI感知反演模型旨在利用图像解释中的空间知识。为了了解哪些解释具有更高的隐私风险，我们分析了各种解释类型和因素如何影响反转性能。尽管有些模型没有提供解释，但我们通过注意转移对替代模型的解释，进一步证明即使对于不可解释的目标模型，反转性能也有所提高。该方法首先从目标预测中反转解释，然后重建目标图像。这些威胁突出了解释的紧迫和重大隐私风险，并呼吁关注新的隐私保护技术，以平衡AI解释性和隐私的双重要求。

Training a neural network model for recognizing multiple labels associated with an image, including identifying unseen labels, is challenging, especially for images that portray numerous semantically diverse labels. As challenging as this task is, it is an essential task to tackle since it represents many real-world cases, such as image retrieval of natural images. We argue that using a single embedding vector to represent an image, as commonly practiced, is not sufficient to rank both relevant seen and unseen labels accurately. This study introduces an end-to-end model training for multi-label zero-shot learning that supports the semantic diversity of the images and labels. We propose to use an embedding matrix having principal embedding vectors trained using a tailored loss function. In addition, during training, we suggest up-weighting in the loss function image samples presenting higher semantic diversity to encourage the diversity of the embedding matrix. Extensive experiments show that our proposed method improves the zero-shot model's quality in tag-based image retrieval achieving SoTA results on several common datasets (NUS-Wide, COCO, Open Images).

训练神经网络模型以识别与图像相关联的多个标签（包括识别看不见的标签）是一项挑战，特别是对于描绘大量语义不同标签的图像。尽管这项任务具有挑战性，但它是一项必须解决的任务，因为它代表了许多实际情况，例如自然图像的图像检索。我们认为，使用一个单一的嵌入向量来表示一幅图像，通常的做法是，不足以准确地对相关的可见和不可见标签进行排序。本研究介绍了一种支持图像和标签语义多样性的多标签零镜头学习的端到端模型训练。我们建议使用一个嵌入矩阵，该矩阵具有使用定制损失函数训练的主嵌入向量。此外，在训练过程中，我们建议在呈现较高语义多样性的损失函数图像样本中增加权重，以鼓励嵌入矩阵的多样性。大量实验表明，我们提出的方法提高了基于标签的图像检索中零镜头模型的质量，在几个常见数据集（NUS-Wide、COCO、Open Images）上实现了SoTA结果。

Existing change captioning studies have mainly focused on a single change. However, detecting and describing multiple changed parts in image pairs is essential for enhancing adaptability to complex scenarios. We solve the above issues from three aspects: (i) we propose a simulation-based multi-change captioning dataset; (ii) we benchmark existing state-of-the-art methods of single change captioning on multi-change captioning; (iii) we further propose Multi-Change Captioning transformers (MCCFormers) that identify change regions by densely correlating different regions in image pairs and dynamically determines the related change regions with words in sentences. The proposed method obtained the highest scores on four conventional change captioning evaluation metrics for multi-change captioning. Additionally, our proposed method can separate attention maps for each change and performs well with respect to change localization. Moreover, the proposed framework outperformed the previous state-of-the-art methods on an existing change captioning benchmark, CLEVR-Change, by a large margin (+6.1 on BLEU-4 and +9.7 on CIDEr scores), indicating its general ability in change captioning tasks. The code and dataset are available at the project page.

现有的变更字幕研究主要集中在单个变更上。然而，检测和描述图像对中的多个变化部分对于增强对复杂场景的适应性至关重要。我们从三个方面解决了上述问题：(i) 提出了一种基于模拟的多变化字幕数据集；(ii) 我们将现有最先进的单变更字幕制作方法与多变更字幕制作方法相比较；(iii) 我们进一步提出了多变化字幕转换器 (MCCFormers)，该转换器通过图像对中不同区域的密集关联来识别变化区域，并通过句子中的单词动态确定相关变化区域。该方法在四个传统的多变化字幕评价指标上获得了最高的分数。此外，我们提出的方法可以为每个变化分离注意图，并在变化定位方面表现良好。此外，提议的框架在现有变更字幕基准CLEVR change上的表现大大优于先前最先进的方法 (BLEU-4为+6.1，苹果酒分为+9.7)，表明其在变更字幕任务方面的总体能力。代码和数据集可在项目页面中找到。

Point clouds acquired from scanning devices are often perturbed by noise, which affects downstream tasks such as surface reconstruction and analysis. The distribution of a noisy point cloud can be viewed as the distribution of a set of noise-free samples  $p(x)$  convolved with some noise model  $n$ , leading to  $(p * n)(x)$  whose mode is the underlying clean surface. To denoise a noisy point cloud, we propose to increase the log-likelihood of each point from  $p * n$  via gradient ascent---iteratively updating each point's position. Since  $p * n$  is unknown at test-time, and we only need the score (i.e., the gradient of the log-probability function) to perform gradient ascent, we propose a neural network architecture to estimate the score of  $p * n$  given only noisy point clouds as input. We derive objective functions for training the network and develop a denoising algorithm leveraging on the estimated scores. Experiments demonstrate that the proposed model outperforms state-of-the-art methods under a variety of noise models, and shows the potential to be applied in other tasks such as point cloud upsampling.

从扫描设备获取的点云经常受到噪声的干扰，这会影响下游任务，如曲面重建和分析。噪声点云的分布可以看作是一组无噪声样本 $p(x)$ 的分布，这些样本与一些噪声模型 $n$ 卷积在一起，导致 $(pn)(x)$ 的模式是底层清洁表面。为了消除噪声点云，我们建议通过梯度上升从 $pn$ 增加每个点的对数似然——迭代更新每个点的位置。由于 $pn$ 在测试时是未知的，并且我们只需要分数（即对数概率函数的梯度）来执行梯度上升，因此我们提出了一种神经网络结构来估计 $pn$ 的分数，仅将噪声点云作为输入。我们推导了训练网络的目标函数，并开发了一种基于估计分数的去噪算法。实验表明，在各种噪声模型下，该模型的性能优于现有的方法，并显示出在其他任务（如点云上采样）中应用的潜力。

Unprecedented access to multi-temporal satellite imagery has opened new perspectives for a variety of Earth observation tasks. Among them, pixel-precise panoptic segmentation of agricultural parcels has major economic and environmental implications. While researchers have explored this problem for single images, we argue that the complex temporal patterns of crop phenology are better addressed with temporal sequences of images. In this paper, we present the first end-to-end, single-stage method for panoptic segmentation of Satellite Image Time Series (SITS). This module can be combined with our novel image sequence encoding network which relies on temporal self-attention to extract rich and adaptive multi-scale spatio-temporal features. We also introduce PASTIS, the first open-access SITS dataset with panoptic annotations. We demonstrate the superiority of our encoder for semantic segmentation against multiple competing network architectures, and set up the first state-of-the-art of panoptic segmentation of SITS. Our implementation and the PASTIS dataset are publicly available at ([link-upon-publication](#)).

前所未有的多时相卫星图像为各种地球观测任务开辟了新的前景。其中，农业地块的像素精确全景分割具有重大的经济和环境影响。虽然研究人员对单个图像的这一问题进行了探索，但我们认为，利用图像的时间序列可以更好地解决作物物候的复杂时间模式。在本文中，我们提出了第一种端到端、单阶段的卫星图像时间序列全景分割方法。该模块可以与我们的新型图像序列编码网络相结合，该网络依赖于时间自我注意来提取丰富且自适应的多尺度时空特征。我们还介绍了PASTIS，这是第一个具有全景注释的开放访问SITS数据集。我们展示了我们的编码器在语义分割方面相对于多种竞争性网络架构的优

势，并建立了第一个最先进的SIT全景分割。我们的实现和PASTIS数据集可在（发布时链接）上公开获取。

Although convolutional neural networks (CNNs) have achieved great success in computer vision, this work investigates a simpler, convolution-free backbone network useful for many dense prediction tasks. Unlike the recently-proposed Vision Transformer (ViT) that was designed for image classification specifically, we introduce the Pyramid Vision Transformer (PVT), which overcomes the difficulties of porting Transformer to various dense prediction tasks. PVT has several merits compared to current state of the arts. (1) Different from ViT that typically yields low-resolution outputs and incurs high computational and memory costs, PVT not only can be trained on dense partitions of an image to achieve high output resolution, which is important for dense prediction, but also uses a progressive shrinking pyramid to reduce the computations of large feature maps. (2) PVT inherits the advantages of both CNN and Transformer, making it a unified backbone for various vision tasks without convolutions, where it can be used as a direct replacement for CNN backbones. (3) we validate PVT through extensive experiments, showing that it boosts the performance of many downstream tasks, including object detection, instance and semantic segmentation. For example, with a comparable number of parameters, PVT+RetinaNet achieves 40.4 AP on the COCO dataset, surpassing ResNet50+RetinaNet (36.3 AP) by 4.1 absolute AP. We hope that PVT could serve as an alternative and useful backbone for pixel-level predictions and facilitate future research.

虽然卷积神经网络 (CNN) 在计算机视觉领域取得了巨大的成功，但本文研究的是一种更简单、无卷积的主干网络，可用于许多密集的预测任务。与最近提出的专门用于图像分类的视觉变换器 (ViT) 不同，我们引入了金字塔视觉变换器 (PVT)，它克服了将变换器移植到各种密集预测任务的困难。和目前的技术相比，PVT有几个优点。 (1) 与通常产生低分辨率输出并产生高计算和内存成本的ViT不同，PVT不仅可以在图像的密集分区上进行训练以获得高输出分辨率，这对于密集预测很重要，而且还使用渐进收缩金字塔来减少大型特征图的计算。 (2) PVT继承了CNN和Transformer的优点，使其成为各种视觉任务的统一主干，无需卷积，可直接替代CNN主干。 (3) 我们通过大量实验证证了PVT，表明它提高了许多下游任务的性能，包括对象检测、实例和语义分割。例如，在参数数量相当的情况下，PVT+RetinaNet在COCO数据集上达到40.4 AP，超过ResNet50+RetinaNet (36.3 AP) 4.1绝对AP。我们希望PVT可以作为像素级预测的替代和有用的主干，并促进未来的研究。

Causally-taken images often suffer from flare artifacts, due to the unintended reflections and scattering of light inside the camera. However, as flares may appear in a variety of shapes, positions, and colors, detecting and removing them entirely from an image is very challenging. Existing methods rely on predefined intensity and geometry priors of flares, and may fail to distinguish the difference between light sources and flare artifacts. We observe that the conditions of the light source in the image play an important role in the resulting flares. In this paper, we present a deep framework with light source aware guidance for single-image flare removal (SIFR). In particular, we first detect the light source regions and the flare regions separately, and then remove the flare artifacts based on the light source aware guidance. By learning the underlying relationships between the two types of regions, our approach can remove different kinds of flares from the image. In addition, instead of using paired training data which are difficult to collect, we propose the first unpaired flare removal dataset and new cycle-consistency constraints to obtain more diverse examples and avoid manual annotations. Extensive experiments demonstrate that our method outperforms the baselines qualitatively and quantitatively. We also show that our model can be applied to flare effect manipulation (e.g., adding or changing image flares).

由于相机内部光线的非预期反射和散射，因果拍摄的图像通常会出现耀斑伪影。然而，由于耀斑可能以各种形状、位置和颜色出现，因此完全从图像中检测和移除它们是非常具有挑战性的。现有方法依赖于耀斑的预定义强度和几何先验，可能无法区分光源和耀斑伪影之间的差异。我们观察到，图像中光源的条件对产生的耀斑起着重要作用。在这篇文章中，我们提出了一个深的框架与光源意识指导单一图像耀斑消除 (SIFR)。特别是，我们首先分别检测光源区域和光斑区域，然后基于光源感知制导去除光斑伪影。通过学习这两类区域之间的潜在关系，我们的方法可以从图像中去除不同类型的耀斑。此外，我们没有使用难以收集的成对训练数据，而是提出了第一个未配对的火炬移除数据集和新的周期一致性约束，以获得更多不同的示例，并避免手动注释。大量实验表明，我们的方法在定性和定量上都优于基线。我们还表明，我们的模型可以应用于耀斑效果操纵（例如，添加或更改图像耀斑）。

Dynamic inference networks, aimed at promoting computational efficiency, go along an adaptive executing path for a given sample. Prevalent methods typically assign a router for each convolutional block and sequentially make block-by-block executing decisions, without considering the relations during the dynamic inference. In this paper, we model the relations for dynamic inference from two aspects: the routers and the samples. We design a novel type of router called the relational router to model the relations among routers for a given sample. In principle, the current relational router aggregates the contextual features of preceding routers by graph convolution and propagates its router features to subsequent ones, making the executing decision for the current block in a long-range manner. Furthermore, we model the relation between samples by introducing a Sample Relation Module (SRM), encouraging correlated samples to go along correlated executing paths. As a whole, we call our method the Relational Dynamic Inference Network (RDI-Net). Extensive experiments on CIFAR-10/100 and ImageNet show that RDI-Net achieves state-of-the-art performance and computational cost reduction. Our code and models will be made publicly available.

为了提高计算效率，动态推理网络对给定的样本沿着一条自适应的执行路径运行。流行的方法通常为每个卷积块分配一个路由器，并依次逐块执行决策，而不考虑动态推理过程中的关系。本文从路由器和样本两个方面对动态推理关系进行建模。我们设计了一种称为关系路由器的新型路由器，以对给定样本的路由器之间的关系进行建模。原则上，当前关系路由器通过图卷积聚合先前路由器的上下文特征，并将其路由器特征传播到后续路由器，以远程方式对当前块执行决策。此外，我们通过引入样本关系模块 (SRM) 对样本之间的关系进行建模，鼓励相关样本沿着相关执行路径进行。作为一个整体，我们称我们的方法为关系动态推理网络 (RDI网络)。在CIFAR-10/100和ImageNet上进行的大量实验表明，RDI-Net实现了最先进的性能并降低了计算成本。我们的代码和模型将公开提供。

This paper introduces an unsupervised loss for training parametric deformation shape generators. The key idea is to enforce the preservation of local rigidity among the generated shapes. Our approach builds on a local approximation of the as-rigid-as possible (or ARAP) deformation energy. We show how to develop the unsupervised loss via a spectral decomposition of the Hessian of the ARAP loss. Our loss nicely decouples pose and shape variations through a robust norm. The loss admits simple closed-form expressions. It is easy to train and can be plugged into any standard generation models, e.g., VAE and GAN. Experimental results show that our approach outperforms existing shape generation approaches considerably across various datasets such as DFAUST, Animal, and Bone.

本文介绍了一种用于训练参数变形形状发生器的无监督损失法。关键思想是在生成的形状中加强局部刚性的保持。我们的方法建立在尽可能刚性（或ARAP）变形能量的局部近似上。我们展示了如何通过ARAP损失的Hessian谱分解来发展无监督损失。我们的损失通过一个稳健的标准很好地解耦了姿势和形状的变化。损失允许简单的闭式表达式。它易于培训，可插入任何标准代机型，如VAE和GAN。实验结果表明，在各种数据集（如DFAUST、动物和骨骼）中，我们的方法比现有的形状生成方法有很大的优势。

We present a two-stage learning framework for weakly supervised object localization (WSOL). While most previous efforts rely on high-level feature based CAMs (Class Activation Maps), this paper proposes to localize objects using the low-level feature based activation maps. In the first stage, an activation map generator produces activation maps based on the low-level feature maps in the classifier, such that rich contextual object information is included in an online manner. In the second stage, we employ an evaluator to evaluate the activation maps predicted by the activation map generator. Based on this, we further propose a weighted entropy loss, an attentive erasing, and an area loss to drive the activation map generator to substantially reduce the uncertainty of activations between object and background, and explore less discriminative regions. Based on the low-level object information preserved in the first stage, the second stage model gradually generates a well-separated, complete, and compact activation map of object in the image, which can be easily thresholded for accurate localization. Extensive experiments on CUB-200-2011 and ImageNet-1K datasets show that our framework surpasses previous methods by a large margin, which sets a new state-of-the-art for WSOL. Code will be available soon.

我们提出了一个用于弱监督目标定位 (WSOL) 的两阶段学习框架。虽然大多数以前的工作依赖于基于高级特征的CAM (类激活映射)，但本文建议使用基于低级特征的激活映射来定位对象。在第一阶段中，激活映射生成器基于分类器中的低级特征映射生成激活映射，从而以在线方式包括丰富的上下文对象信息。在第二阶段，我们使用评估器评估激活图生成器预测的激活图。在此基础上，我们进一步提出了加权熵损失、注意擦除和面积损失来驱动激活图生成器，以大幅降低对象和背景之间激活的不确定性，并探索较少辨别的区域。第二阶段模型基于第一阶段保存的低层目标信息，逐步生成图像中分离良好、完整且紧凑的目标激活图，该激活图易于阈值化以实现精确定位。在CUB-200-2011和ImageNet-1K数据集上的大量实验表明，我们的框架大大超过了以前的方法，这为WSOL创造了一个新的技术水平。代码将很快提供。

Research in unpaired video translation has mainly focused on short-term temporal consistency by conditioning on neighboring frames. However for transfer from simulated to photorealistic sequences, available information on the underlying geometry offers potential for achieving global consistency across views. We propose a novel approach which combines unpaired image translation with neural rendering to transfer simulated to photorealistic surgical abdominal scenes. By introducing global learnable textures and a lighting-invariant view-consistency loss, our method produces consistent translations of arbitrary views and thus enables long-term consistent video synthesis. We design and test our model to generate video sequences from minimally-invasive surgical abdominal scenes. Because labeled data is often limited in this domain, photorealistic data where ground truth information from the simulated domain is preserved is especially relevant. By extending existing image-based methods to view-consistent videos, we aim to impact the applicability of simulated training and evaluation environments for surgical applications. Code and data: <http://opencas.dkfz.de/video-sim2real>.

未配对视频翻译的研究主要集中在通过对相邻帧进行条件处理来实现短时时间一致性。但是，对于从模拟序列到真实照片序列的传输，有关底层几何体的可用信息提供了实现视图全局一致性的可能性。我们提出了一种新的方法，该方法将未成对图像转换与神经渲染相结合，以将模拟的腹部手术场景转换为照片级真实感场景。通过引入全局可学习纹理和光照不变的视图一致性损失，我们的方法生成任意视图的一致性转换，从而实现长期一致的视频合成。我们设计并测试我们的模型，从微创外科腹部场景生成视频序列。由于标记数据通常局限于该领域，因此保存模拟领域地面真实信息的照片级真实感数据尤其重要。通过扩展现有的基于图像的方法以查看一致的视频，我们旨在影响模拟培训和评估环境在外科应用中的适用性。代码和数据：<http://opencas.dkfz.de/video-sim2real>。

Online semantic segmentation on a time series of point cloud frames is an essential task in autonomous driving. Existing models focus on single-frame segmentation, which cannot achieve satisfactory segmentation accuracy and offer unstably flicker among frames. In this paper, we propose a light-weight semantic segmentation framework for large-scale point cloud series, called TempNet, which can improve both the accuracy and the stability of existing semantic segmentation models by combining a novel frame aggregation scheme. To be computational cost efficient, feature extraction and aggregation are only conducted on a small portion of key frames via a temporal feature aggregation (TFA) network using an attentional pooling mechanism, and such enhanced features are propagated to the intermediate non-key frames. To avoid information loss from non-key frames, a partial feature update (PFU) network is designed to partially update the propagated features with the local features extracted on a non-key frame if a large disparity between the two is quickly assessed. As a result, consistent and information-rich features can be obtained for each frame. We implement TempNet on five state-of-the-art (SOTA) point cloud segmentation models and conduct extensive experiments on the SemanticKITTI dataset. Results demonstrate that TempNet outperforms SOTA competitors by wide margins with little extra computational cost.

基于点云帧时间序列的在线语义分割是自动驾驶中的一项重要任务。现有的模型侧重于单帧分割，不能达到令人满意的分割精度，并且帧间闪烁不稳定。在本文中，我们提出了一种用于大规模点云序列的轻量级语义分割框架TempNet，该框架通过结合一种新的框架聚合方案，可以提高现有语义分割模型的准确性和稳定性。为了节省计算成本，特征提取和聚合仅通过使用注意池机制的时间特征聚合（TFA）网络在关键帧的一小部分上进行，并且这种增强的特征被传播到中间非关键帧。为了避免非关键帧的信息丢失，设计了一个部分特征更新（PFU）网络，在快速评估非关键帧之间的较大差异时，使用在非关键帧上提取的局部特征部分更新传播的特征。因此，可以为每个帧获得一致且信息丰富的特征。我们在五种最先进的（SOTA）点云分割模型上实现了TempNet，并在SemanticKITTI数据集上进行了大量实验。结果表明，TempNet在几乎不增加额外计算成本的情况下比SOTA竞争对手有较大的优势。

Few-shot object detection (FSOD) aims to detect never-seen objects using few examples. This field sees recent improvement owing to the meta-learning techniques by learning how to match between the query image and few-shot class examples, such that the learned model can generalize to few-shot novel classes. However, currently, most of the meta-learning-based methods perform pairwise matching between query image regions (usually proposals) and novel classes separately, therefore failing to take into account multiple relationships among them. In this paper, we propose a novel FSOD model using heterogeneous graph convolutional networks. Through efficient message passing among all the proposal and class nodes with three different types of edges, we could obtain context-aware proposal features and query-adaptive, multiclass-enhanced prototype representations for each class, which could help promote the pairwise matching and improve final FSOD accuracy. Extensive experimental results show that our proposed model, denoted as QA-FewDet, outperforms the current state-of-the-art approaches on the PASCAL VOC and MSCOCO FSOD benchmarks under different shots and evaluation metrics.

少镜头目标检测（FSOD）旨在使用很少的示例来检测从未见过的目标。由于元学习技术的出现，该领域最近得到了改进，通过学习如何在查询图像和少数镜头类示例之间进行匹配，使得学习的模型可以推广到少数镜头类。然而，目前大多数基于元学习的方法在查询图像区域（通常是建议）和新类之间分别执行pairwise匹配，因此没有考虑它们之间的多种关系。本文提出了一种基于异构图卷积网络的FSOD模型。通过在具有三种不同边缘类型的所有提议和类节点之间高效地传递消息，我们可以获得上下文感知的提议特征，并为每个类查询自适应的、多类增强的原型表示，这有助于促进成对匹配并提高最终的

FSOD精度。大量的实验结果表明，我们提出的模型（称为QA FewDet）在PASCAL VOC和MSCOCO FSOD基准上，在不同的放炮和评估指标下，优于当前最先进的方法。

We propose ResRep, a novel method for lossless channel pruning (a.k.a. filter pruning), which slims down a CNN by reducing the width (number of output channels) of convolutional layers. Inspired by the neurobiology research about the independence of remembering and forgetting, we propose to re-parameterize a CNN into the remembering parts and forgetting parts, where the former learn to maintain the performance and the latter learn to prune. via training with regular SGD on the former but a novel update rule with penalty gradients on the latter, we realize structured sparsity. Then we equivalently merge the remembering and forgetting parts into the original architecture with narrower layers. In this sense, ResRep can be viewed as a successful application of Structural Re-parameterization. Such a methodology distinguishes ResRep from the traditional learning-based pruning paradigm that applies a penalty on parameters to produce sparsity, which may suppress the parameters essential for the remembering. ResRep slims down a standard ResNet-50 with 76.15% accuracy on ImageNet to a narrower one with only 45% FLOPs and no accuracy drop, which is the first to achieve lossless pruning with such a high compression ratio. The code and models are at <https://github.com/DingXiaoH/ResRep>.

我们提出了一种新的无损信道修剪方法ResRep（也称为滤波器修剪），它通过减小卷积层的宽度（输出信道的数量）来精简CNN。受关于记忆和遗忘独立性的神经生物学研究的启发，我们建议将CNN重新参数化为记忆部分和遗忘部分，前者学习保持表现，后者学习修剪。通过在前者上使用常规SGD进行训练，而在后者上使用带有惩罚梯度的新更新规则，我们实现了结构化稀疏性。然后，我们将记忆和遗忘部分等效地合并到具有较窄层的原始体系结构中。从这个意义上讲，ResRep可以看作是结构再参数化的成功应用。这种方法不同于传统的基于学习的剪枝范式，传统的剪枝范式对参数施加惩罚以产生稀疏性，稀疏性可能会抑制记忆所必需的参数。ResRep将ImageNet上具有76.15%准确度的标准ResNet-50精简为仅具有45%浮点且无准确度下降的更窄版本，这是第一个以如此高的压缩比实现无损修剪的版本。代码和模型位于<https://github.com/DingXiaoH/ResRep>。

Unsupervised domain adaptation (DA) has gained substantial interest in semantic segmentation. However, almost all prior arts assume concurrent access to both labeled source and unlabeled target, making them unsuitable for scenarios demanding source-free adaptation. In this work, we enable source-free DA by partitioning the task into two: a) source-only domain generalization and b) source-free target adaptation. Towards the former, we provide theoretical insights to develop a multi-head framework trained with a virtually extended multi-source dataset, aiming to balance generalization and specificity. Towards the latter, we utilize the multi-head framework to extract reliable target pseudo-labels for self-training. Additionally, we introduce a novel conditional prior-enforcing auto-encoder that discourages spatial irregularities, thereby enhancing the pseudo-label quality. Experiments on the standard GTA5-to-Cityscapes and SYNTHIA-to-Cityscapes benchmarks show our superiority even against the non-source-free prior-arts. Further, we show our compatibility with online adaptation enabling deployment in a sequentially changing environment.

无监督领域自适应（DA）在语义分割中得到了广泛的关注。然而，几乎所有现有技术都假设同时访问标记源和未标记目标，这使得它们不适合要求无源适配的场景。在这项工作中，我们通过将任务分为两部分来实现无源DA：a) 仅源域泛化和b) 无源目标自适应。对于前者，我们提供了理论见解，以开发一个使用虚拟扩展的多源数据集训练的多头框架，旨在平衡泛化和专用性。对于后者，我们利用多头框架提取可靠的目标伪标签进行自我训练。此外，我们还引入了一种新的条件先验强制自动编码器，该编码器可防止空间不规则，从而提高伪标签质量。在标准GTA5到Cityscapes和SYNTHIA到Cityscapes基准上的

实验表明，即使与无源代码的现有技术相比，我们也具有优势。此外，我们还展示了我们与在线自适应的兼容性，从而能够在连续变化的环境中进行部署。

We develop a conceptually simple, flexible, and effective framework (named T-Net) for two-view correspondence learning. Given a set of putative correspondences, we reject outliers and regress the relative pose encoded by the essential matrix, by an end-to-end framework, which is consisted of two novel structures: “-” structure and “|” structure. “-” structure adopts an iterative strategy to learn correspondence features. “|” structure integrates all the features of the iterations and outputs the correspondence weight. In addition, we introduce Permutation-Equivariant Context Squeeze-and-Excitation module, an adapted version of SE module, to process sparse correspondences in a permutation-equivariant way and capture both global and channel-wise contextual information. Extensive experiments on outdoor and indoor scenes show that the proposed T-Net achieves state-of-the-art performance. On outdoor scenes (YFCC100M dataset), T-Net achieves an mAP of 52.28%, a 34.22% precision increase from the best-published result (38.95%). On indoor scenes (SUN3D dataset), T-Net (19.71%) obtains a 21.82% precision increase from the best-published result (16.18%).

我们开发了一个概念简单、灵活、有效的双视图对应学习框架 (T-Net)。给定一组假定的对应关系，我们拒绝异常值，并通过端到端框架回归基本矩阵编码的相对姿势，该框架由两种新结构组成：“-”结构和“|”结构。“-”结构采用迭代策略学习对应特征。“|”结构集成了迭代的所有特征，并输出对应的权重。此外，我们还引入了置换等变上下文压缩和激发模块 (SE模块的一个改进版本)，以置换等变的方式处理稀疏对应，并捕获全局和通道上下文信息。对室外和室内场景的大量实验表明，所提出的T-Net达到了最先进的性能。在室外场景 (YFCC100M数据集) 上，T-Net获得了52.28%的地图，比最佳公布结果 (38.95%) 提高了34.22%。在室内场景 (SUN3D数据集) 上，T-Net (19.71%) 的精度比最佳公布结果 (16.18%) 提高了21.82%。

Video highlight detection plays an increasingly important role in social media content filtering, however, it remains highly challenging to develop automated video highlight detection methods because of the lack of temporal annotations (i.e., where the highlight moments are in long videos) for supervised learning. In this paper, we propose a novel weakly supervised method that can learn to detect highlights by mining video characteristics with video level annotations (topic tags) only. Particularly, we exploit audio-visual features to enhance video representation and take temporal cues into account for improving detection performance. Our contributions are threefold: 1) we propose an audio-visual tensor fusion mechanism that efficiently models the complex association between two modalities while reducing the gap of the heterogeneity between the two modalities; 2) we introduce a novel hierarchical temporal context encoder to embed local temporal clues in between neighboring segments; 3) finally, we alleviate the gradient vanishing problem theoretically during model optimization with attention-gated instance aggregation. Extensive experiments on two benchmark datasets (YouTube Highlights and TVSum) have demonstrated our method outperforms other state-of-the-art methods with remarkable improvements.

视频高光检测在社交媒体内容过滤中扮演着越来越重要的角色，然而，由于缺乏用于监督学习的时间注释（即，高光时刻位于长视频中），因此开发自动视频高光检测方法仍然是一项高度挑战。在本文中，我们提出了一种新的弱监督方法，该方法可以通过仅使用视频级别注释（主题标签）挖掘视频特征来学习检测高光。特别地，我们利用视听特征来增强视频表示，并考虑时间线索来提高检测性能。我们的贡献有三个方面：1) 我们提出了一种视听张量融合机制，有效地模拟了两种模式之间的复杂关联，同时缩小了两种模式之间的异质性差距；2) 我们引入了一种新的分层时间上下文编码器，将局部时间线索嵌入相邻片段之间；3) 最后，我们从理论上缓解了基于注意门的实例聚合模型优化过程中的梯度消失问

题。在两个基准数据集 (YouTube Highlights和TVSum) 上的大量实验表明，我们的方法优于其他最先进的方法，并有显著的改进。

Benefiting from large-scale pre-training, we have witnessed significant performance boost on the popular Visual Question Answering (VQA) task. Despite rapid progress, it remains unclear whether these state-of-the-art (SOTA) models are robust when encountering examples in the wild. To study this, we introduce Adversarial VQA, a new large-scale VQA benchmark, collected iteratively via an adversarial human-and-model-in-the-loop procedure. Through this new benchmark, we discover several interesting findings. (i) Surprisingly, we find that during dataset collection, non-expert annotators can easily attack SOTA VQA models successfully. (ii) Both large-scale pre-trained models and adversarial training methods achieve far worse performance on the new benchmark than over standard VQA v2 dataset, revealing the fragility of these models while demonstrating the effectiveness of our adversarial dataset. (iii) when used for data augmentation, our dataset can effectively boost model performance on other robust VQA benchmarks. We hope our Adversarial VQA dataset can shed new light on robustness study in the community and serve as a valuable benchmark for future work.

得益于大规模的预培训，我们见证了流行的视觉问答 (VQA) 任务的显著性能提升。尽管取得了快速的进展，但这些最先进的 (SOTA) 模型在野外遇到例子时是否可靠仍不清楚。为了研究这一点，我们引入了对抗式VQA，这是一种新的大规模VQA基准测试，通过对抗式人和模型在环过程迭代收集。通过这个新的基准，我们发现了几个有趣的发现。 (i) 令人惊讶的是，我们发现在数据集收集过程中，非专家注释者可以轻松地成功地攻击SOTA VQA模型。 (ii) 大规模预训练模型和对抗性训练方法在新基准上的性能远远低于标准VQA v2数据集，这表明了这些模型的脆弱性，同时也证明了我们对抗性数据集的有效性。 (iii) 当用于数据扩充时，我们的数据集可以在其他健壮的VQA基准上有效地提高模型性能。我们希望我们的对抗性VQA数据集能够为社区中的稳健性研究提供新的思路，并为未来的工作提供有价值的基准。

Unsupervised domain adaptation (DA) methods have focused on achieving maximal performance through aligning features from source and target domains without using labeled data in the target domain. Whereas, in the real-world scenario's it might be feasible to get labels for a small proportion of target data. In these scenarios, it is important to select maximally-informative samples to label and find an effective way to combine them with the existing knowledge from source data. Towards achieving this, we propose S<sup>3</sup>VAAADA which i) introduces a novel submodular criterion to select a maximally informative subset to label and ii) enhances a cluster-based DA procedure through novel improvements to effectively utilize all the available data for improving generalization on target. Our approach consistently outperforms the competing state-of-the-art approaches on datasets with varying degrees of domain shifts. The project page with additional details is available here: <https://sites.google.com/iisc.ac.in/s3vaada-iccv2021>.

无监督域自适应 (DA) 方法的重点是通过对齐源域和目标域的特征而不使用目标域中的标记数据来实现最大性能。然而，在现实场景中，为一小部分目标数据获取标签可能是可行的。在这些场景中，重要的是选择信息量最大的样本进行标记，并找到一种有效的方法将它们与源数据中的现有知识相结合。为了实现这一点，我们提出了S<sup>3</sup>VAAADA，其中i) 引入了一个新的子模块标准来选择一个信息量最大的子集进行标记，ii) 通过新的改进来增强基于集群的DA过程，以有效地利用所有可用的数据来改进目标泛化。我们的方法在具有不同程度域转移的数据集上始终优于竞争的最新方法。包含其他详细信息的项目页面可在以下位置获得：<https://sites.google.com/iisc.ac.in/s3vaada-iccv2021>。

Video activity localisation has recently attained increasing attention due to its practical values in automatically localising the most salient visual segments corresponding to their language descriptions (sentences) from untrimmed and unstructured videos. For supervised model training, a temporal annotation of both the start and end time index of each video segment for a sentence (a video moment) must be given. This is not only very expensive but also sensitive to ambiguity and subjective annotation bias, a much harder task than image labelling. In this work, we develop a more accurate weakly-supervised solution by introducing Cross-Sentence Relations Mining (CRM) in video moment proposal generation and matching when only a paragraph description of activities without per-sentence temporal annotation is available. Specifically, we explore two cross-sentence relational constraints: (1) Temporal ordering and (2) semantic consistency among sentences in a paragraph description of video activities. Existing weakly-supervised techniques only consider within-sentence video segment correlations in training without considering cross-sentence paragraph context. This can mislead due to ambiguous expressions of individual sentences with visually indiscriminate video moment proposals in isolation. Experiments on two publicly available activity localisation datasets show the advantages of our approach over the state-of-the-art weakly supervised methods, especially so when the video activity descriptions become more complex.

视频活动定位最近受到越来越多的关注，因为它在自动定位与未剪辑和非结构化视频中的语言描述（句子）相对应的最显著的视觉片段方面具有实用价值。对于监督模型训练，必须给出句子（视频时刻）每个视频段的开始和结束时间索引的时间注释。这不仅非常昂贵，而且对模糊性和主观注释偏差也很敏感，这比图像标记困难得多。在这项工作中，我们开发了一个更精确的弱监督解决方案，在视频矩建议生成和匹配中引入了跨句子关系挖掘（CRM），当只有活动的段落描述而没有句子时态注释时。具体来说，我们探讨了两个跨句子关系约束：（1）时间顺序和（2）视频活动段落描述中句子之间的语义一致性。现有的弱监督技术只考虑训练中的句子视频段相关性，而不考虑跨句段落上下文。这可能导致误导，因为单独使用视频片段建议时，个别句子的表达模棱两可。在两个公开的活动定位数据集上的实验表明，我们的方法比最先进的弱监督方法具有优势，尤其是当视频活动描述变得更加复杂时。

Point cloud registration is the process of using the common structures in two point clouds to splice them together. To find out these common structures and make these structures match more accurately, we investigate the direction of interacting information of the source and target point clouds. To this end, we propose a Feature Interactive Representation learning Network (FIRE-Net), which can explore feature interaction among the source and target point clouds from different levels. Specifically, we first introduce a Combined Feature Encoder (CFE) based on feature interaction intra point cloud. CFE extracts interactive features intra each point cloud and combines them to enhance the ability of the network to describe the local geometric structure. Then, we propose a feature interaction mechanism inter point clouds which includes a Local Interaction Unit (LIU) and a Global Interaction Unit (GIU). The former is used to interact information between point pairs across two point clouds, thus the point features in one point cloud and its similar point features in another point cloud can be aware of each other. The latter is applied to change the per-point features depending on the global cross information of two point clouds, thus one point cloud has the global perception of another. Extensive experiments on partially overlapping point cloud registration show that our method achieves state-of-the-art performance.

点云注册是使用两个点云中的公共结构将它们拼接在一起的过程。为了找出这些共同的结构并使这些结构更精确地匹配，我们研究了源点云和目标点云相互作用信息的方向。为此，我们提出了一个特征交互表示学习网络（FIRE Net），它可以从不同的层次探索源点云和目标点云之间的特征交互。具体来说，我们首先介绍了一种基于点云内特征交互的组合特征编码器（CFE）。CFE在每个点云内提取交互特征，

并将其组合以增强网络描述局部几何结构的能力。然后，我们提出了一种点云间的特征交互机制，该机制包括一个局部交互单元（LIU）和一个全局交互单元（GIU）。前者用于跨两个点云的点对之间的信息交互，因此一个点云中的点特征和另一个点云中的相似点特征可以相互感知。后者用于根据两个点云的全局交叉信息更改每点特征，因此一个点云具有另一个点云的全局感知。大量的部分重叠点云配准实验表明，我们的方法达到了最先进的性能。

Discrete point cloud objects lack sufficient shape descriptors of 3D geometries. In this paper, we present a novel method for aggregating hypothetical curves in point clouds. Sequences of connected points (curves) are initially grouped by taking guided walks in the point clouds, and then subsequently aggregated back to augment their point-wise features. We provide an effective implementation of the proposed aggregation strategy including a novel curve grouping operator followed by a curve aggregation operator. Our method was benchmarked on several point cloud analysis tasks where we achieved the state-of-the-art classification accuracy of 94.2% on the ModelNet40 classification task, instance IoU of 86.8% on the ShapeNetPart segmentation task and cosine error of 0.11 on the ModelNet40 normal estimation task.

离散点云对象缺少足够的三维几何图形形状描述符。在本文中，我们提出了一种在点云中聚合假设曲线的新方法。连接点（曲线）的序列最初通过在点云中进行引导行走进行分组，然后再聚合回来以增强其逐点特征。我们提供了一个有效的实现所提出的聚合策略，包括一个新的曲线分组操作符和一个曲线聚合操作符。我们的方法在多个点云分析任务上进行了基准测试，其中我们在ModelNet40分类任务上达到了94.2%的最先进分类精度，在ShapeNetPart分割任务上达到了86.8%的实例IoU，在ModelNet40正常估计任务上达到了0.11的余弦误差。

We consider using untrained neural networks to solve the reconstruction problem of snapshot compressive imaging (SCI), which uses a two-dimensional (2D) detector to capture a high-dimensional (usually 3D) data-cube in a compressed manner. Various SCI systems have been built in recent years to capture data such as high-speed videos, hyperspectral images, and the state-of-the-art reconstruction is obtained by the deep neural networks. However, most of these networks are trained in an end-to-end manner by a large amount of corpus with sometimes simulated ground truth, measurement pairs. In this paper, inspired by the untrained neural networks such as deep image priors (DIP) and deep decoders, we develop a framework by integrating DIP into the plug-and-play regime, leading to a self-supervised network for spectral SCI reconstruction. Extensive synthetic and real data results show that the proposed algorithm without training is capable of achieving competitive results to the training based networks. Furthermore, by integrating the proposed method with a pre-trained deep denoising prior, we have achieved higher performance than existing state-of-the-art.

我们考虑使用未经训练的神经网络来解决快照压缩成像（SCI）的重建问题，其使用二维（2D）检测器以压缩方式捕获高维（通常为3D）数据立方体。近年来，人们建立了各种SCI系统来捕获高速视频、高光谱图像等数据，并通过深度神经网络实现了最先进的重建。然而，这些网络中的大多数都是由大量语料库以端到端的方式进行训练的，这些语料库有时带有模拟的地面真相、测量对。在本文中，受未经训练的神经网络（如深度图像先验（DIP）和深度解码器）的启发，我们开发了一个框架，将DIP集成到即插即用模式中，从而形成一个用于光谱SCI重建的自监督网络。大量的合成和实际数据结果表明，该算法在不进行训练的情况下，能够获得与基于训练的网络相媲美的结果。此外，通过将所提出的方法与预先训练好的深度去噪先验知识相结合，我们获得了比现有技术更高的性能。

Assessing action quality is challenging due to the subtle differences between videos and large variations in scores. Most existing approaches tackle this problem by regressing a quality score from a single video, suffering a lot from the large inter-video score variations. In this paper, we show that the relations among videos can provide important clues for more accurate action quality assessment during both training and inference. Specifically, we reformulate the problem of action quality assessment as regressing the relative scores with reference to another video that has shared attributes (e.g. category and difficulty), instead of learning unreference scores. Following this formulation, we propose a new contrastive regression (CoRe) framework to learn the relative scores by pair-wise comparison, which highlights the differences between videos and guides the models to learn the key hints for assessment. In order to further exploit the relative information between two videos, we devise a group-aware regression tree to convert the conventional score regression into two easier sub-problems: coarse-to-fine classification and regression in small intervals. To demonstrate the effectiveness of CoRe, we conduct extensive experiments on three mainstream AQA datasets including AQA-7, MTL-AQA, and JIGSAWS. Our approaches outperform previous methods by a large margin and establish new state-of-the-art on all three benchmarks.

由于视频之间的细微差异和分数的巨大差异，评估动作质量具有挑战性。大多数现有的方法都是通过从单个视频中回归质量分数来解决这个问题，因为视频间的分数变化很大。在本文中，我们发现视频之间的关系可以为训练和推理过程中更准确的动作质量评估提供重要线索。具体而言，我们将动作质量评估问题重新表述为参考另一个具有共同属性（例如类别和难度）的视频回归相对分数，而不是学习未参考分数。根据这个公式，我们提出了一个新的对比回归（CoRe）框架，通过成对比较来学习相对分数，该框架突出了视频之间的差异，并指导模型学习评估的关键提示。为了进一步利用两个视频之间的相关信息，我们设计了一个群体感知回归树，将传统的分数回归转化为两个更简单的子问题：从粗到细的分类和小间隔回归。为了证明CoRe的有效性，我们在三个主流AQA数据集上进行了广泛的实验，包括AQA-7、MTL-AQA和JIGSAWS。我们的方法大大优于以前的方法，并在所有三个基准上建立了新的最先进水平。

Vision-and-Language Navigation (VLN) requires an agent to find a path to a remote location on the basis of natural-language instructions and a set of photo-realistic panoramas. Most existing methods take the words in the instructions and the discrete views of each panorama as the minimal unit of encoding. However, this requires a model to match different nouns (e.g., TV, table) against the same input view feature. In this work, we propose an object-informed sequential BERT to encode visual perceptions and linguistic instructions at the same fine-grained level, namely objects and words. Our sequential BERT also enables the visual-textual clues to be interpreted in light of the temporal context, which is crucial to multi-round VLN tasks. Additionally, we enable the model to identify the relative direction (e.g., left/right/front/back) of each navigable location and the room type (e.g., bedroom, kitchen) of its current and final navigation goal, as such information is widely mentioned in instructions implying the desired next and final locations. We thus enable the model to know-where the objects lie in the images, and to know-where they stand in the scene. Extensive experiments demonstrate the effectiveness compared against several state-of-the-art methods on three indoor VLN tasks: REVERIE, NDH, and R2R. Project repository: <https://github.com/YuankaiQi/ORIST>

视觉和语言导航（VLN）需要一个代理根据自然语言指令和一组照片逼真的全景图查找到远程位置的路径。大多数现有方法将指令中的单词和每个全景图的离散视图作为最小编码单元。但是，这需要一个模型将不同的名词（例如，TV、表格）与相同的输入视图功能相匹配。在这项工作中，我们提出了一种对象通知的顺序伯特编码视觉感知和语言指令在同一细粒度的水平，即对象和单词。我们的顺序BERT还可以根据时间上下文解释视觉文本线索，这对于多轮VLN任务至关重要。此外，我们使模型能够识别每个

可导航位置的相对方向（例如，左/右/前/后）及其当前和最终导航目标的房间类型（例如，卧室、厨房），因为此类信息在说明中广泛提及，暗示了所需的下一个和最终位置。因此，我们使模型能够知道对象在图像中的位置，并知道它们在场景中的位置。大量的实验表明，与几种最先进的方法相比，该方法在三种室内VLN任务（幻想、NDH和R2R）上的有效性。项目存储库：<https://github.com/YuankaiQi/ORIST>

In this paper, by modeling the point cloud registration task as a Markov decision process, we propose an end-to-end deep model embedded with the cross-entropy method (CEM) for unsupervised 3D registration. Our model consists of a sampling network module and a differentiable CEM module. In our sampling network module, given a pair of point clouds, the sampling network learns a prior sampling distribution over the transformation space. The learned sampling distribution can be used as a "good"" initialization of the differentiable CEM module. In our differentiable CEM module, we first propose a maximum consensus criterion based alignment metric as the reward function for the point cloud registration task. Based on the reward function, for each state, we then construct a fused score function to evaluate the sampled transformations, where we weight the current and future rewards of the transformations. Particularly, the future rewards of the sampled transforms are obtained by performing the iterative closest point (ICP) algorithm on the transformed state. By selecting the top-k transformations with the highest scores, we iteratively update the sampling distribution. Furthermore, in order to make the CEM differentiable, we use the sparsemax function to replace the hard top-k selection. Finally, we formulate a Geman-McClure estimator based loss to train our end-to-end registration model. Extensive experimental results demonstrate the good registration performance of our method on benchmark datasets.

在本文中，通过将点云配准任务建模为一个马尔可夫决策过程，我们提出了一种嵌入交叉熵方法(CEM)的端到端深度模型，用于无监督三维配准。我们的模型由一个采样网络模块和一个可微CEM模块组成。在我们的采样网络模块中，给定一对点云，采样网络学习变换空间上的先验采样分布。学到的采样分布可用作“好的”可微CEM模块的初始化。在我们的可微CEM模块中，我们首先提出一个基于最大一致性标准的对齐度量作为点云注册任务的奖励函数。基于奖励函数，对于每个状态，我们然后构造一个融合分数函数来评估采样的变换，其中我们权衡当前和未来的转型回报。特别地，通过对变换状态执行迭代最近点(ICP)算法来获得采样变换的未来回报。通过选择得分最高的top-k变换，我们迭代地更新采样分布。此外，为了使CEM可微，我们使用sparsemax函数来代替硬top-k选择。最后，我们提出了一个基于Geman-McClure估计的损失来训练我们的端到端注册模型。大量的实验结果表明，该方法在基准数据集上具有良好的配准性能。

Most recent approaches for online action detection tend to apply Recurrent Neural Network (RNN) to capture long-range temporal structure. However, RNN suffers from non-parallelism and gradient vanishing, hence it is hard to be optimized. In this paper, we propose a new encoder-decoder framework based on Transformers, named OadTR, to tackle these problems. The encoder attached with a task token aims to capture the relationships and global interactions between historical observations. The decoder extracts auxiliary information by aggregating anticipated future clip representations. Therefore, OadTR can recognize current actions by encoding historical information and predicting future context simultaneously. We extensively evaluate the proposed OadTR on three challenging datasets: HDD, TVSeries, and THUMOS14. The experimental results show that OadTR achieves higher training and inference speeds than current RNN based approaches, and significantly outperforms the state-of-the-art methods in terms of both mAP and mcAP. Code is available at <https://github.com/wangxiang1230/OadTR>.

最新的在线行为检测方法倾向于使用递归神经网络（RNN）来捕获长程时间结构。然而，RNN存在非并行性和梯度消失问题，因此很难进行优化。在本文中，我们提出了一种新的基于变压器的编解码框架OadTR来解决这些问题。带有任务标记的编码器旨在捕获历史观测之间的关系和全局交互。解码器通过聚合预期的未来片段表示来提取辅助信息。因此，OadTR可以通过同时编码历史信息和预测未来上下文来识别当前行为。我们在三个具有挑战性的数据集（HDD、TVSeries和THUMOS14）上广泛评估了提议的OadTR。实验结果表明，OadTR比现有的基于RNN的方法具有更高的训练和推理速度，并且在mAP和mcAP方面显著优于最新的方法。代码位于<https://github.com/wangsiang1230/OadTR>。

Detection and segmentation of nuclei are fundamental analysis operations in pathology images, the assessments derived from which serve as the gold standard for cancer diagnosis. Manual segmenting nuclei is expensive and time-consuming. what's more, accurate segmentation detection of nuclei can be challenging due to the large appearance variation, conjoined and overlapping nuclei, and serious degeneration of histological structures. Supervised methods highly rely on massive annotated samples. The existing two unsupervised methods are prone to failure on degenerated samples. This paper proposes a Mutual-Complementing Framework (MCF) for nuclei detection and segmentation in pathology images. Two branches of MCF are trained in the mutual-complementing manner, where the detection branch complements the pseudo mask of the segmentation branch, while the progressive trained segmentation branch complements the missing nucleus templates through calculating the mask residual between the predicted mask and detected result. In the detection branch, two response map fusion strategies and gradient direction based postprocessing are devised to obtain the optimal detection response. Furthermore, the confidence loss combined with the synthetic samples and self-finetuning is adopted to train the segmentation network with only high confidence areas. Extensive experiments demonstrate that MCF achieves comparable performance with only a few nucleus patches as supervision. Especially, MCF possesses good robustness (only dropping by about 6%) on degenerated samples, which are critical and common cases in clinical diagnosis.

细胞核的检测和分割是病理图像中的基本分析操作，由此产生的评估是癌症诊断的金标准。人工分割细胞核既昂贵又耗时。更重要的是，由于大的外观变化、连体和重叠的细胞核以及组织结构的严重退化，细胞核的精确分割检测可能是一个挑战。监督方法高度依赖于大量带注释的样本。现有的两种无监督方法在退化样本上容易失效。提出了一种用于病理图像细胞核检测和分割的互补框架（MCF）。MCF的两个分支以互补的方式进行训练，其中检测分支对分割分支的伪掩模进行互补，而渐进训练的分割分支通过计算预测掩模与检测结果之间的掩模残差对缺失的核模板进行互补。在检测分支中，设计了两种响应图融合策略和基于梯度方向的后处理，以获得最佳的检测响应。此外，采用置信度损失结合合成样本和自微调的方法来训练只有高置信区域的分割网络。大量的实验表明，MCF只需几个nucleus补丁作为监控，即可获得相当的性能。特别是，MCF对退化样本具有良好的鲁棒性（仅下降约6%），这是临床诊断中的关键和常见情况。

Recently, transfer subspace learning based approaches have shown to be a valid alternative to unsupervised subspace clustering and temporal data clustering for human motion segmentation (HMS). These approaches leverage prior knowledge from a source domain to improve clustering performance on a target domain, and currently they represent the state of the art in HMS. Bucking this trend, in this paper, we propose a novel unsupervised model that learns a representation of the data and digs clustering information from the data itself. Our model is reminiscent of temporal subspace clustering, but presents two critical differences. First, we learn an auxiliary data matrix that can deviate from the initial data, hence confers more degrees of freedom to the coding matrix. Second, we introduce a regularization term for this auxiliary data matrix that preserves the local geometrical structure present in the high-dimensional space. The proposed model is efficiently optimized by using an original Alternating Direction Method of Multipliers (ADMM) formulation allowing to learn jointly the auxiliary data representation, a nonnegative dictionary and a coding matrix. Experimental results on four benchmark datasets for HMS demonstrate that our approach achieves significantly better clustering performance than state-of-the-art methods, including both unsupervised and more recent semi-supervised transfer learning approaches.

近年来，基于转移子空间学习的方法已被证明是无监督子空间聚类和时间数据聚类用于人体运动分割 (HMS) 的有效替代方法。这些方法利用来自源域的先验知识来提高目标域上的集群性能，目前它们代表了HMS的最新技术。与此相反，在本文中，我们提出了一种新的无监督模型，该模型学习数据的表示，并从数据本身挖掘聚类信息。我们的模型让人联想到时间子空间聚类，但有两个关键区别。首先，我们学习一个辅助数据矩阵，它可以偏离初始数据，从而赋予编码矩阵更多的自由度。其次，我们为这个辅助数据矩阵引入一个正则化项，它保持了高维空间中存在的局部几何结构。该模型通过使用原始交替方向乘子法 (ADMM) 公式进行有效优化，允许联合学习辅助数据表示、非负字典和编码矩阵。在四个HMS基准数据集上的实验结果表明，我们的方法比最先进的方法（包括无监督和最近的半监督转移学习方法）取得了显著更好的聚类性能。

While recent studies on pedestrian attribute recognition have shown remarkable progress in leveraging complicated networks and attention mechanisms, most of them neglect the inter-image relations and an important prior: spatial consistency and semantic consistency of attributes under surveillance scenarios. The spatial locations of the same attribute should be consistent between different pedestrian images, e.g., the "hat" attribute and the "boots" attribute are always located at the top and bottom of the picture respectively. In addition, the inherent semantic feature of the "hat" attribute should be consistent, whether it is a baseball cap, beret, or helmet. To fully exploit inter-image relations and aggregate human prior in the model learning process, we construct a Spatial and Semantic Consistency (SSC) framework that consists of two complementary regularizations to achieve spatial and semantic consistency for each attribute. Specifically, we first propose a spatial consistency regularization to focus on reliable and stable attribute-related regions. Based on the precise attribute locations, we further propose a semantic consistency regularization to extract intrinsic and discriminative semantic features. We conduct extensive experiments on popular benchmarks including PA100K, RAP, and PETA. Results show that the proposed method performs favorably against state-of-the-art methods without increasing parameters.

虽然最近关于行人属性识别的研究表明，在利用复杂的网络和注意机制方面取得了显著的进展，但大多数研究忽略了图像间的关系和一个重要的优先事项：监视场景下属性的空间一致性和语义一致性。相同属性的空间位置应在不同的行人图像之间保持一致，例如，“帽子”属性和“靴子”属性始终分别位于图片的顶部和底部。此外，“帽子”属性的固有语义特征应该是一致的，无论是棒球帽、贝雷帽还是头盔。为了在模型学习过程中充分利用图像间的关系并聚合人类先验知识，我们构建了一个由两个互补正则化组成

的空间和语义一致性 (SSC) 框架，以实现每个属性的空间和语义一致性。具体来说，我们首先提出了一种空间一致性正则化方法，以关注可靠和稳定的属性相关区域。基于精确的属性位置，我们进一步提出了一种语义一致性正则化方法来提取内在的和有区别的语义特征。我们在流行的基准上进行了广泛的实验，包括PA100K、RAP和PETA。结果表明，在不增加参数的情况下，该方法优于现有的方法。

Many emerging applications of intelligent robots need to explore and understand new environments, where it is desirable to detect objects of novel categories on the fly with minimum online efforts. This is an object detection on demand (ODOD) task. It is challenging, because it is impossible to annotate large data on the fly, and the embedded systems are usually unable to perform back-propagation which is essential for training. Most existing few-shot detection methods are confronted here as they need extra training. We propose a novel morphable detector (MD), that simply "morphs" some of its changeable parameters online estimated from the few samples, so as to detect novel categories without any extra training. The MD has two sets of parameters, one for the feature embedding and the other for category representation (called "prototypes"). Each category is associated with a hidden prototype to be learned by integrating the visual and semantic embeddings. The learning of the MD is based on the alternate learning of the feature embedding and the prototypes in an EM-like approach which allows the recovery of an unknown prototype from a few samples of a novel category. Once an MD is learned, it is able to use a few samples of a novel category to directly compute its prototype to fulfill the online morphing process. We have shown the superiority of the MD in Pascal, COCO and FSOD datasets.

智能机器人的许多新兴应用需要探索和理解新的环境，在这些环境中，人们希望用最少的在线努力在飞行中检测新类别的物体。这是一项按需对象检测 (ODOD) 任务。这是一个挑战，因为不可能动态地对大数据进行注释，并且嵌入式系统通常无法执行训练所必需的反向传播。大多数现有的少数镜头检测方法都需要额外的培训，因此在这里面临。我们提出了一种新的可变形检测器 (MD)，该检测器简单地“变形”从少数样本在线估计的一些可变参数，以便在不进行任何额外训练的情况下检测新类别。MD有两组参数，一组用于特征嵌入，另一组用于类别表示（称为“原型”）。每个类别都与一个隐藏的原型相关联，通过集成视觉和语义嵌入来学习。MD的学习基于特征嵌入和原型的交替学习，采用类似EM的方法，允许从新类别的几个样本中恢复未知原型。一旦学习了MD，它就能够使用一个新类别的几个样本直接计算其原型来完成在线变形过程。我们已经在Pascal、COCO和FSOD数据集中展示了MD的优越性。

Multi-view pedestrian detection aims to predict a bird's eye view (BEV) occupancy map from multiple camera views. This task is confronted with two challenges: how to establish the 3D correspondences from views to the BEV map and how to assemble occupancy information across views. In this paper, we propose a novel stacked Homography Transformations (SHOT) approach, which is motivated by approximating projections in 3D world coordinates via a stack of homographies. We first construct a stack of transformations for projecting views to the ground plane at different height levels. Then we design a soft selection module so that the network learns to predict the likelihood of the stack of transformations. Moreover, we provide an in-depth theoretical analysis on constructing SHOT and how well SHOT approximates projections in 3D world coordinates. SHOT is empirically verified to be capable of estimating accurate correspondences from individual views to the BEV map, leading to new state-of-the-art performance on standard evaluation benchmarks.

多视角行人检测旨在从多个摄像机视角预测鸟瞰图 (BEV) 占用地图。这项任务面临两个挑战：如何建立从视图到BEV地图的3D对应关系，以及如何跨视图收集占用信息。在本文中，我们提出了一种新的叠加单应变换 (SHOT) 方法，其动机是通过叠加单应来近似3D世界坐标中的投影。我们首先构造一组变换，用于将视图投影到不同高度级别的地平面。然后，我们设计了一个软选择模块，以便网络学习预测

变换堆栈的可能性。此外，我们提供了一个深入的理论分析构造镜头和如何以及镜头近似投影在三维世界坐标。通过经验验证，SHOT能够估计从单个视图到BEV地图的准确对应关系，从而在标准评估基准上实现新的最先进性能。

Visual understanding goes well beyond the study of images or videos on the web. To achieve complex tasks in volatile situations, the human can deeply understand the environment, quickly perceive events happening around, and continuously track objects' state changes, which are still challenging for current AI systems. To equip AI system with the ability to understand dynamic environments, we build a video Question Answering dataset named Env-QA. Env-QA contains 23K egocentric videos, where each video is composed of a series of events about exploring and interacting in the environment. It also provides 85K questions to evaluate the ability of understanding the composition, layout, and state changes of the environment presented by the events in videos. Moreover, we propose a video QA model, Temporal Segmentation and Event Attention network (TSEA), which introduces event-level video representation and corresponding attention mechanisms to better extract environment information and answer questions. Comprehensive experiments demonstrate the effectiveness of our framework and show the formidable challenges of Env-QA in terms of long-term state tracking, multi-event temporal reasoning and event counting, etc.

视觉理解远远超出了对网络上图像或视频的研究。为了在多变的环境中完成复杂的任务，人类可以深入了解环境，快速感知周围发生的事件，并持续跟踪对象的状态变化，这对于当前的人工智能系统来说仍然是一个挑战。为了使人工智能系统具备理解动态环境的能力，我们构建了一个名为Env-QA的视频问答数据集。Env QA包含23K个以自我为中心的视频，每个视频都由一系列关于在环境中探索和互动的事件组成。它还提供85K个问题，以评估理解视频中事件所呈现环境的组成、布局和状态变化的能力。此外，我们还提出了一个视频质量保证模型——时间分割和事件注意网络（TSEA），该模型引入了事件级视频表示和相应的注意机制，以更好地提取环境信息和回答问题。综合实验证明了该框架的有效性，并显示了Env-QA在长期状态跟踪、多事件时态推理和事件计数等方面面临的巨大挑战。

We extend the task of composed image retrieval, where an input query consists of an image and short textual description of how to modify the image. Existing methods have only been applied to non-complex images within narrow domains, such as fashion products, thereby limiting the scope of study on in-depth visual reasoning in rich image and language contexts. To address this issue, we collect the Compose Image Retrieval on Real-life images (CIRR) dataset, which consists of over 36,000 pairs of crowd-sourced, open-domain images with human-generated modifying text. To extend current methods to the open-domain, we propose CIRPLANT, a transformer based model that leverages rich pre-trained vision-and-language (V&L) knowledge for modifying visual features conditioned on natural language. Retrieval is then done by nearest neighbor lookup on the modified features. We demonstrate that with a relatively simple architecture, CIRPLANT outperforms existing methods on open-domain images, while matching state-of-the-art accuracy on the existing narrow datasets, such as fashion. Together with the release of CIRR, we believe this work will inspire further research on composed image retrieval. Our dataset, code and pre-trained models are available at <https://cuberick-orion.github.io/CIRR/>.

我们扩展了合成图像检索的任务，其中输入查询由图像和如何修改图像的简短文本描述组成。现有方法仅适用于狭窄领域内的非复杂图像，如时装产品，从而限制了丰富图像和语言背景下深入视觉推理的研究范围。为了解决这个问题，我们收集了真实图像上的合成图像检索（CIRR）数据集，该数据集由36000多对众包、开放域图像和人工生成的修改文本组成。为了将当前的方法扩展到开放领域，我们提出了CIRPLANT，这是一种基于转换器的模型，它利用丰富的预先训练的视觉和语言（V&L）知识来修改以自然语言为条件的视觉特征。然后通过最近邻查找修改后的特征来进行检索。我们证明，通过相对简单的体系结构，CIRPLANT在开放域图像上优于现有方法，同时在现有狭窄数据集（如fashion）上匹配

最先进的精度。随着CIRR的发布，我们相信这项工作将激发合成图像检索的进一步研究。我们的数据集、代码和预先训练的模型可在<https://cuberick-orion.github.io/CIRR/>.

Transferability of adversarial examples is of central importance for attacking an unknown model, which facilitates adversarial attacks in more practical scenarios, e.g., blackbox attacks. Existing transferable attacks tend to craft adversarial examples by indiscriminately distorting features to degrade prediction accuracy in a source model without aware of intrinsic features of objects in the images. We argue that such brute-force degradation would introduce model-specific local optimum into adversarial examples, thus limiting the transferability. By contrast, we propose the Feature Importance-aware Attack (FIA), which disrupts important object-aware features that dominate model decisions consistently. More specifically, we obtain feature importance by introducing the aggregate gradient, which averages the gradients with respect to feature maps of the source model, computed on a batch of random transforms of the original clean image. The gradients will be highly correlated to objects of interest, and such correlation presents invariance across different models. Besides, the random transforms will preserve intrinsic features of objects and suppress model-specific information. Finally, the feature importance guides to search for adversarial examples towards disrupting critical features, achieving stronger transferability. Extensive experimental evaluation demonstrates the effectiveness and superior performance of the proposed FIA, i.e., improving the success rate by 9.5% against normally trained models and 12.8% against defense models as compared to the state-of-the-art transferable attacks. Code is available at: <https://github.com/hcguo00/FIA>

对抗性示例的可转移性对于攻击未知模型至关重要，这有助于在更实际的场景中进行对抗性攻击，例如黑盒攻击。现有的可转移攻击倾向于通过不加区别地扭曲特征来制造对抗性示例，从而降低源模型中的预测精度，而不知道图像中对象的固有特征。我们认为，这种蛮力退化会在对抗性示例中引入特定于模型的局部最优，从而限制可转移性。相比之下，我们提出了特征重要性感知攻击（FIA），它破坏了一致支配模型决策的重要对象感知特征。更具体地说，我们通过引入聚合梯度来获得特征重要性，聚合梯度是在原始干净图像的一批随机变换上计算的相对于源模型的特征映射的梯度的平均值。梯度将与感兴趣的对象高度相关，并且这种相关性在不同的模型中呈现不变性。此外，随机变换将保留对象的固有特征并抑制特定于模型的信息。最后，功能重要性指导搜索对抗性示例，以破坏关键功能，实现更强的可转移性。广泛的实验评估证明了所提出的FIA的有效性和优越性能，即与最先进的可转移攻击相比，与正常训练模型相比，成功率提高了9.5%，与防御模型相比，成功率提高了12.8%。代码可从以下网址获取：<https://github.com/hcguo00/FIA>

The conventional detectors tend to make imbalanced classification and suffer performance drop, when the distribution of the training data is severely skewed. In this paper, we propose to use the mean classification score to indicate the classification accuracy for each category during training. Based on this indicator, we balance the classification via an Equilibrium Loss (EBL) and a Memory-augmented Feature Sampling (MFS) method. Specifically, EBL increases the intensity of the adjustment of the decision boundary for the weak classes by a designed score-guided loss margin between any two classes. On the other hand, MFS improves the frequency and accuracy of the adjustments of the decision boundary for the weak classes through over-sampling the instance features of those classes. Therefore, EBL and MFS work collaboratively for finding the classification equilibrium in long-tailed detection, and dramatically improve the performance of tail classes while maintaining or even improving the performance of head classes. We conduct experiments on LVIS using Mask R-CNN with various backbones including ResNet-50-FPN and ResNet-101-FPN to show the superiority of the proposed method. It improves the detection performance of tail classes by 15.6 AP, and outperforms the most recent long-tailed object detectors by more than 1 AP. Code is available at <https://github.com/fcjian/LOCE>.

当训练数据分布严重偏斜时，传统的检测器往往会导致分类不平衡，性能下降。在本文中，我们建议使用平均分类分数来表示训练期间每个类别的分类精度。基于这个指标，我们通过平衡损失（EBL）和记忆增强特征抽样（MFS）方法来平衡分类。具体而言，EBL通过在任意两个类别之间设计分数引导损失裕度，增加了弱类别决策边界调整的强度。另一方面，MFS通过对弱类的实例特征进行过采样，提高了弱类决策边界调整的频率和精度。因此，EBL和MFS协同工作，在长尾检测中找到分类均衡，并显著提高尾部类的性能，同时保持甚至提高头部类的性能。我们使用Mask R-CNN和ResNet-50-FPN、ResNet-101-FPN等多种主干在LVIS上进行了实验，证明了该方法的优越性。它将尾部类的检测性能提高了15.6 AP，比最新的长尾目标检测器提高了1 AP以上。代码可在<https://github.com/fcjian/LOCE>。

In recent years, research on adversarial attacks has become a hot spot. Although current literature on the transfer-based adversarial attack has achieved promising results for improving the transferability to unseen black-box models, it still leaves a long way to go. Inspired by the idea of meta-learning, this paper proposes a novel architecture called Meta Gradient Adversarial Attack (MGAA), which is plug-and-play and can be integrated with any existing gradient-based attack method for improving the cross-model transferability. Specifically, we randomly sample multiple models from a model zoo to compose different tasks and iteratively simulate a white-box attack and a black-box attack in each task. By narrowing the gap between the gradient directions in white-box and black-box attacks, the transferability of adversarial examples on the black-box setting can be improved. Extensive experiments on the CIFAR10 and ImageNet datasets show that our architecture outperforms the state-of-the-art methods for both black-box and white-box attack settings.

近年来，对抗性攻击的研究成为一个热点。尽管当前关于基于转移的对抗性攻击的文献在提高对不可见黑盒模型的可转移性方面取得了有希望的结果，但它仍有很长的路要走。受元学习思想的启发，本文提出了一种新的体系结构，称为元梯度对抗攻击（MGAA），它是即插即用的，可以与任何现有的基于梯度的攻击方法集成，以提高跨模型的可转移性。具体来说，我们从一个模型动物园随机抽取多个模型来组成不同的任务，并在每个任务中迭代模拟白盒攻击和黑盒攻击。通过缩小白盒攻击和黑盒攻击中梯度方向之间的差距，可以提高黑盒设置上对抗性示例的可转移性。在CIFAR10和ImageNet数据集上进行的大量实验表明，对于黑盒和白盒攻击设置，我们的体系结构都优于最先进的方法。

Exploiting convolutional neural networks for point cloud processing is quite challenging, due to the inherent irregular distribution and discrete shape representation of point clouds. To address these problems, many handcrafted convolution variants have sprung up in recent years. Though with elaborate design, these variants could be far from optimal in sufficiently capturing diverse shapes formed by discrete points. In this paper, we propose PointSeaConv, i.e., a novel differential convolution search paradigm on point clouds. It can work in a purely data-driven manner and thus is capable of auto-creating a group of suitable convolutions for geometric shape modeling. We also propose a joint optimization framework for simultaneous search of internal convolution and external architecture, and introduce epsilon-greedy algorithm to alleviate the effect of discretization error. As a result, PointSeaNet, a deep network that is sufficient to capture geometric shapes at both convolution level and architecture level, can be searched out for point cloud processing. Extensive experiments strongly evidence that our proposed PointSeaNet surpasses current handcrafted deep models on challenging benchmarks across multiple tasks with remarkable margins.

由于点云固有的不规则分布和离散形状表示，利用卷积神经网络进行点云处理具有很大的挑战性。为了解决这些问题，近年来出现了许多手工制作的卷积变体。尽管设计精巧，但这些变体在充分捕捉离散点形成的各种形状方面可能远远不够理想。在本文中，我们提出了PointSeaConv，即一种新的点云微分卷积搜索范式。它可以以纯数据驱动的方式工作，因此能够为几何形状建模自动创建一组合适的卷积。我

们还提出了一个内部卷积和外部结构同时搜索的联合优化框架，并引入了epsilon贪婪算法来缓解离散化误差的影响。因此，PointSeaNet是一个深度网络，足以在卷积级别和体系结构级别捕获几何形状，可用于搜索点云处理。大量的实验有力地证明，我们提出的PointSeaNet在跨多个任务的具有挑战性的基准测试上超越了当前手工制作的深度模型，具有显著的优势。

we explore the zero-shot setting for day-night domain adaptation. The traditional domain adaptation setting is to train on one domain and adapt to the target domain by exploiting unlabeled data samples from the test set. As gathering relevant test data is expensive and sometimes even impossible, we do not rely on test data and instead exploit a visual inductive prior derived from physics-based reflection models for domain adaptation. We cast a number of color invariant edge detectors as trainable layers in a convolutional neural network and evaluate their robustness to illumination changes. We show that the color invariant layer reduces the day-night distribution shift in feature map activations throughout the network. we demonstrate improved performance for zero-shot day to night domain adaptation on both synthetic as well as natural datasets in various tasks, including classification, segmentation and place recognition.

我们探索昼夜域自适应的零镜头设置。传统的域自适应设置是在一个域上进行训练，并通过利用测试集中未标记的数据样本来适应目标域。由于收集相关测试数据非常昂贵，有时甚至是不可能的，因此我们不依赖测试数据，而是利用基于物理的反射模型得出的视觉归纳先验来进行域自适应。我们将许多颜色不变的边缘检测器作为卷积神经网络中的可训练层，并评估它们对光照变化的鲁棒性。我们表明，颜色不变层减少了整个网络中特征地图激活的昼夜分布偏移。我们在各种任务（包括分类、分割和位置识别）中展示了在合成数据集和自然数据集上改进的零镜头昼夜域自适应性能。

when capturing panoramas, people tend to align their cameras with the vertical axis, i.e., the direction of gravity. Moreover, modern devices, e.g. smartphones and tablets, are equipped with an IMU (Inertial Measurement Unit) that can measure the gravity vector accurately. Using this prior, the y-axes of the cameras can be aligned or assumed to be already aligned, reducing the relative orientation to 1-DOF (degree of freedom). Exploiting this assumption, we propose new minimal solutions to panoramic stitching of images taken by cameras with coinciding optical centers, i.e. undergoing pure rotation. we consider six practical camera configurations, from fully calibrated ones up to a camera with unknown fixed or varying focal length and with or without radial distortion. The solvers are tested both on synthetic scenes, on more than 500k real image pairs from the Sun360 dataset, and from scenes captured by us using two smartphones equipped with IMUS. The new solvers have similar or better accuracy than the state-of-the-art ones and outperform them in terms of processing time.

拍摄全景时，人们倾向于将相机与垂直轴对齐，即重力方向。此外，智能手机和平板电脑等现代设备都配备了IMU（惯性测量单元），可以精确测量重力矢量。使用此先验知识，可以对齐或假设相机的y轴已经对齐，从而将相对方向减少到1-DOF（自由度）。利用这一假设，我们提出了新的最小解决方案，用于具有重合光学中心（即经历纯旋转）的相机拍摄的图像的全景拼接。我们考虑六个实际的相机配置，从完全校准的相机到具有未知的固定或变化焦距和有或没有径向畸变。解算器在合成场景、Sun360数据集中超过500k的真实图像对以及我们使用配备IMU的两部智能手机拍摄的场景上进行了测试。新的解算器与最先进的解算器具有相似或更好的精度，并且在处理时间方面优于它们。

In this paper, we introduce a novel self-supervised visual representation learning method which understands both images and videos in a joint learning fashion. The proposed neural network architecture and objectives are designed to obtain two different Convolutional Neural Networks for solving visual recognition tasks in the domain of videos and images. Our method called Video/Image for Visual Contrastive Learning of Representation(Vi2CLR) uses unlabeled videos to exploit dynamic and static visual cues for self-supervised and instances similarity/dissimilarity learning. Vi2CLR optimization pipeline consists of visual clustering part and representation learning based on groups of similar positive instances within a cluster and negative ones from other clusters and learning visual clusters and their distances. We show how a joint self-supervised visual clustering and instance similarity learning with 2D (image) and 3D (video) CovNet encoders yields such robust and near to supervised learning performance. We extensively evaluate the method on downstream tasks like large scale action recognition and image and object classification on datasets like Kinetics, ImageNet, Pascal VOC'07 and UCF101 and achieve outstanding results compared to state-of-the-art self-supervised methods. To the best of our knowledge, the Vi2CLR is the first of its kind self-supervised neural network to tackle both video and image recognition task simultaneously by only using one source of data.

在本文中，我们介绍了一种新的自监督视觉表征学习方法，它以联合学习的方式理解图像和视频。提出的神经网络结构和目标旨在获得两种不同的卷积神经网络，用于解决视频和图像领域的视觉识别任务。我们的方法称为视频/图像表示的视觉对比学习（Vi2CLR），使用未标记的视频利用动态和静态视觉线索进行自我监督和实例相似性/差异性学习。Vi2CLR优化管道包括视觉聚类部分和基于集群内相似正实例组和来自其他集群的负实例组的表示学习，以及学习视觉集群及其距离。我们展示了如何使用2D（图像）和3D（视频）CovNet编码器进行联合自监督视觉聚类和实例相似性学习，从而产生如此强健和接近监督的学习性能。我们对下游任务（如大规模动作识别）以及数据集（如Kinetics、ImageNet、Pascal VOC'07和UCF101）上的图像和对象分类广泛评估了该方法，并与最先进的自我监督方法相比取得了优异的结果。据我们所知，Vi2CLR是第一个通过使用一个数据源同时处理视频和图像识别任务的自监督神经网络。

In this paper, we investigate the knowledge distillation (KD) strategy for object detection and propose an effective framework applicable to both homogeneous and heterogeneous student-teacher pairs. The conventional feature imitation paradigm introduces imitation masks to focus on informative foreground areas while excluding the background noises. However, we find that those methods fail to fully utilize the semantic information in all feature pyramid levels, which leads to inefficiency for knowledge distillation between FPN-based detectors. To this end, we propose a novel semantic-guided feature imitation technique, which automatically performs soft matching between feature pairs across all pyramid levels to provide the optimal guidance to the student. To push the envelop even further, we introduce contrastive distillation to effectively capture the information encoded in the relationship between different feature regions. Finally, we propose a generalized detection KD pipeline, which is capable of distilling both homogeneous and heterogeneous detector pairs. Our method consistently outperforms the existing detection KD techniques, and works when (1) components in the framework are used separately and in conjunction; (2) for both homogeneous and heterogeneous student-teacher pairs and (3) on multiple detection benchmarks. With a powerful x101-FasterRCNN-Instaboost detector as the teacher, R50-FasterRCNN reaches 44.0% AP, R50-RetinaNet reaches 43.3% AP and R50-FCOS reaches 43.1% AP on COCO dataset.

在本文中，我们研究了用于目标检测的知识提取（KD）策略，并提出了一个适用于同质和异质师生对的有效框架。传统的特征模拟范式引入模拟遮罩，在排除背景噪声的同时，将重点放在信息丰富的前景区域。然而，我们发现这些方法没有充分利用所有特征金字塔层次的语义信息，这导致基于FPN的检测器之间的知识提取效率低下。为此，我们提出了一种新的语义引导的特征模拟技术，该技术可以自动在所有金字塔级别的特征对之间进行软匹配，从而为学生提供最佳的指导。为了进一步推进包络，我们引入了对比蒸馏来有效地捕获编码在不同特征区域之间关系中的信息。最后，我们提出了一种广义检测KD管道，它能够同时提取同质和异质检测器对。我们的方法始终优于现有的检测KD技术，并且在（1）框架中的组件单独使用或结合使用时有效；（2）对于同质和异质学生-教师对和（3）多重检测基准。借助功能强大的X101 FasterRCNN Instaboost探测器，R50 FasterRCNN达到44.0%的AP，R50 RetinaNet达到43.3%的AP，R50-FCOS在COCO数据集上达到43.1%的AP。

weakly-supervised object localization (WSOL) enables finding an object using a dataset without any localization information. By simply training a classification model using only image-level annotations, the feature map of a model can be utilized as a score map for localization. In spite of many WSOL methods proposing novel strategies, there has not been any de facto standards about how to normalize the class activation map (CAM). Consequently, many WSOL methods have failed to fully exploit their own capacity because of the misuse of a normalization method. In this paper, we review many existing normalization methods and point out that they should be used according to the property of the given dataset. Additionally, we propose a new normalization method which substantially enhances the performance of any CAM-based WSOL methods. Using the proposed normalization method, we provide a comprehensive evaluation over three datasets (CUB, ImageNet and OpenImages) on three different architectures and observe significant performance gains over the conventional normalization methods in all the evaluated cases.

弱监督对象定位（WSOL）允许在没有任何定位信息的情况下使用数据集查找对象。通过仅使用图像级注释简单地训练分类模型，模型的特征映射可以用作用于定位的得分映射。尽管有许多WSOL方法提出了新的策略，但对于如何规范类激活映射（CAM）还没有任何事实上的标准。因此，由于误用了规范化方法，许多WSOL方法未能充分利用其自身的能力。在本文中，我们回顾了许多现有的规范化方法，并指出它们应该根据给定数据集的属性来使用。此外，我们还提出了一种新的标准化方法，大大提高了任何基于CAM的WSOL方法的性能。使用提出的标准化方法，我们对三种不同体系结构上的三个数据集（CUB、ImageNet和OpenImages）进行了综合评估，并在所有评估案例中观察到与传统标准化方法相比的显著性能提升。

This paper studies the problem of learning self-supervised representations on videos. In contrast to image modality that only requires appearance information on objects or scenes, video needs to further explore the relations between multiple frames/clips along the temporal dimension. However, the recent proposed contrastive-based self-supervised frameworks do not grasp such relations explicitly since they simply utilize two augmented clips from the same video and compare their distance without referring to their temporal relation. To address this, we present a contrast-and-order representation (CORP) framework for learning self-supervised video representations that can automatically capture both the appearance information within each frame and temporal information across different frames. In particular, given two video clips, our model first predicts whether they come from the same input video, and then predict the temporal ordering of the clips if they come from the same video. We also propose a novel decoupling attention method to learn symmetric similarity (contrast) and anti-symmetric patterns (order). Such design involves neither extra parameters nor computation, but can speed up the learning process and improve accuracy compared to the vanilla multi-head attention. We extensively validate the representation ability of our learned video features for the downstream action recognition task on Kinetics-400 and Something-something v2. Our method outperforms previous state-of-the-arts by a significant margin.

本文研究了视频自监督表示的学习问题。与仅需要对象或场景的外观信息的图像模式不同，视频需要进一步探索多个帧/片段之间沿时间维度的关系。然而，最近提出的基于对比的自监督框架并没有明确地把握这种关系，因为它们只是利用来自同一视频的两个增强片段，比较它们之间的距离，而不参考它们之间的时间关系。为了解决这个问题，我们提出了一个对比度和顺序表示（CORP）框架，用于学习自监督视频表示，该框架可以自动捕获每个帧内的外观信息和不同帧之间的时间信息。特别是，给定两个视频剪辑，我们的模型首先预测它们是否来自同一输入视频，然后预测剪辑的时间顺序（如果它们来自同一视频）。我们还提出了一种新的解耦注意方法来学习对称相似性（对比度）和反对称模式（顺序）。这种设计既不需要额外的参数，也不需要计算，但与普通的多头注意相比，可以加快学习过程并提高准确性。我们在Kinetics-400和某物V2上广泛验证了我们学习的视频特征在下游动作识别任务中的表现能力。我们的方法在很大程度上优于以前的先进水平。

We present an out-of-core variational approach for surface reconstruction from a set of aligned depth maps. Input depth maps are supposed to be reconstructed from regular photos or/and can be a representation of terrestrial LIDAR point clouds. Our approach is based on surface reconstruction via total generalized variation minimization (TGV) because of its strong visibility-based noise-filtering properties and GPU-friendliness. Our main contribution is an out-of-core OpenCL-accelerated adaptation of this numerical algorithm which can handle arbitrarily large real-world scenes with scale diversity.

我们提出了一种从一组对齐的深度图重建曲面的核外变分方法。输入深度图应根据常规照片或/或地面激光雷达点云进行重建。我们的方法是基于总广义变异最小化（TGV）的曲面重建，因为它具有强大的基于可见性的噪声滤波特性和GPU友好性。我们的主要贡献是对该数值算法的核心外OpenCL加速适应，该算法可以处理具有尺度多样性的任意大型真实场景。

The minimum graph cut and minimum s-t-cut problems are important primitives in the modeling of combinatorial problems in computer science, including in computer vision and machine learning. Some of the most efficient algorithms for finding global minimum cuts are randomized algorithms based on Karger's groundbreaking contraction algorithm. Here, we study whether Karger's algorithm can be successfully generalized to other cut problems. We first prove that a wide class of natural generalizations of Karger's algorithm cannot efficiently solve the s-t-mincut or the normalized cut problem to optimality. However, we then present a simple new algorithm for seeded segmentation / graph-based semi-supervised learning that is closely based on Karger's original algorithm, showing that for these problems, extensions of Karger's algorithm can be useful. The new algorithm has linear asymptotic runtime and yields a potential that can be interpreted as the posterior probability of a sample belonging to a given seed / class. We clarify its relation to the random walker algorithm / harmonic energy minimization in terms of distributions over spanning forests. On classical problems from seeded image segmentation and graph-based semi-supervised learning on image data, the method performs at least as well as the random walker / harmonic energy minimization / Gaussian processes.

最小图割和最小s-t割问题是计算机科学中组合问题建模的重要基础，包括计算机视觉和机器学习。寻找全局最小割集的一些最有效的算法是基于Karger开创性收缩算法的随机算法。在这里，我们研究Karger算法是否可以成功地推广到其他切割问题。我们首先证明了Karger算法的一大类自然推广不能有效地解决s-t-mincut或规范化割问题的最优性。然而，我们随后提出了一种简单的基于种子分割/图的半监督学习新算法，该算法紧密地基于Karger的原始算法，表明对于这些问题，Karger算法的扩展是有用的。新算法具有线性渐近运行时间，并产生一个可解释为属于给定种子/类的样本的后验概率的势。我们澄清了它与随机walker算法/谐波能量最小化在跨越森林分布方面的关系。对于种子图像分割和基于图的图像数据半监督学习等经典问题，该方法的性能至少与随机walker/谐波能量最小化/高斯过程相当。

Image translation methods typically aim to manipulate a set of labeled attributes (given as supervision at training time e.g. domain label) while leaving the unlabeled attributes intact. Current methods achieve either: (i) disentanglement, which exhibits low visual fidelity and can only be satisfied where the attributes are perfectly uncorrelated. (ii) visually-plausible translations, which are clearly not disentangled. In this work, we propose OverLORD, a single framework for disentangling labeled and unlabeled attributes as well as synthesizing high-fidelity images, which is composed of two stages; (i) Disentanglement: Learning disentangled representations with latent optimization. Differently from previous approaches, we do not rely on adversarial training or any architectural biases. (ii) Synthesis: Training feed-forward encoders for inferring the learned attributes and tuning the generator in an adversarial manner to increase the perceptual quality. When the labeled and unlabeled attributes are correlated, we model an additional representation that accounts for the correlated attributes and improves disentanglement. We highlight that our flexible framework covers multiple settings as disentangling labeled attributes, pose and appearance, localized concepts, and shape and texture. We present significantly better disentanglement with higher translation quality and greater output diversity than state-of-the-art methods.

图像转换方法通常旨在操作一组标记的属性（在训练时作为监督提供，例如域标签），同时保持未标记的属性完好无损。目前的方法实现了以下两种方法之一：(i) 解纠缠，其表现出较低的视觉保真度，并且只能在属性完全不相关的情况下满足。(ii) 在视觉上看似合理的翻译，这些翻译显然没有被解开。在这项工作中，我们提出了一个用于分离标记和未标记属性以及合成高保真图像的单一框架，该框架由两个阶段组成；(i) 解纠缠：通过潜在优化学习解纠缠表示。与以前的方法不同，我们不依赖对抗性训练或任何架构偏见。(ii) 综合：训练前馈编码器，以推断学习的属性，并以对抗的方式调整生成器，以提高感知质量。当标记属性和未标记属性相关时，我们对一个额外的表示进行建模，该表示说明了相关属

性并改进了解纠缠。我们强调，我们灵活的框架涵盖了多种设置，如分离标签属性、姿势和外观、本地化概念以及形状和纹理。与最先进的方法相比，我们提供了更高的翻译质量和更大的输出多样性。

This paper presents a generic method for generating full facial 3D animation from speech. Existing approaches to audio-driven facial animation exhibit uncanny or static upper face animation, fail to produce accurate and plausible co-articulation or rely on person-specific models that limit their scalability. To improve upon existing models, we propose a generic audio-driven facial animation approach that achieves highly realistic motion synthesis results for the entire face. At the core of our approach is a categorical latent space for facial animation that disentangles audio-correlated and audio-uncorrelated information based on a novel cross-modality loss. Our approach ensures highly accurate lip motion, while also synthesizing plausible animation of the parts of the face that are uncorrelated to the audio signal, such as eye blinks and eye brow motion. We demonstrate that our approach outperforms several baselines and obtains state-of-the-art quality both qualitatively and quantitatively. A perceptual user study demonstrates that our approach is deemed more realistic than the current state-of-the-art in over 75% of cases. We recommend watching the supplemental video before reading the paper: <https://github.com/facebookresearch/meshtalk>

本文提出了一种从语音中生成全人脸三维动画的通用方法。现有的音频驱动面部动画方法表现出神秘的或静态的上面部动画，无法产生准确和合理的共同发音，或者依赖于限制其可伸缩性的特定于人的模型。为了改进现有模型，我们提出了一种通用的音频驱动的人脸动画方法，该方法可以为整个人脸获得高度逼真的运动合成结果。我们的方法的核心是一个分类的面部动画潜在空间，该空间基于一种新的跨模态丢失分离音频相关和音频不相关信息。我们的方法确保了高度精确的嘴唇运动，同时还合成了与音频信号不相关的面部部分的合理动画，如眨眼和眉毛运动。我们证明了我们的方法优于几个基线，并在定性和定量上获得了最先进的质量。一项感性用户研究表明，在超过75%的情况下，我们的方法被认为比目前最先进的方法更现实。我们建议在阅读论文之前先观看补充视频：<https://github.com/facebookresearch/meshtalk>

Recently, the performance of single image super-resolution (SR) has been significantly improved with powerful networks. However, these networks are developed for image SR with specific integer scale factors (e.g., x2/3/4), and cannot handle non-integer and asymmetric SR. In this paper, we propose to learn a scale-arbitrary image SR network from scale-specific networks. Specifically, we develop a plug-in module for existing SR networks to perform scale-arbitrary SR, which consists of multiple scale-aware feature adaption blocks and a scale-aware upsampling layer. Moreover, conditional convolution is used in our plug-in module to generate dynamic scale-aware filters, which enables our network to adapt to arbitrary scale factors. Our plug-in module can be easily adapted to existing networks to realize scale-arbitrary SR with a single model. These networks plugged with our module can produce promising results for non-integer and asymmetric SR while maintaining state-of-the-art performance for SR with integer scale factors. Besides, the additional computational and memory cost of our module is very small.

近年来，由于强大的网络，单图像超分辨率（SR）的性能得到了显著提高。然而，这些网络是为具有特定整数比例因子（例如x2/3/4）的图像SR开发的，并且不能处理非整数和不对称SR。在本文中，我们建议从特定比例网络中学习任意比例的图像SR网络。具体来说，我们为现有SR网络开发了一个插件模块，用于执行任意规模的SR，该模块由多个规模感知特征自适应块和一个规模感知上采样层组成。此外，我们的插件模块中使用了条件卷积来生成动态比例感知滤波器，这使我们的网络能够适应任意比例因子。我们的插件模块可以很容易地适应现有的网络，用一个模型实现任意规模的SR。使用我们的模块插入的这些网络可以为非整数和不对称SR产生有希望的结果，同时保持整数比例因子SR的最新性能。此外，我们模块的额外计算和内存成本非常小。

Neural painting refers to the procedure of producing a series of strokes for a given image and non-photo-realistically recreating it using neural networks. While reinforcement learning (RL) based agents can generate a stroke sequence step by step for this task, it is not easy to train a stable RL agent. On the other hand, stroke optimization methods search for a set of stroke parameters iteratively in a large search space; such low efficiency significantly limits their prevalence and practicality. Different from previous methods, in this paper, we formulate the task as a set prediction problem and propose a novel Transformer-based framework, dubbed Paint Transformer, to predict the parameters of a stroke set with a feed forward network. This way, our model can generate a set of strokes in parallel and obtain the final painting of size 512x512 in near real time. More importantly, since there is no dataset available for training the Paint Transformer, we devise a self-training pipeline such that it can be trained without any off-the-shelf dataset while still achieving excellent generalization capability. Experiments demonstrate that our method achieves better painting performance than previous ones with cheaper training and inference costs. Codes and models will be available.

神经绘画是指为给定的图像生成一系列笔划，并使用神经网络以非照片真实的方式重新创建该图像的过程。虽然基于强化学习 (RL) 的agent可以为该任务逐步生成笔划序列，但训练稳定的RL agent并不容易。另一方面，笔划优化方法在较大的搜索空间内迭代搜索一组笔划参数；这种低效率极大地限制了它们的普遍性和实用性。与以往的方法不同，在本文中，我们将该任务描述为一个集合预测问题，并提出了一个新的基于变换器的框架，称为Paint-Transformer，用前馈网络预测笔划集合的参数。通过这种方式，我们的模型可以并行生成一组笔划，并几乎实时地获得大小为512x512的最终绘制。更重要的是，由于没有数据集可用于训练Paint Transformer，因此我们设计了一个自训练管道，这样就可以在不使用任何现成数据集的情况下对其进行训练，同时仍然可以实现出色的泛化能力。实验表明，该方法比以前的方法具有更好的绘制性能，且训练和推理成本更低。将提供代码和型号。

We propose PR-RRN, a novel neural-network based method for Non-rigid Structure-from-Motion (NRSfM). PR-RRN consists of Residual-Recursive Networks (RRN) and two extra regularization losses. RRN is designed to effectively recover 3D shape and camera from 2D keypoints with novel residual-recursive structure. As NRSfM is a highly under-constrained problem, we propose two new pairwise regularization to further regularize the reconstruction. The Rigidity-based Pairwise Contrastive Loss regularizes the shape representation by encouraging higher similarity between the representations of high-rigidity pairs of frames than low-rigidity pairs. We propose minimum singular-value ratio to measure the pairwise rigidity. The Pairwise Consistency Loss enforces the reconstruction to be consistent when the estimated shapes and cameras are exchanged between pairs. Our approach achieves state-of-the-art performance on CMU MOCAP and PASCAL3D+ dataset.

我们提出了PR-RRN，一种新的基于神经网络的非刚性结构运动识别方法（NRSfM）。PR-RRN由剩余递归网络（RRN）和两个额外的正则化损失组成。RRN采用新的残差递归结构，从二维关键点有效恢复三维形状和相机。由于NRSfM是一个高度欠约束的问题，我们提出了两种新的成对正则化方法来进一步正则化重建。基于刚度的成对对比损失通过鼓励高刚度框架对的表示之间比低刚度框架对的表示之间具有更高的相似性来规范形状表示。我们提出了最小奇异值比来度量两两刚度。当估计的形状和相机在对之间交换时，两两一致性损失强制重建是一致的。我们的方法在CMU MOCAP和PASCAL3D+数据集上实现了最先进的性能。

We study the machine's understanding of embodied reference: One agent uses both language and gesture to refer to an object to another agent in a shared physical environment. Of note, this new visual task requires understanding multimodal cues with perspective-taking to identify which object is being referred to. To tackle this problem, we introduce YouRefIt, a new crowd-sourced dataset of embodied reference collected in various physical scenes; the dataset contains 4,195 unique reference clips in 432 indoor scenes. To the best of our knowledge, this is the first embodied reference dataset that allows us to study referring expressions in daily physical scenes to understand referential behavior, human communication, and human-robot interaction. We further devise two benchmarks for image-based and video-based embodied reference understanding. Comprehensive baselines and extensive experiments provide the very first result of machine perception on how the referring expressions and gestures affect the embodied reference understanding. Our results provide essential evidence that gestural cues are as critical as language cues in understanding the embodied reference.

我们研究了机器对具象参照的理解：在共享的物理环境中，一个代理使用语言和手势将对象参照给另一个代理。值得注意的是，这项新的视觉任务需要通过透视来理解多模态线索，以识别所指的对象。为了解决这个问题，我们引入YouRefIt，这是一个新的众包数据集，收集了各种物理场景中的具体参考；该数据集包含432个室内场景中的4195个独特参考片段。据我们所知，这是第一个具体化的参考数据集，允许我们研究日常物理场景中的参考表达，以了解参考行为、人类通信和人类-机器人交互。我们进一步设计了基于图像和基于视频的具体参考理解的两个基准。全面的基线和广泛的实验提供了机器感知关于指称表达和手势如何影响具体指称理解的第一个结果。我们的研究结果提供了必要的证据，表明在理解具体指称时，手势线索和语言线索一样重要。

Most recent video super-resolution (SR) methods either adopt an iterative manner to deal with low-resolution (LR) frames from a temporally sliding window, or leverage the previously estimated SR output to help reconstruct the current frame recurrently. A few studies try to combine these two structures to form a hybrid framework but have failed to give full play to it. In this paper, we propose an omniscient framework to not only utilize the preceding SR output, but also leverage the SR outputs from the present and future. The omniscient framework is more generic because the iterative, recurrent and hybrid frameworks can be regarded as its special cases. The proposed omniscient framework enables a generator to behave better than its counterparts under other frameworks. Abundant experiments on public datasets show that our method is superior to the state-of-the-art methods in objective metrics, subjective visual effects and complexity.

最新的视频超分辨率 (SR) 方法要么采用迭代方式来处理来自时间滑动窗口的低分辨率 (LR) 帧，要么利用先前估计的SR输出来帮助重复地重建当前帧。一些研究试图将这两种结构结合起来形成一个混合框架，但未能充分发挥其作用。在本文中，我们提出了一个全知框架，不仅可以利用前面的SR输出，还可以利用现在和将来的SR输出。全知框架更为通用，因为迭代、循环和混合框架可以视为其特例。建议的全知框架使生成器比其他框架下的生成器表现得更好。在公共数据集上的大量实验表明，我们的方法在客观度量、主观视觉效果和复杂性方面优于现有的方法。

Gradient-based meta-learning relates task-specific models to a meta-model by gradients. By this design, an algorithm first optimizes the task-specific models by an inner loop and then backpropagates meta-gradients through the loop to update the meta-model. The number of inner-loop optimization steps has to be small (e.g., one step) to avoid high-order derivatives, big memory footprints, and the risk of vanishing or exploding meta-gradients. We propose an intuitive teacher-student scheme to enable the gradient-based meta-learning algorithms to explore long horizons by the inner loop. The key idea is to employ a student network to adequately explore the search space of task-specific models (e.g., by more than ten steps), and a teacher then takes a "leap" toward the regions probed by the student. The teacher not only arrives at a high-quality model but also defines a lightweight computation graph for meta-gradients. Our approach is generic; it performs well when applied to four meta-learning algorithms over three tasks: few-shot learning, long-tailed classification, and meta-attack.

基于梯度的元学习通过梯度将特定于任务的模型与元模型联系起来。通过这种设计，算法首先通过内部循环优化特定于任务的模型，然后通过循环反向传播元梯度以更新元模型。内循环优化步骤的数量必须很小（例如，一个步骤），以避免高阶导数、大内存占用以及元梯度消失或爆炸的风险。我们提出了一个直观的师生方案，使基于梯度的元学习算法能够通过内循环探索长视野。关键思想是利用学生网络充分探索任务特定模型的搜索空间（例如，通过十多个步骤），然后教师向学生探索的区域“跳跃”。老师不仅得到了一个高质量的模型，而且还为元梯度定义了一个轻量级的计算图。我们的方法是通用的；它在三个任务中应用于四种元学习算法时表现良好：少镜头学习、长尾分类和元攻击。

Recently, the encoder-decoder and intensity transformation approaches lead to impressive progress in image enhancement. However, the encoder-decoder often loses details in input images during down-sampling and up-sampling processes. Also, the intensity transformation has a limited capacity to cover color transformation between low-quality and high-quality images. In this paper, we propose a novel approach, called representative color transform (RCT), to tackle these issues in existing methods. RCT determines different representative colors specialized in input images and estimates transformed colors for the representative colors. It then determines enhanced colors using these transformed colors based on the similarity between input and representative colors. Extensive experiments demonstrate that the proposed algorithm outperforms recent state-of-the-art algorithms on various image enhancement problems.

最近，编码器-解码器和强度变换方法在图像增强方面取得了令人瞩目的进展。然而，在下采样和上采样过程中，编码器-解码器经常丢失输入图像中的细节。此外，强度变换覆盖低质量和高质量图像之间的颜色变换的能力有限。在本文中，我们提出了一种新的方法，称为代表性颜色变换（RCT），以解决现有方法中的这些问题。RCT确定输入图像中的不同代表性颜色，并估计代表性颜色的变换颜色。然后，它根据输入颜色和代表颜色之间的相似性，使用这些变换后的颜色确定增强颜色。大量的实验表明，该算法在各种图像增强问题上都优于目前最新的算法。

This paper presents a new Vision Transformer (ViT) architecture Multi-scale Vision Longformer, which significantly enhances the ViT of [??] for encoding high-resolution images using two techniques. The first is the multi-scale model structure, which provides image encodings at multiple scales with manageable computational cost. The second is the attention mechanism of Vision Longformer, which is a variant of Longformer [??], originally developed for natural language processing, and achieves a linear complexity w.r.t. the number of input tokens. A comprehensive empirical study shows that the new ViT significantly outperforms several strong baselines, including the existing ViT models and their ResNet counterparts, and the Pyramid Vision Transformer from a concurrent work [??], on a range of vision tasks, including image classification, object detection, and segmentation. The models and source code are released at <https://github.com/microsoft/vision-longformer>.

本文提出了一种新的视觉变换器 (ViT) 结构——多尺度视觉长形器，它显著提高了[? ? ]的ViT用于使用两种技术对高分辨率图像进行编码。第一种是多尺度模型结构，它以可管理的计算成本提供多尺度的图像编码。第二个是视觉长形器的注意机制，它是长形器的变体[? ? ]，最初是为自然语言处理而开发的，并且实现了线性复杂度w.r.t.输入标记的数量。一项全面的实证研究表明，新的ViT显著优于几个强基线，包括现有的ViT模型及其ResNet对应模型，以及并行工作中的金字塔视觉转换器[? ? ]，关于一系列视觉任务，包括图像分类、目标检测和分割。模型和源代码发布于<https://github.com/microsoft/vision-longformer>。

In structure-from-motion the viewing graph is a graph where vertices correspond to cameras and edges represent fundamental matrices. We provide a new formulation and an algorithm for establishing whether a viewing graph is solvable, i.e. it uniquely determines a set of projective cameras. Known theoretical conditions either do not fully characterize the solvability of all viewing graphs, or are exceedingly hard to compute for they involve solving a system of polynomial equations with a large number of unknowns. The main result of this paper is a method for reducing the number of unknowns by exploiting the cycle consistency. We advance the understanding of the solvability by (i) finishing the classification of all previously undecided minimal graphs up to 9 nodes, (ii) extending the practical solvability testing up to minimal graphs with up to 90 nodes, and (iii) definitely answering an open research question by showing that the finite solvability is not equivalent to the solvability. Finally, we present an experiment on real data showing that unsolvable graphs are appearing in practical situations.

在“运动结构”中，观察图是一个顶点对应于摄影机且边表示基本矩阵的图。我们提供了一个新的公式和算法来确定一个观察图是否可解，即它唯一地确定一组投影相机。已知的理论条件要么不能完全描述所有观察图的可解性，要么计算起来非常困难，因为它们涉及到求解含有大量未知量的多项式方程组。本文的主要结果是利用循环一致性减少未知量的方法。我们通过 (i) 完成所有之前未确定的最小图（最多9个节点）的分类， (ii) 将实际可解性测试扩展到最多90个节点的最小图，从而提高了对可解性的理解， (iii) 通过证明有限可解性不等同于可解性，明确回答了一个开放的研究问题。最后，我们给出了一个真实数据的实验，表明在实际情况下出现了不可解图。

There has been a booming demand for integrating Convolutional Neural Networks (CNNs) powered functionalities into Internet-of-Thing (IoT) devices to enable ubiquitous intelligent "IoT cameras". However, more extensive applications of such IoT systems are still limited by two challenges. First, some applications, especially medicine- and wearable-related ones, impose stringent requirements on the camera form factor. Second, powerful CNNs often require considerable storage and energy cost, whereas IoT devices often suffer from limited resources.

PhlatCam, with its form factor potentially reduced by orders of magnitude, has emerged as a promising solution to the first aforementioned challenge, while the second one remains a bottleneck. Existing compression techniques, which can potentially tackle the second challenge, are far from realizing the full potential in storage and energy reduction, because they mostly focus on the CNN algorithm itself. To this end, this work proposes SACoD, a Sensor Algorithm Co-Design framework to develop more efficient CNN-powered PhlatCam. In particular, the mask coded in the PhlatCam sensor and the backend CNN model are jointly optimized in terms of both model parameters and architectures via differential neural architecture search. Extensive experiments including both simulation and physical measurement on manufactured masks show that the proposed SACoD framework achieves aggressive model compression and energy savings while maintaining or even boosting the task accuracy, when benchmarking over two state-of-the-art (SOTA) designs with six datasets across four different vision tasks including classification, segmentation, image translation, and face recognition. Our codes are available at: <https://github.com/RICE-EIC/SACoD>.

将卷积神经网络 (CNN) 驱动的功能集成到物联网 (IoT) 设备中，以实现无处不在的智能“IoT摄像机”的需求日益旺盛。然而，此类物联网系统的更广泛应用仍然受到两个挑战的限制。首先，一些应用，特别是医药和可穿戴相关的应用，对相机的外形提出了严格的要求。其次，强大的CNN通常需要相当大的存储和能源成本，而物联网设备往往资源有限。PhlatCam的形状因子可能会减少几个数量级，它已成为上述第一个挑战的一个有希望的解决方案，而第二个挑战仍然是一个瓶颈。现有的压缩技术可能会解决第二个挑战，但远未实现存储和节能方面的全部潜力，因为它们主要集中在CNN算法本身。为此，本工作提出了SACoD，一个传感器算法协同设计框架，以开发更高效的CNN驱动的PhlatCam。特别是，PhlatCam传感器和后端CNN模型中编码的掩模通过差分神经结构搜索在模型参数和结构方面联合优化。大量实验（包括对制造的掩模的模拟和物理测量）表明，提出的SACoD框架在保持甚至提高任务精度的同时，实现了积极的模型压缩和节能，在对两个最先进的 (SOTA) 设计进行基准测试时，使用六个数据集，完成四个不同的视觉任务，包括分类、分割、图像翻译和人脸识别。我们的代码可从以下网址获得：<https://github.com/RICE-EIC/SACoD>.

Point cloud completion aims to predict a complete shape in high accuracy from its partial observation. However, previous methods usually suffered from discrete nature of point cloud and unstructured prediction of points in local regions, which makes it hard to reveal fine local geometric details on the complete shape. To resolve this issue, we propose SnowflakeNet with Snowflake Point Deconvolution (SPD) to generate the complete point clouds. The SnowflakeNet models the generation of complete point clouds as the snowflake-like growth of points in 3D space, where the child points are progressively generated by splitting their parent points after each SPD. Our insight of revealing detailed geometry is to introduce skip-transformer in SPD to learn point splitting patterns which can fit local regions the best. Skip-transformer leverages attention mechanism to summarize the splitting patterns used in the previous SPD layer to produce the splitting in the current SPD layer. The locally compact and structured point cloud generated by SPD is able to precisely capture the structure characteristic of 3D shape in local patches, which enables the network to predict highly detailed geometries, such as smooth regions, sharp edges and corners. Our experimental results outperform the state-of-the-art point cloud completion methods under widely used benchmarks. Code will be available at <https://github.com/AllenXiangX/SnowflakeNet>.

点云完成的目的是从局部观测中高精度地预测一个完整的形状。然而，以往的方法往往存在点云的离散性和局部区域点的非结构化预测等问题，难以揭示完整形状上的局部几何细节。为了解决这个问题，我们提出了雪花网和雪花点反褶积 (SPD) 来生成完整的点云。雪花网将完整点云的生成建模为三维空间中点的雪花状增长，其中子点通过在每个SPD后拆分其父点逐步生成。我们揭示细节几何的见解是在SPD中引入跳跃变压器，以学习最适合局部区域的点分裂模式。Skip transformer利用注意机制总结前一个SPD层中使用的分裂模式，以在当前SPD层中产生分裂。SPD生成的局部紧凑结构化点云能够精确捕捉局部面片中三维形状的结构特征，从而使网络能够预测高度详细的几何图形，如平滑区域、锐角和锐角。在广泛使用的基准下，我们的实验结果优于最先进的点云完成方法。代码将在<https://github.com/AllenXiangX/SnowflakeNet>。

Multi-modal learning, which focuses on utilizing various modalities to improve the performance of a model, is widely used in video recognition. While traditional multi-modal learning offers excellent recognition results, its computational expense limits its impact for many real-world applications. In this paper, we propose an adaptive multi-modal learning framework, called AdaMML, that selects on-the-fly the optimal modalities for each segment conditioned on the input for efficient video recognition. Specifically, given a video segment, a multi-modal policy network is used to decide what modalities should be used for processing by the recognition model, with the goal of improving both accuracy and efficiency. We efficiently train the policy network jointly with the recognition model using standard back-propagation. Extensive experiments on four challenging diverse datasets demonstrate that our proposed adaptive approach yields 35%-55% reduction in computation when compared to the traditional baseline that simply uses all the modalities irrespective of the input, while also achieving consistent improvements in accuracy over the state-of-the-art methods. Project page: <https://rpand002.github.io/adamml.html>.

多模态学习在视频识别中有着广泛的应用，它关注于利用各种模式来提高模型的性能。虽然传统的多模态学习提供了出色的识别结果，但其计算开销限制了其对许多实际应用的影响。在本文中，我们提出了一个称为AdaMML的自适应多模式学习框架，该框架根据输入动态地为每个片段选择最佳模式，以实现高效的视频识别。具体而言，给定一个视频片段，使用多模式策略网络来决定识别模型应使用哪些模式进行处理，目的是提高准确性和效率。我们使用标准的反向传播技术，结合识别模型，有效地训练策略网络。在四个具有挑战性的不同数据集上进行的大量实验表明，与传统基线相比，我们提出的自适应方

法在计算量上减少了35%-55%，而传统基线仅使用所有模式，而不考虑输入，同时与最先进的方法相比，在准确性方面也取得了一致的改进。项目页面：<https://rpand002.github.io/adamml.html>.

We propose 3DETR, an end-to-end Transformer based object detection model for 3D point clouds. Compared to existing detection methods that employ a number of 3D-specific inductive biases, 3DETR requires minimal modifications to the vanilla Transformer block. Specifically, we find that a standard Transformer with non-parametric queries and Fourier positional embeddings is competitive with specialized architectures that employ libraries of 3D-specific operators with hand-tuned hyperparameters. Nevertheless, 3DETR is conceptually simple and easy to implement, enabling further improvements by incorporating 3D domain knowledge. Through extensive experiments, we show 3DETR outperforms the well-established and highly optimized VoteNet baselines on the challenging ScanNetV2 dataset by 9.5%. Furthermore, we show 3DETR is applicable to 3D tasks beyond detection, and can serve as a building block for future research.

我们提出了3DETR，一种基于端到端转换器的三维点云目标检测模型。与使用大量3D特定感应偏压的现有检测方法相比，3DETR只需对普通变压器块进行最小修改。具体而言，我们发现，具有非参数查询和傅里叶位置嵌入的标准转换器与使用具有手动调整超参数的三维特定运算符库的专用体系结构具有竞争力。尽管如此，3DETR在概念上简单且易于实现，通过结合3D领域知识实现了进一步的改进。通过大量的实验，我们发现3DETR在具有挑战性的ScanNetV2数据集上的性能比完善且高度优化的VoteNet基线高9.5%。此外，我们还表明3DETR适用于检测不到的3D任务，可以作为未来研究的基础。

It is common practice to represent spoken languages at their phonetic level. However, for sign languages, this implies breaking motion into its constituent motion primitives. Avatar based Sign Language Production (SLP) has traditionally done just this, building up animation from sequences of hand motions, shapes and facial expressions. However, more recent deep learning based solutions to SLP have tackled the problem using a single network that estimates the full skeletal structure. We propose splitting the SLP task into two distinct jointly-trained sub-tasks. The first translation sub-task translates from spoken language to a latent sign language representation, with gloss supervision. Subsequently, the animation sub-task aims to produce expressive sign language sequences that closely resemble the learnt spatio-temporal representation. Using a progressive transformer for the translation sub-task, we propose a novel Mixture of Motion Primitives (MoMP) architecture for sign language animation. A set of distinct motion primitives are learnt during training, that can be temporally combined at inference to animate continuous sign language sequences. We evaluate on the challenging RWTH-PHOENIX-Weather-2014T(PHOENIX14T) dataset, presenting extensive ablation studies and showing that MoMP outperforms baselines in user evaluations. We achieve state-of-the-art back translation performance with an 11% improvement over competing results. Importantly, and for the first time, we showcase stronger performance for a full translation pipeline going from spoken language to sign, than from gloss to sign.

在语音水平上表现口语是一种常见的做法。然而，对于手语来说，这意味着将运动分解为其组成的运动原语。基于化身的手语制作（SLP）传统上就是这样做的，通过手部动作、形状和面部表情序列构建动画。然而，最近针对SLP的基于深度学习的解决方案使用一个单一网络来估计整个骨骼结构，从而解决了这个问题。我们建议将SLP任务分为两个不同的联合训练子任务。第一个翻译子任务将口语翻译成潜在的手语表达，并进行监督。随后，动画子任务旨在生成与所学时空表示非常相似的表达性手语序列。使用渐进式变换器作为翻译子任务，我们提出了一种新的混合运动原语（MoMP）手语动画体系结构。在训练过程中学习一组不同的运动原语，这些运动原语可以在推理时进行临时组合，从而为连续的手语序列制作动画。我们在具有挑战性的RWTH-PHOENIX-Weather-2014T (PHOENIX14T) 数据集上进行评估，展示了广泛的消融研究，并表明MoMP在用户评估中优于基线。我们实现了最先进的回译性能，

比竞争结果提高了11%。重要的是，我们首次展示了从口头语言到手语的完整翻译管道比从手语到手语的完整翻译管道具有更强的性能。

We present DocFormer – a multi-modal transformer based architecture for the task of Visual Document Understanding (VDU). VDU is a challenging problem which aims to understand documents in their varied formats(forms, receipts etc.) and layouts. In addition, DocFormer is pre-trained in an unsupervised fashion using carefully designed tasks which encourage multi-modal interaction. DocFormer uses text, vision and spatial features and combines them using a novel multi-modal self-attention layer. DocFormer also shares learned spatial embeddings across modalities which makes it easy for the model to correlate text to visual tokens and vice versa. DocFormer is evaluated on 4 different datasets each with strong baselines. DocFormer achieves state-of-the-art results on all of them, sometimes beating models 4x its size (in no. of parameters)

我们提出DocFormer——一种基于多模式转换器的体系结构，用于可视化文档理解（VDU）任务。VDU是一个具有挑战性的问题，旨在理解各种格式（表格、收据等）和布局的文档。此外，DocFormer使用精心设计的任务以无监督的方式进行预培训，这些任务鼓励多模态交互。DocFormer使用文本、视觉和空间特征，并使用新颖的多模式自我注意层将它们结合起来。DocFormer还跨模式共享学习到的空间嵌入，这使得模型很容易将文本与视觉标记关联起来，反之亦然。DocFormer在4个不同的数据集上进行评估，每个数据集都有强大的基线。DocFormer在所有这些产品上都达到了最先进的效果，有时甚至超过了其尺寸的4倍（参数数量）

We present a general learning-based solution for restoring images suffering from spatially-varying degradations. Prior approaches are typically degradation-specific and employ the same processing across different images and different pixels within. However, we hypothesize that such spatially rigid processing is suboptimal for simultaneously restoring the degraded pixels as well as reconstructing the clean regions of the image. To overcome this limitation, we propose SPAIR, a network design that harnesses distortion-localization information and dynamically adjusts computation to difficult regions in the image. SPAIR comprises of two components, (1) a localization network that identifies degraded pixels, and (2) a restoration network that exploits knowledge from the localization network in filter and feature domain to selectively and adaptively restore degraded pixels. Our key idea is to exploit the non-uniformity of heavy degradations in spatial-domain and suitably embed this knowledge within distortion-guided modules performing sparse normalization, feature extraction and attention. Our architecture is agnostic to physical formation model and generalizes across several types of spatially-varying degradations. We demonstrate the efficacy of SPAIR individually on four restoration tasks- removal of rain-streaks, raindrops, shadows and motion blur. Extensive qualitative and quantitative comparisons with prior art on 11 benchmark datasets demonstrate that our degradation-agnostic network design offers significant performance gains over state-of-the-art degradation-specific architectures. Code available at <https://github.com/human-analysis/spatially-adaptive-image-restoration>.

我们提出了一个通用的基于学习的解决方案，用于恢复遭受空间变化退化的图像。先前的方法通常是特定于退化的，并且在不同的图像和图像中的不同像素上采用相同的处理。然而，我们假设这样的空间刚性处理对于同时恢复退化像素以及重建图像的干净区域而言是次优的。为了克服这一局限性，我们提出了SPAIR，这是一种利用失真定位信息并根据图像中的困难区域动态调整计算的网络设计。SPAIR由两部分组成：（1）一个用于识别退化像素的定位网络，以及（2）一个恢复网络，该网络利用来自滤波器和特征域中的定位网络的知识来选择性和自适应地恢复退化像素。我们的关键思想是利用空间域中严重退化的不均匀性，并将这些知识适当地嵌入到执行稀疏归一化、特征提取和注意的失真引导模块中。我们的体系结构与物理形成模型无关，并概括了几种类型的空间变化退化。我们分别展示了SPAIR在四项恢复任务中的功效——去除雨条纹、雨滴、阴影和运动模糊。在11个基准数据集上与现有技术进行了广泛

的定性和定量比较，结果表明，与最先进的特定于退化的体系结构相比，我们的退化不可知网络设计提供了显著的性能提升。代码可在<https://github.com/human-analysis/spatially-adaptive-image-restoration>.

Crowd counting is a difficult task because of the diversity of scenes. Most of the existing crowd counting methods adopt complex structures with massive backbones to enhance the generalization ability. Unfortunately, the performance of existing methods on large-scale data sets is not satisfactory. In order to handle various scenarios with less complex network, we explored how to efficiently use the multi-expert model for crowd counting tasks. We mainly focus on how to train more efficient expert networks and how to choose the most suitable expert. Specifically, we propose a task-driven similarity metric based on sample's mutual enhancement, referred as co-fine-tune similarity, which can find a more efficient subset of data for training the expert network. Similar samples are considered as a cluster which is used to obtain parameters of an expert. Besides, to make better use of the proposed method, we design a simple network called FPN with Deconvolution Counting Network, which is a more suitable base model for the multi-expert counting network. Experimental results show that multiple experts FDC (MFDC) achieves the best performance on four public data sets, including the large scale NWPU-Crowd data set. Furthermore, the MFDC trained on an extensive dense crowd data set can generalize well on the other data sets without extra training or fine-tuning.

由于场景的多样性，人群计数是一项困难的任务。现有的人群计数方法大多采用复杂的结构和大量的主干来提高泛化能力。不幸的是，现有方法在大规模数据集上的性能并不令人满意。为了处理网络不太复杂的各种场景，我们探索了如何有效地使用多专家模型进行人群计数任务。我们主要关注如何训练更高效的专家网络以及如何选择最合适专家。具体来说，我们提出了一种基于样本互增强的任务驱动相似度度量，称为协同微调相似度，它可以找到更有效的数据子集来训练专家网络。相似样本被视为一个聚类，用于获取专家的参数。此外，为了更好地利用所提出的方法，我们设计了一个简单的带反褶积计数网络的FPN网络，这是一个更适合于多专家计数网络的基础模型。实验结果表明，多专家FDC (MFDC) 在四个公共数据集（包括大规模NWPU群组数据集）上的性能最好。此外，在大量密集人群数据集上训练的MFDC可以很好地推广到其他数据集，而无需额外训练或微调。

Deep features have been proven powerful in building accurate dense semantic correspondences in various previous works. However, the multi-scale and pyramidal hierarchy of convolutional neural networks has not been well studied to learn discriminative pixel-level features for semantic correspondence. In this paper, we propose a multiscale matching network that is sensitive to tiny semantic differences between neighboring pixels. We follow the coarse-to-fine matching strategy, and build a top-down feature and matching enhancement scheme that is coupled with the multi-scale hierarchy of deep convolutional neural networks. During feature enhancement, intra-scale enhancement fuses same-resolution feature maps from multiple layers together via local self-attention, and cross-scale enhancement hallucinates higher resolution feature maps along the top-down hierarchy. Besides, we learn complementary matching details at different scales, and thus the overall matching score is refined by features at different semantic levels gradually. Our multi-scale matching network can be trained end-to-end easily with few additional learnable parameters. Experimental results demonstrate the proposed method achieves state-of-the-art performance on three popular benchmarks with high computational efficiency.

在以前的各种工作中，深度特征在构建精确的密集语义对应方面已经被证明是强大的。然而，卷积神经网络的多尺度和金字塔层次结构还没有得到很好的研究来学习用于语义对应的区分像素级特征。在本文中，我们提出了一种多尺度匹配网络，它对相邻像素之间的微小语义差异非常敏感。我们遵循从粗到精的匹配策略，构建了一个自顶向下的特征和匹配增强方案，该方案与深度卷积神经网络的多尺度层次结

构相耦合。在特征增强过程中，尺度内增强通过局部自关注将来自多个层的相同分辨率的特征映射融合在一起，而尺度间增强则沿着自上而下的层次结构产生更高分辨率的特征映射。此外，我们还学习了不同尺度下的互补匹配细节，从而根据不同语义层次的特征逐步细化整体匹配分数。我们的多尺度匹配网络可以很容易地进行端到端的训练，只需要很少的额外可学习参数。实验结果表明，该方法在三个流行的基准测试上都取得了最新的性能，并且具有较高的计算效率。

Increasing demands for understanding the internal behavior of convolutional neural networks (CNNs) have led to remarkable improvements in explanation methods. Particularly, several class activation mapping (CAM) based methods, which generate visual explanation maps by a linear combination of activation maps from CNNs, have been proposed. However, the majority of the methods lack a clear theoretical basis on how they assign the coefficients of the linear combination. In this paper, we revisit the intrinsic linearity of CAM with respect to the activation maps; we construct an explanation model of CNN as a linear function of binary variables that denote the existence of the corresponding activation maps. With this approach, the explanation model can be determined by additive feature attribution methods in an analytic manner. We then demonstrate the adequacy of SHAP values, which is a unique solution for the explanation model with a set of desirable properties, as the coefficients of CAM. Since the exact SHAP values are unattainable, we introduce an efficient approximation method, LIFT-CAM, based on DeepLIFT. Our proposed LIFT-CAM can estimate the SHAP values of the activation maps with high speed and accuracy. Furthermore, it greatly outperforms other previous CAM-based methods in both qualitative and quantitative aspects.

对卷积神经网络 (CNN) 内部行为的理解需求不断增加，这导致了解释方法的显著改进。特别是，已经提出了几种基于类激活映射 (CAM) 的方法，这些方法通过CNN激活映射的线性组合生成可视化解释映射。然而，大多数方法在如何分配线性组合的系数方面缺乏明确的理论基础。在本文中，我们重新讨论了CAM关于激活映射的固有线性；我们构造了一个CNN的解释模型，作为表示相应激活映射存在的二元变量的线性函数。通过这种方法，可以通过分析的方式通过附加特征属性方法确定解释模型。然后，我们证明了形状值的充分性，这是解释模型的唯一解决方案，具有一组理想的特性，作为CAM的系数。由于无法获得精确的形状值，我们介绍了一种基于DeepLIFT的高效近似方法LIFT-CAM。我们提出的升降凸轮可以快速准确地估计激活图的形状值。此外，它在定性和定量方面都大大优于其他基于CAM的方法。

video semantic segmentation is an essential task for the analysis and understanding of videos. Recent efforts largely focus on supervised video segmentation by learning from fully annotated data, but the learnt models often experience clear performance drop while applied to videos of a different domain. This paper presents DA-VSN, a domain adaptive video segmentation network that addresses domain gaps in videos by temporal consistency regularization (TCR) for consecutive frames of target-domain videos. DA-VSN consists of two novel and complementary designs. The first is cross-domain TCR that guides the prediction of target frames to have similar temporal consistency as that of source frames (learnt from annotated source data) via adversarial learning. The second is intra-domain TCR that guides unconfident predictions of target frames to have similar temporal consistency as confident predictions of target frames. Extensive experiments demonstrate the superiority of our proposed domain adaptive video segmentation network which outperforms multiple baselines consistently by large margins.

视频语义分割是分析和理解视频的一项重要任务。最近的研究主要集中在通过从完全注释的数据中学习来进行有监督的视频分割，但是学到的模型在应用于不同领域的视频时通常会经历明显的性能下降。本文提出了一种域自适应视频分割网络DA-VSN，它通过对目标域视频连续帧的时间一致性正则化 (TCR) 来解决视频中的域间隙问题。DA-VSN由两种新颖且互补的设计组成。第一种是跨域TCR，它通

通过对抗性学习引导目标帧的预测具有与源帧（从注释源数据学习）相似的时间一致性。第二种是域内TCR，它引导目标帧的不确定预测具有与目标帧的可靠预测相似的时间一致性。大量的实验证明了我们提出的域自适应视频分割网络的优越性，它在很大程度上优于多个基线。

Personalized video highlight detection aims to shorten a long video to interesting moments according to a user's preference, which has recently raised the community's attention. Current methods regard the user's history as holistic information to predict the user's preference but negating the inherent diversity of the user's interests, resulting in vague preference representation. In this paper, we propose a simple yet efficient preference reasoning framework (PR-Net) to explicitly take the diverse interests into account for frame-level highlight prediction. Specifically, distinct user-specific preferences for each input query frame are produced, presented as the similarity weighted sum of history highlights to the corresponding query frame. Next, distinct comprehensive preferences are formed by the user-specific preferences and a learnable generic preference for more overall highlight measurement. Lastly, the degree of highlight and non-highlight for each query frame is calculated as semantic similarity to its comprehensive and non-highlight preferences, respectively. Besides, to alleviate the ambiguity due to the incomplete annotation, a new bi-directional contrastive loss is proposed to ensure a compact and differentiable metric space. In this way, our method significantly outperforms state-of-the-art methods with a relative improvement of 12% in mean accuracy precision.

个性化视频高光检测旨在根据用户的偏好将长视频缩短为有趣的时刻，这一点最近引起了社区的关注。当前的方法将用户的历史视为预测用户偏好的整体信息，但否定了用户兴趣的固有多样性，导致偏好表达模糊。在本文中，我们提出了一个简单而有效的偏好推理框架（PR-Net），以明确考虑帧级高光预测的不同兴趣。具体而言，为每个输入查询帧生成不同的特定于用户的首选项，作为对应查询帧的历史突出显示的相似度加权和表示。接下来，不同的综合偏好由用户特定偏好和可学习的通用偏好形成，用于更全面的高光测量。最后，计算每个查询帧的高亮和非高亮程度，分别作为其综合偏好和非高亮偏好的语义相似度。此外，为了减少由于注释不完整造成的歧义，提出了一种新的双向对比损失，以确保度量空间的紧凑性和可微性。这样，我们的方法明显优于最先进的方法，平均准确度相对提高12%。

Model quantization is an important mechanism for energy-efficient deployment of deep neural networks on resource-constrained devices by reducing the bit precision of weights and activations. However, it remains challenging to maintain high accuracy as bit precision decreases, especially for low-precision networks (e.g., 2-bit MobileNetV2). Existing methods have explored to address this problem by minimizing the quantization error or mimicking the data distribution of full-precision networks. In this work, we propose a novel weight regularization algorithm for improving low-precision network quantization. Instead of constraining the overall data distribution, we separably optimize all elements in each quantization bin to be as close to the target quantized value as possible. Such bin regularization (BR) mechanism encourages the weight distribution of each quantization bin to be sharp and approximate to a Dirac delta distribution ideally. Experiments demonstrate that our method achieves consistent improvements over the state-of-the-art quantization-aware training methods for different low-precision networks. Particularly, our bin regularization improves LSQ for 2-bit MobileNetV2 and MobileNetV3-Small by 3.9% and 4.9% top-1 accuracy on ImageNet, respectively.

通过降低权重和激活的比特精度，模型量化是在资源受限设备上高效部署深度神经网络的重要机制。然而，随着位精度的降低，保持高精度仍然是一个挑战，特别是对于低精度网络（例如，2位MobileNetV2）。现有的方法已经探索通过最小化量化误差或模仿全精度网络的数据分布来解决这个问题。在这项工作中，我们提出了一种新的权重正则化算法来改进低精度网络量化。我们不限制总体数据分布，而是对每个量化单元中的所有元素进行分离优化，使其尽可能接近目标量化值。这种面元正则化

(BR) 机制鼓励每个量化面元的权重分布是尖锐的，并且在理想情况下近似于Dirac delta分布。实验表明，对于不同的低精度网络，我们的方法比最先进的量化感知训练方法取得了一致的改进。特别是，我们的bin正则化将ImageNet上2位MobileNetV2和MobileNetV3的LSQ分别提高了3.9%和4.9%的top-1精度。

We tackle the long-tailed visual recognition problem from the knowledge distillation perspective by proposing a Distill the virtual Examples (DiVE) method. Specifically, by treating the predictions of a teacher model as virtual examples, we prove that distilling from these virtual examples is equivalent to label distribution learning under certain constraints. We show that when the virtual example distribution becomes flatter than the original input distribution, the under-represented tail classes will receive significant improvements, which is crucial in long-tailed recognition. The proposed DiVE method can explicitly tune the virtual example distribution to become flat. Extensive experiments on three benchmark datasets, including the large-scale iNaturalist ones, justify that the proposed DiVE method can significantly outperform state-of-the-art methods. Furthermore, additional analyses and experiments verify the virtual example interpretation, and demonstrate the effectiveness of tailored designs in DiVE for long-tailed problems.

我们从知识提取的角度解决了长尾视觉识别问题，提出了一种提取虚拟示例 (DiVE) 的方法。具体来说，通过将教师模型的预测视为虚拟样本，我们证明了从这些虚拟样本中提取的信息等价于在某些约束条件下的标签分布学习。我们表明，当虚拟示例分布变得比原始输入分布更平坦时，表示不足的尾部类将得到显著改进，这在长尾识别中至关重要。所提出的DiVE方法可以显式地调整虚拟样本分布，使其变得平坦。在三个基准数据集（包括大规模iNaturalist数据集）上进行的大量实验证明，所提出的潜水方法可以显著优于最先进的方法。此外，额外的分析和实验证了虚拟示例的解释，并证明了针对长尾问题的定制设计的有效性。

We introduce a Large Scale Multi-Illuminant (LSMI) Dataset that contains 7,486 images, captured with three different cameras on more than 2,700 scenes with two or three illuminants. For each image in the dataset, the new dataset provides not only the pixel-wise ground truth illumination but also the chromaticity of each illuminant in the scene and the mixture ratio of illuminants per pixel. Images in our dataset are mostly captured with illuminants existing in the scene, and the ground truth illumination is computed by taking the difference between the images with different illumination combination. Therefore, our dataset captures natural composition in the real-world setting with wide field-of-view, providing more extensive dataset compared to existing datasets for multi-illumination white balance. As conventional single illuminant white balance algorithms cannot be directly applied, we also apply per-pixel DNN-based white balance algorithm and show its effectiveness against using patch-wise white balancing. We validate the benefits of our dataset through extensive analysis including a user-study, and expect the dataset to make meaningful contribution for future work in white balancing.

我们介绍了一个包含7486幅图像的大规模多光源 (LSMI) 数据集，这些图像由三个不同的摄像机在2700多个场景中使用两个或三个光源拍摄。对于数据集中的每个图像，新数据集不仅提供像素级地面真实照明，还提供场景中每个光源的色度以及每个像素的光源混合比。我们的数据集中的图像大多是在场景中存在光源的情况下捕获的，并且通过获取具有不同照明组合的图像之间的差异来计算地面真实照明。因此，我们的数据集在真实环境中以宽视野捕捉自然成分，与现有的多照明白平衡数据集相比，提供了更广泛的数据集。由于传统的单光源白平衡算法不能直接应用，我们也采用了基于每像素DNN的白平衡算法，并证明了其对采用分片白平衡的有效性。我们通过广泛的分析（包括用户研究）验证了数据集的好处，并期望该数据集为白平衡的未来工作做出有意义的贡献。

The performance of surface registration relies heavily on the metric used for the alignment error between the source and target shapes. Traditionally, such a metric is based on the point-to-point or point-to-plane distance from the points on the source surface to their closest points on the target surface, which is susceptible to failure due to instability of the closest-point correspondence. In this paper, we propose a novel metric based on the intersection points between the two shapes and a random straight line, which does not assume a specific correspondence. We verify the effectiveness of this metric by extensive experiments, including its direct optimization for a single registration problem as well as unsupervised learning for a set of registration problems. The results demonstrate that the algorithms utilizing our proposed metric outperforms the state-of-the-art optimization-based and unsupervised learning-based methods.

曲面配准的性能在很大程度上依赖于用于源形状和目标形状之间对齐误差的度量。传统上，此类度量基于从源表面上的点到目标表面上的最近点的点到点或点到平面的距离，由于最近点对应的不稳定性，该距离容易发生故障。在本文中，我们提出了一种基于两个形状之间的交点和一条随机直线的新度量，该度量不假设特定的对应关系。我们通过大量实验验证了该度量的有效性，包括对单个注册问题的直接优化以及对一组注册问题的无监督学习。结果表明，利用我们提出的度量的算法优于基于优化和无监督学习的最新方法。

We present PICCOLO, a simple and efficient algorithm for omnidirectional localization. Given a colored point cloud and a 360 panorama image of a scene, our objective is to recover the camera pose at which the panorama image is taken. Our pipeline works in an off-the-shelf manner with a single image given as a query and does not require any training of neural networks or collecting ground-truth poses of images. Instead, we match each point cloud color to the holistic view of the panorama image with gradient-descent optimization to find the camera pose. Our loss function, called sampling loss, is point cloud-centric, evaluated at the projected location of every point in the point cloud. In contrast, conventional photometric loss is image-centric, comparing colors at each pixel location. with a simple change in the compared entities, sampling loss effectively overcomes the severe visual distortion of omnidirectional images, and enjoys the global context of the 360 view to handle challenging scenarios for visual localization. PICCOLO outperforms existing omnidirectional localization algorithms in both accuracy and stability when evaluated in various environments.

我们提出了一种简单有效的全方位定位算法PICCOLO。给定场景的彩色点云和360度全景图像，我们的目标是恢复拍摄全景图像时的相机姿态。我们的管道以现成的方式工作，将单个图像作为查询，不需要任何神经网络训练或收集图像的地面真实姿势。相反，我们通过梯度下降优化将每个点云颜色与全景图像的整体视图相匹配，以找到相机姿势。我们的损失函数称为采样损失，以点云为中心，在点云中每个点的投影位置进行评估。相比之下，传统的光度损失以图像为中心，比较每个像素位置的颜色。通过对比较实体的简单更改，采样丢失有效地克服了全向图像的严重视觉失真，并享受360 view的全局环境，以处理视觉定位的挑战性场景。在各种环境下评估时，PICCOLO在准确性和稳定性方面都优于现有的全方位定位算法。

The existence of noisy data is prevalent in both the training and testing phases of machine learning systems, which inevitably leads to the degradation of model performance. There have been plenty of works concentrated on learning with in-distribution (IND) noisy labels in the last decade, i.e., some training samples are assigned incorrect labels that do not correspond to their true classes. Nonetheless, in real application scenarios, it is necessary to consider the influence of out-of-distribution (OOD) samples, i.e., samples that do not belong to any known classes, which has not been sufficiently explored yet. To remedy this, we study a new problem setup, namely Learning with Open-world Noisy Data (LOND). The goal of LOND is to simultaneously learn a classifier and an OOD detector from datasets with mixed IND and OOD noise. In this paper, we propose a new graph-based framework, namely Noisy Graph Cleaning (NGC), which collects clean samples by leveraging geometric structure of data and model predictive confidence. Without any additional training effort, NGC can detect and reject the OOD samples based on the learned class prototypes directly in testing phase. We conduct experiments on multiple benchmarks with different types of noise and the results demonstrate the superior performance of our method against state of the arts.

在机器学习系统的训练和测试阶段，噪声数据的存在是普遍存在的，这不可避免地导致模型性能的下降。在过去的十年中，有大量的工作集中在使用分布内（IND）噪声标签的学习上，即，一些训练样本被分配了不正确的标签，这些标签与它们的真实类别不符。然而，在实际的应用场景中，有必要考虑不分配（OOD）样本的影响，即不属于任何已知类别的样本，这些样本尚未被充分探索。为了解决这个问题，我们研究了一个新的问题设置，即开放世界噪声数据学习（LOND）。LOND的目标是从具有IND和OOD混合噪声的数据集中同时学习分类器和OOD检测器。在本文中，我们提出了一种新的基于图的框架，即噪声图清理（NGC），该框架利用数据的几何结构和模型预测置信度来收集干净的样本。NGC不需要任何额外的训练，就可以在测试阶段直接基于所学的类原型检测和拒绝OOD样本。我们在具有不同类型噪声的多个基准上进行了实验，结果证明了我们的方法在对抗最新技术方面的优越性能。

Superpoints are formed by grouping similar points with local geometric structures, which can effectively reduce the number of primitives of point clouds for subsequent point cloud processing. Existing superpoint methods mainly focus on employing clustering or graph partition to generate superpoints with handcrafted or learned features. Nonetheless, these methods cannot learn superpoints of point clouds with an end-to-end network. In this paper, we develop a new deep iterative clustering network to directly generate superpoints from irregular 3D point clouds in an end-to-end manner. Specifically, in our clustering network, we first jointly learn a soft point-superpoint association map from the coordinate and feature spaces of point clouds, where each point is assigned to the superpoint with a learned weight. Furthermore, we then iteratively update the association map and superpoint centers so that we can more accurately group the points into the corresponding superpoints with locally similar geometric structures. Finally, by predicting the pseudo labels of the superpoint centers, we formulate a label consistency loss on the points and superpoint centers to train the network. Extensive experiments on various datasets indicate that our method not only achieves the state-of-the-art on superpoint generation but also improves the performance of point cloud semantic segmentation. Code is available at <https://github.com/fpthink/SPNet>.

通过将相似点与局部几何结构分组形成超点，可有效减少点云的基元数量，便于后续点云处理。现有的超点方法主要集中在使用聚类或图划分来生成具有手工或学习特征的超点。尽管如此，这些方法无法通过端到端网络学习点云的重叠。在本文中，我们开发了一种新的深度迭代聚类网络，以端到端的方式直接从不规则的三维点云生成超点。具体地说，在我们的聚类网络中，我们首先从点云的坐标空间和特征空间联合学习一个软点超点关联图，其中每个点都被分配给具有学习权重的超点。此外，我们接着迭代更新关联映射和超点中心，以便我们能够更准确地将点分组到具有局部相似几何结构的对应超点中。最

后，通过预测超点中心的伪标签，我们在点和超点中心上建立了标签一致性损失来训练网络。在各种数据集上的大量实验表明，该方法不仅实现了最新的超点生成技术，而且提高了点云语义分割的性能。代码可在<https://github.com/fpthink/SPNet>.

Gaze following, i.e., detecting the gaze target of a human subject, in 2D images has become an active topic in computer vision. However, it usually suffers from the out of frame issue due to the limited field-of-view (FoV) of 2D images. In this paper, we introduce a novel task, gaze following in 360-degree images which provide an omnidirectional FoV and can alleviate the out of frame issue. We collect the first dataset, "GazeFollow360", for this task, containing around 10,000 360-degree images with complex gaze behaviors under various scenes. Existing 2D gaze following methods suffer from performance degradation in 360-degree images since they may use the assumption that a gaze target is in the 2D gaze sight line. However, this assumption is no longer true for long-distance gaze behaviors in 360-degree images, due to the distortion brought by sphere-to-plane projection. To address this challenge, we propose a 3D sight line guided dual-pathway framework, to detect the gaze target within a local region (here) and from a distant region (there), parallelly. Specifically, the local region is obtained as a 2D cone-shaped field along the 2D projection of the sight line starting at the human subject's head position, and the distant region is obtained by searching along the sight line in 3D sphere space. Finally, the location of the gaze target is determined by fusing the estimations from both the local region and the distant region. Experimental results show that our method achieves significant improvements over previous 2D gaze following methods on our GazeFollow360 dataset.

在二维图像中，视线跟踪，即检测人类目标的视线，已经成为计算机视觉中的一个活跃话题。然而，由于二维图像的视野 (FoV) 有限，它通常会遇到帧外问题。在本文中，我们介绍了一种新的任务，360度图像中的凝视跟踪，它提供了一个全方位的视野，可以缓解帧外问题。我们为此任务收集了第一个数据集“GazeFollow360”，其中包含大约10000张360度的图像，在各种场景下具有复杂的凝视行为。现有的二维注视跟踪方法在360度图像中存在性能退化问题，因为它们可能使用注视目标位于二维注视视线中的假设。然而，对于360度图像中的远距离凝视行为，由于球体到平面投影带来的失真，这种假设不再成立。为了应对这一挑战，我们提出了一种三维视线引导的双路径框架，用于并行地检测局部区域（此处）和远处区域（此处）内的凝视目标。具体地说，局部区域作为沿视线的2D投影的2D锥形场从人类被摄体的头部位置开始获得，并且通过在3D球面空间中沿视线搜索获得远处区域。最后，通过融合来自局部区域和远处区域的估计来确定凝视目标的位置。实验结果表明，在我们的GazeFollow360数据集上，我们的方法比以前的2D注视跟踪方法取得了显著的改进。

When a camera is pointed at a strong light source, the resulting photograph may contain lens flare artifacts. Flares appear in a wide variety of patterns (halos, streaks, color bleeding, haze, etc.) and this diversity in appearance makes flare removal challenging. Existing analytical solutions make strong assumptions about the artifact's geometry or brightness, and therefore only work well on a small subset of flares. Machine learning techniques have shown success in removing other types of artifacts, like reflections, but have not been widely applied to flare removal due to the lack of training data. To solve this problem, we explicitly model the optical causes of flare either empirically or using wave optics, and generate semi-synthetic pairs of flare-corrupted and clean images. This enables us to train neural networks to remove lens flare for the first time. Experiments show our data synthesis approach is critical for accurate flare removal, and that models trained with our technique generalize well to real lens flares across different scenes, lighting conditions, and cameras.

当相机指向强光源时，产生的照片可能包含镜头光斑伪影。耀斑以各种各样的模式出现（光晕、条纹、渗色、薄雾等），这种外观上的多样性使耀斑消除具有挑战性。现有的分析解决方案对人造物体的几何结构或亮度做出了强有力假设，因此只适用于一小部分耀斑。机器学习技术已成功地去除了其他类型的工件，如反射，但由于缺乏训练数据，尚未广泛应用于光斑去除。为了解决这个问题，我们通过经验或使用波动光学明确地模拟耀斑的光学原因，并生成耀斑损坏和干净图像的半合成对。这使我们能够训练神经网络来第一次去除镜头光斑。实验表明，我们的数据合成方法对于精确消除光斑至关重要，并且使用我们的技术训练的模型能够很好地推广到不同场景、照明条件和相机的真实镜头光斑。

We propose DeepMultiCap, a novel method for multi-person performance capture using sparse multi-view cameras. Our method can capture time varying surface details without the need of using pre-scanned template models. To tackle with the serious occlusion challenge for close interacting scenes, we combine a recently proposed pixel-aligned implicit function with parametric model for robust reconstruction of the invisible surface areas. An effective attention-aware module is designed to obtain the fine-grained geometry details from multi-view images, where high-fidelity results can be generated. In addition to the spatial attention method, for video inputs, we further propose a novel temporal fusion method to alleviate the noise and temporal inconsistencies for moving character reconstruction. For quantitative evaluation, we contribute a high quality multi-person dataset, MultiHuman, which consists of 150 static scenes with different levels of occlusions and ground truth 3D human models. Experimental results demonstrate the state-of-the-art performance of our method and the well generalization to real multiview video data, which outperforms the prior works by a large margin.

我们提出了DeepMultiCap，这是一种使用稀疏多视图相机进行多人性能捕获的新方法。我们的方法可以捕获随时间变化的表面细节，而无需使用预扫描模板模型。为了解决近距离交互场景中严重的遮挡问题，我们将最近提出的像素对齐隐式函数与参数化模型相结合，用于不可见表面区域的鲁棒重建。设计了一个有效的注意感知模块，用于从多视图图像中获取细粒度的几何细节，从而生成高保真的结果。除了空间注意方法外，对于视频输入，我们进一步提出了一种新的时间融合方法，以缓解噪声和时间不一致性，从而重建运动字符。对于定量评估，我们提供了一个高质量的多人数据集MultiHuman，该数据集由150个具有不同遮挡级别的静态场景和地面真实三维人体模型组成。实验结果表明，我们的方法具有最先进的性能，并且能够很好地推广到真实的多视点视频数据，其性能大大优于以前的工作。

Batch normalization (BN) has been widely used in modern deep neural networks (DNNs) due to improved convergence. BN is observed to increase the model accuracy while at the cost of adversarial robustness. There is an increasing interest in the ML community to understand the impact of BN on DNNs, especially related to the model robustness. This work attempts to understand the impact of BN on DNNs from a non-robust feature perspective. Straightforwardly, the improved accuracy can be attributed to the better utilization of useful features. It remains unclear whether BN mainly favors learning robust features (RFs) or non-robust features (NRFs). Our work presents empirical evidence that supports that BN shifts a model towards being more dependent on NRFs. To facilitate the analysis of such a feature robustness shift, we propose a framework for disentangling robust usefulness into robustness and usefulness. Extensive analysis under the proposed framework yields valuable insight on the DNN behavior regarding robustness, e.g. DNNs first mainly learn RFs and then NRFs. The insight that RFs transfer better than NRFs, further inspires simple techniques to strengthen transfer-based black-box attacks.

批处理规范化 (BN) 由于具有更好的收敛性，在现代深度神经网络 (DNN) 中得到了广泛的应用。观察到BN提高了模型的准确性，同时以对抗性稳健性为代价。ML社区对理解BN对DNN的影响越来越感兴趣，尤其是与模型稳健性相关的问题。这项工作试图从非稳健特征的角度理解BN对DNN的影响。直截了当地说，精度的提高可以归因于对有用特征的更好利用。目前尚不清楚BN主要倾向于学习稳健特征 (RFs) 还是非稳健特征 (NRF)。我们的工作提供了经验证据，支持BN将模型转向更加依赖NRF。为了便于分析这种特征稳健性变化，我们提出了一个框架，将稳健性有用性分解为稳健性和有用性。在所提议的框架下进行的广泛分析产生了关于DNN鲁棒性行为的有价值的见解，例如，DNN首先主要学习RFs，然后学习NRFs。RFs传输优于NRFs这一观点进一步启发了加强基于传输的黑盒攻击的简单技术。

weakly supervised object localization (WSOL) aims to localize objects with only image-level labels, which has better scalability and practicability than fully supervised methods in the actual deployment. However, with only image-level labels, learning object classification models tends to activate object parts and ignore the whole object, while expanding object parts into the whole object may deteriorate classification performance. To alleviate this problem, we propose foreground activation maps (FAM), whose aim is to optimize object localization and classification jointly via an object-aware attention module and a part-aware attention module in a unified model, where the two tasks can complement and enhance each other. To the best of our knowledge, this is the first work that can achieve remarkable performance for both tasks by optimizing them jointly via FAM for WSOL. Besides, the designed two modules can effectively highlight foreground objects for localization and discover discriminative parts for classification. Extensive experiments with four backbones on two standard benchmarks demonstrate that our FAM performs favorably against state-of-the-art WSOL methods.

弱监督目标定位 (WSOL) 的目标是仅使用图像级标签对目标进行定位，在实际部署中比完全监督方法具有更好的可扩展性和实用性。然而，仅使用图像级标签，学习对象分类模型往往激活对象部分而忽略整个对象，而将对象部分扩展到整个对象可能会降低分类性能。为了缓解这一问题，我们提出了前景激活图 (FAM)，其目的是在统一的模型中通过对对象感知注意模块和部分感知注意模块联合优化对象定位和分类，这两个任务可以相互补充和增强。据我们所知，这是第一项通过FAM for WSOL联合优化两项任务，从而实现出色性能的工作。此外，所设计的两个模块可以有效地突出前景对象进行定位，并发现有区别的部分进行分类。在两个标准基准上使用四个主干进行的大量实验表明，我们的FAM相对于最先进的WSOL方法表现良好。

We introduce the first Neural Architecture Search (NAS) method to find a better transformer architecture for image recognition. Recently, transformers without CNN-based backbones are found to achieve impressive performance for image recognition. However, the transformer is designed for NLP tasks and thus could be sub-optimal when directly used for image recognition. In order to improve the visual representation ability for transformers, we propose a new search space and searching algorithm. Specifically, we introduce a locality module that models the local correlations in images explicitly with fewer computational cost. With the locality module, our search space is defined to let the search algorithm freely trade off between global and local information as well as optimizing the low-level design choice in each module. To tackle the problem caused by huge search space, a hierarchical neural architecture search method is proposed to search the optimal vision transformer from two levels separately with the evolutionary algorithm. Extensive experiments on the ImageNet dataset demonstrate that our method can find more discriminative and efficient transformer variants than the ResNet family (e.g., ResNet101) and the baseline ViT for image classification.

我们介绍了第一种神经结构搜索 (NAS) 方法，以找到更好的变压器结构用于图像识别。最近，发现没有基于CNN的主干的变压器在图像识别方面具有令人印象深刻的性能。然而，转换器是为NLP任务设计的，因此直接用于图像识别时可能是次优的。为了提高变压器的视觉表示能力，提出了一种新的搜索空间和搜索算法。具体地说，我们引入了一个局部性模块，该模块以较少的计算量显式地对图像中的局部相关性进行建模。通过局部性模块，我们定义了搜索空间，使搜索算法能够在全局和局部信息之间自由权衡，并优化每个模块中的底层设计选择。针对搜索空间大的问题，提出了一种分层神经结构搜索方法，利用进化算法从两个层次分别搜索最优视觉变换器。在ImageNet数据集上的大量实验表明，我们的方法可以找到比ResNet系列（例如ResNet101）和用于图像分类的基线ViT更具辨别力和效率的变压器变体。

Deep generative models of 3D shapes have received a great deal of research interest. Yet, almost all of them generate discrete shape representations, such as voxels, point clouds, and polygon meshes. We present the first 3D generative model for a drastically different shape representation --- describing a shape as a sequence of computer-aided design (CAD) operations. Unlike meshes and point clouds, CAD models encode the user creation process of 3D shapes, widely used in numerous industrial and engineering design tasks. However, the sequential and irregular structure of CAD operations poses significant challenges for existing 3D generative models. Drawing an analogy between CAD operations and natural language, we propose a CAD generative network based on the Transformer. We demonstrate the performance of our model for both shape autoencoding and random shape generation. To train our network, we create a new CAD dataset consisting of 178,238 models and their CAD construction sequences. We have made this dataset publicly available to promote future research on this topic.

三维形状的深层生成模型已经引起了广泛的研究兴趣。然而，几乎所有这些都会生成离散的形状表示，例如体素、点云和多边形网格。我们提出了第一个完全不同形状表示的3D生成模型——将形状描述为一系列计算机辅助设计 (CAD) 操作。与网格和点云不同，CAD模型对用户创建3D形状的过程进行编码，广泛用于许多工业和工程设计任务。然而，CAD操作的顺序和不规则结构对现有的三维生成模型提出了重大挑战。将CAD操作与自然语言进行类比，提出了一种基于变压器的CAD生成网络。我们展示了我们的模型在形状自动编码和随机形状生成方面的性能。为了训练我们的网络，我们创建了一个由178238个模型及其CAD构造序列组成的新CAD数据集。我们已经公开了这个数据集，以促进这一主题的未来研究。

In this paper, we focus on recognizing 3D shapes from arbitrary views, i.e., arbitrary numbers and positions of viewpoints. It is a challenging and realistic setting for view-based 3D shape recognition. We propose a canonical view representation to tackle this challenge. We first transform the original features of arbitrary views to a fixed number of view features, dubbed canonical view representation, by aligning the arbitrary view features to a set of learnable reference view features using optimal transport. In this way, each 3D shape with arbitrary views is represented by a fixed number of canonical view features, which are further aggregated to generate a rich and robust 3D shape representation for shape recognition. We also propose a canonical view feature separation constraint to enforce that the view features in canonical view representation can be embedded into scattered points in a Euclidean space. Experiments on the ModelNet40, ScanObjectNN, and RGBD datasets show that our method achieves competitive results under the fixed viewpoint settings, and significantly outperforms the applicable methods under the arbitrary view setting.

在本文中，我们着重于从任意视图（即任意数量和位置的视点）识别三维形状。对于基于视图的三维形状识别来说，这是一个具有挑战性和现实性的设置。我们提出了一种规范化的视图表示来应对这一挑战。我们首先将任意视图的原始特征转换为固定数量的视图特征，称为规范视图表示，方法是使用最佳传输将任意视图特征与一组可学习的参考视图特征对齐。通过这种方式，具有任意视图的每个三维形状由固定数量的规范视图特征表示，这些特征进一步聚合以生成用于形状识别的丰富而健壮的三维形状表示。我们还提出了一个规范视图特征分离约束，以强制将规范视图表示中的视图特征嵌入到欧氏空间中的散乱点中。在ModelNet40、ScanObjectNN和RGBD数据集上的实验表明，我们的方法在固定视点设置下取得了有竞争力的结果，并且在任意视图设置下显著优于适用的方法。

Many super-resolution (SR) models are optimized for high performance only and therefore lack efficiency due to large model complexity. As large models are often not practical in real-world applications, we investigate and propose novel loss functions, to enable SR with high perceptual quality from much more efficient models. The representative power for a given low-complexity generator network can only be fully leveraged by strong guidance towards the optimal set of parameters. We show that it is possible to improve the performance of a recently introduced efficient generator architecture solely with the application of our proposed loss functions. In particular, we use a Fourier space supervision loss for improved restoration of missing high-frequency (HF) content from the ground truth image and design a discriminator architecture working directly in the Fourier domain to better match the target HF distribution. We show that our losses' direct emphasis on the frequencies in Fourier-space significantly boosts the perceptual image quality, while at the same time retaining high restoration quality in comparison to previously proposed loss functions for this task. The performance is further improved by utilizing a combination of spatial and frequency domain losses, as both representations provide complementary information during training. On top of that, the trained generator achieves comparable results with and is 2.4x and 48x faster than state-of-the-art perceptual SR methods RankSRGAN and SRFflow respectively.

许多超分辨率 (SR) 模型仅针对高性能进行了优化，因此由于模型复杂度大而缺乏效率。由于大型模型在实际应用中往往不实用，我们研究并提出了新的损失函数，以便从更有效的模型中获得高感知质量的 SR。给定的低复杂度发电机网络的代表性功率只能通过对最佳参数集的有力指导来充分利用。我们表明，仅通过应用我们提出的损耗函数，就有可能提高最近引入的高效发电机结构的性能。特别是，我们使用傅里叶空间监督损耗来改进从地面真值图像中恢复丢失的高频 (HF) 内容，并设计了一种直接在傅里叶域工作的鉴别器结构，以更好地匹配目标HF分布。我们表明，与之前提出的该任务的损失函数相比，我们的损失直接强调傅里叶空间中的频率显著提高了感知图像质量，同时保持了较高的恢复质量。由于两种表示在训练期间提供互补信息，因此通过结合使用空间域和频域损耗，性能得到进一步改善。最重要的是，经过训练的生成器实现了与最先进的感知SR方法RankSRGAN和SRFlow相当的结果，分别比最先进的感知SR方法快2.4倍和48倍。

Vulnerability of 3D point cloud (PC) classifiers has become a grave concern due to the popularity of 3D sensors in safety-critical applications. Existing adversarial attacks against 3D PC classifiers are all test-time evasion (TTE) attacks that aim to induce test-time misclassifications using knowledge of the classifier. But since the victim classifier is usually not accessible to the attacker, the threat is largely diminished in practice, as PC TTEs typically have poor transferability. Here, we propose the first backdoor attack (BA) against PC classifiers. Originally proposed for images, BAs poison the victim classifier's training set so that the classifier learns to decide to the attacker's target class whenever the attacker's backdoor pattern is present in a given input sample. Significantly, BAs do not require knowledge of the victim classifier. Different from image BAs, we propose to insert a cluster of points into a PC as a robust backdoor pattern customized for 3D PCs. Such clusters are also consistent with a physical attack (i.e., with a captured object in a scene). We optimize the cluster's location using an independently trained surrogate classifier and choose the cluster's local geometry to evade possible PC preprocessing and PC anomaly detectors (ADs). Experimentally, our BA achieves a uniformly high success rate ( $\geq 87\%$ ) and shows evasiveness against state-of-the-art PC ADs. Code is available at <https://github.com/zhenxianglance/PCBA>.

由于3D传感器在安全关键应用中的普及，3D点云（PC）分类器的脆弱性已成为一个严重问题。现有针对3D PC分类器的对抗性攻击都是测试时规避（TTE）攻击，其目的是利用分类器的知识导致测试时错误分类。但是，由于攻击者通常无法访问受害者分类器，因此在实践中，威胁会大大降低，因为PC TTE通常具有较差的可转移性。在这里，我们提出了针对PC分类器的第一种后门攻击（BA）。最初是针对图像提出的，BAs会毒害受害者分类器的训练集，以便分类器学会在给定输入样本中存在攻击者的后门模式时决定攻击目标类。值得注意的是，BAs不需要受害者分类器的知识。与图像BAs不同，我们建议将一组点插入PC中，作为为3D PC定制的健壮后门模式。此类集群还与物理攻击（即，场景中捕获的对象）一致。我们使用一个独立训练的代理分类器来优化聚类的位置，并选择聚类的局部几何结构来避免可能的PC预处理和PC异常检测器（ADs）。通过实验，我们的BA获得了一致的高成功率（ $\geq 87\%$ ），并对最先进的PC广告表现出回避态度。代码可在<https://github.com/zhenxianglance/PCBA>.

Rapid progress in 3D semantic segmentation is inseparable from the advances of deep network models, which highly rely on large-scale annotated data for training. To address the high cost and challenges of 3D point-level labeling, we present a method for semi-supervised point cloud semantic segmentation to adopt unlabeled point clouds in training to boost the model performance. Inspired by the recent contrastive loss in self-supervised tasks, we propose the guided point contrastive loss to enhance the feature representation and model generalization ability in semi-supervised setting. Semantic predictions on unlabeled point clouds serve as pseudo-label guidance in our loss to avoid negative pairs in the same category. Also, we design the confidence guidance to ensure high-quality feature learning. Besides, a category-balanced sampling strategy is proposed to collect positive and negative samples to mitigate the class imbalance problem. Extensive experiments on three datasets (ScanNet V2, S3DIS, and SemanticKITTI) show the effectiveness of our semi-supervised method to improve the prediction quality with unlabeled data.

三维语义分割的快速发展与深度网络模型的发展密不可分，深度网络模型高度依赖大规模标注数据进行训练。为了解决三维点级标记的高成本和挑战，我们提出了一种半监督点云语义分割方法，在训练中采用未标记的点云来提高模型性能。受最近在自我监督任务中对比损失的启发，我们提出了引导点对比损失来增强半监督环境下的特征表示和模型泛化能力。未标记点云的语义预测在我们的损失中充当伪标记指导，以避免同一类别中的负对。此外，我们还设计了信心指导，以确保高质量的特征学习。此外，还提出了一种类别平衡抽样策略来收集正样本和负样本，以缓解类别不平衡问题。在三个数据集

(ScanNet V2、S3DIS和SemanticKITTI) 上的大量实验表明，我们的半监督方法可以有效地提高未标记数据的预测质量。

This paper proposes a novel location-aware deep-learning-based single image reflection removal method. Our network has a reflection detection module to regress a probabilistic reflection confidence map, taking multi-scale Laplacian features as inputs. This probabilistic map tells if a region is reflection-dominated or transmission-dominated, and it is used as a cue for the network to control the feature flow when predicting the reflection and transmission layers. We design our network as a recurrent network to progressively refine reflection removal results at each iteration. The novelty is that we leverage Laplacian kernel parameters to emphasize the boundaries of strong reflections. It is beneficial to strong reflection detection and substantially improves the quality of reflection removal results. Extensive experiments verify the superior performance of the proposed method over state-of-the-art approaches. Our code and the pre-trained model can be found at <https://github.com/zdlarr/Location-aware-SIRR>.

提出了一种基于位置感知深度学习的单图像反射去除方法。我们的网络有一个反射检测模块，以多尺度拉普拉斯特征作为输入，回归概率反射置信图。这个概率图告诉我们一个区域是以反射为主还是以透射为主，当预测反射层和透射层时，它被用作网络控制特征流的线索。我们将我们的网络设计为一个循环网络，以在每次迭代中逐步完善反射消除结果。新颖之处在于我们利用拉普拉斯核参数来强调强反射的边界。这有利于强反射检测，并显著提高反射去除结果的质量。大量实验证明了该方法优于现有方法的性能。我们的代码和预先训练的模型可以在<https://github.com/zdlarr/Location-aware-SIRR>.

Test-time augmentation---the aggregation of predictions across transformed versions of a test input---is a common practice in image classification. Traditionally, predictions are combined using a simple average. In this paper, we present 1) experimental analyses that shed light on cases in which the simple average is suboptimal and 2) a method to address these shortcomings. A key finding is that even when test-time augmentation produces a net improvement in accuracy, it can change many correct predictions into incorrect predictions. We delve into when and why test-time augmentation changes a prediction from being correct to incorrect and vice versa. Building on these insights, we present a learning-based method for aggregating test-time augmentations. Experiments across a diverse set of models, datasets, and augmentations show that our method delivers consistent improvements over existing approaches.

测试时间增加——在测试输入的转换版本之间聚合预测——是图像分类中的常见做法。传统上，预测使用简单的平均值进行组合。在本文中，我们提出了1) 实验分析，阐明了简单平均值次优的情况，以及2) 解决这些缺点的方法。一个关键的发现是，即使测试时间的增加产生了准确度的净提高，它也可以将许多正确的预测变成错误的预测。我们深入研究了何时以及为什么测试时间增加会将预测从正确更改为不正确，反之亦然。基于这些见解，我们提出了一种基于学习的方法来聚合测试时间扩展。在一组不同的模型、数据集和扩充中进行的实验表明，我们的方法比现有方法提供了一致的改进。

Neural trees aim at integrating deep neural networks and decision trees so as to bring the best of the two worlds, including representation learning from the former and faster inference from the latter. In this paper, we introduce a novel approach, termed as Self-born Wiring (SeBow), to learn neural trees from a mother deep neural network. In contrast to prior neural-tree approaches that either adopt a pre-defined structure or grow hierarchical layers in a progressive manner, task-adaptive neural trees in SeBow evolve from a deep neural network through a construction-by-destruction process, enabling a global-level parameter optimization that further yields favorable results. Specifically, given a designated network configuration like VGG, SeBow disconnects all the layers and derives isolated filter groups, based on which a global-level wiring process is conducted to attach a subset of filter groups, eventually bearing a lightweight neural tree. Extensive experiments demonstrate that, with a lower computational cost, SeBow outperforms all prior neural trees by a significant margin and even achieves results on par with predominant non-tree networks like ResNets. Moreover, SeBow proves its scalability to large-scale datasets like ImageNet, which has been barely explored by prior tree networks.

神经树的目标是将深层神经网络和决策树结合起来，从而实现两个世界的最佳结合，包括前者的表示学习和后者的快速推理。在本文中，我们介绍了一种新的方法，称为自生布线（SeBoW），从母体深层神经网络学习神经树。与先前采用预定义结构或以渐进方式增长层次结构的神经树方法不同，SeBoW中的任务自适应神经树通过分解构建过程从深层神经网络进化而来，实现了全局级参数优化，从而进一步产生了有利的结果。具体地说，给定一个指定的网络配置（如VGG），SeBoW断开所有层并导出隔离的过滤器组，在此基础上执行全局级布线过程以连接过滤器组的子集，最终生成一个轻量级神经树。大量的实验表明，在较低的计算成本下，SeBoW比所有先验神经树都有显著的裕度，甚至达到了与ReNETs等非树型网络一样的结果。此外，SeBoW还证明了其对大规模数据集（如ImageNet）的可扩展性，而以前的树网络几乎没有对其进行过研究。

Nowadays, there is an abundance of data involving images and surrounding free-form text weakly corresponding to those images. Weakly Supervised phrase-Grounding (WSG) deals with the task of using this data to learn to localize (or to ground) arbitrary text phrases in images without any additional annotations. However, most recent SotA methods for WSG assume an existence of a pre-trained object detector, relying on it to produce the ROIs for localization. In this work, we focus on the task of Detector-Free WSG (DF-WSG) to solve WSG without relying on a pre-trained detector. We directly learn everything from the images and associated free-form text pairs, thus potentially gaining advantage on the categories unsupported by the detector. The key idea behind our proposed Grounding by Separation (GbS) method is synthesizing 'text to image-regions' associations by random alpha-blending of arbitrary image pairs and using the corresponding texts of the pair as conditions to recover the alpha map from the blended image via a segmentation network. At test time, this allows using the query phrase as a condition for a non-blended query image, thus interpreting the test image as a composition of a region corresponding to the phrase and the complement region. Using this approach we demonstrate a significant accuracy improvement, up to 8.5% over previous DF-WSG SotA, for a range of benchmarks including Flickr30K, Visual Genome, and ReferIt, as well as a significant complementary improvement (above 7%) over the detector-based approaches for WSG.

如今，有大量的数据涉及图像和周围的自由格式文本，这些文本与这些图像弱对应。弱监督短语接地（WSG）处理的任务是使用这些数据来学习在图像中定位（或接地）任意文本短语，而无需任何附加注释。然而，最近用于WSG的SotA方法假设存在一个预先训练的目标检测器，依靠它产生ROI进行定位。在这项工作中，我们专注于无检测器WSG（DF-WSG）的任务，以解决WSG问题，而不依赖预先训练的检测器。我们直接从图像和相关的自由形式文本对中学习所有内容，从而潜在地获得检测器不支持的类别的优势。我们提出的分离接地（GbS）方法背后的关键思想是通过任意图像对的随机alpha混合合成

“文本到图像区域”关联，并使用该图像对的对应文本作为条件，通过分割网络从混合图像恢复alpha映射。在测试时，这允许使用查询短语作为非混合查询图像的条件，从而将测试图像解释为与短语和补码区域相对应的区域的组成。使用这种方法，我们证明了一系列基准（包括Flickr30K、Visual Genome和ReferIt）的准确度显著提高，比以前的DF-WSG SotA高达8.5%，并且与基于检测器的WSG方法相比，具有显著的互补性提高（超过7%）。

Contrastive learning allows us to flexibly define powerful losses by contrasting positive pairs from sets of negative samples. Recently, the principle has also been used to learn cross-modal embeddings for video and text, yet without exploiting its full potential. In particular, previous losses do not take the intra-modality similarities into account, which leads to inefficient embeddings, as the same content is mapped to multiple points in the embedding space. With CrossCLR, we present a contrastive loss that fixes this issue. Moreover, we define sets of highly related samples in terms of their input embeddings and exclude them from the negative samples to avoid issues with false negatives. We show that these principles consistently improve the quality of the learned embeddings. The joint embeddings learned with CrossCLR extend the state of the art in video-text retrieval on Youcook2 and LSMDC datasets and in video captioning on the Youcook2 dataset by a large margin. We also demonstrate the generality of the concept by learning improved joint embeddings for other pairs of modalities.

对比学习允许我们通过对一组负样本中的正对，灵活地定义强大的损失。最近，该原理也被用于学习视频和文本的跨模态嵌入，但没有充分发挥其潜力。特别是，以前的损失没有考虑模态内的相似性，这导致了低效的嵌入，因为相同的内容映射到嵌入空间中的多个点。使用CrossCLR，我们提供了一个对比损失，解决了这个问题。此外，我们根据输入嵌入定义了高度相关的样本集，并将其从否定样本中排除，以避免出现假否定问题。我们表明，这些原则不断提高学习嵌入的质量。通过CrossCLR学习到的联合嵌入大大扩展了Youcook2和LSMDC数据集上视频文本检索以及Youcook2数据集上视频字幕的最新技术。我们还通过学习其他模式对的改进的联合嵌入来证明该概念的通用性。

This paper proposes a novel deep learning approach for single image defocus deblurring based on inverse kernels. In a defocused image, the blur shapes are similar among pixels although the blur sizes can spatially vary. To utilize the property with inverse kernels, we exploit the observation that when only the size of a defocus blur changes while keeping the shape, the shape of the corresponding inverse kernel remains the same and only the scale changes. Based on the observation, we propose a kernel-sharing parallel atrous convolutional (KPAC) block specifically designed by incorporating the property of inverse kernels for single image defocus deblurring. To effectively simulate the invariant shapes of inverse kernels with different scales, KPAC shares the same convolutional weights among multiple atrous convolution layers. To efficiently simulate the varying scales of inverse kernels, KPAC consists of only a few atrous convolution layers with different dilations and learns per-pixel scale attentions to aggregate the outputs of the layers. KPAC also utilizes the shape attention to combine the outputs of multiple convolution filters in each atrous convolution layer, to deal with defocus blur with a slightly varying shape. We demonstrate that our approach achieves state-of-the-art performance with a much smaller number of parameters than previous methods.

提出了一种基于逆核的单幅图像离焦去模糊深度学习方法。在离焦图像中，像素之间的模糊形状相似，尽管模糊大小可以在空间上变化。为了利用逆核的特性，我们利用了这样一个观察结果：在保持形状的同时，只有散焦模糊的大小发生变化，相应逆核的形状保持不变，只有尺度发生变化。基于观察结果，我们提出了一种核共享并行阿托斯卷积（KPAC）块，该块通过结合逆核的特性专门设计用于单图像散焦去模糊。为了有效地模拟具有不同尺度的逆核的不变形状，KPAC在多个Atrous卷积层之间共享相同的卷积权重。为了有效地模拟不同尺度的逆核，KPAC仅由几个具有不同膨胀度的阿托斯卷积层组成，并学习

每像素尺度的注意来聚合层的输出。KPAC还利用形状注意来组合每个Atrus卷积层中多个卷积滤波器的输出，以处理形状略有变化的散焦模糊。我们证明，与以前的方法相比，我们的方法在参数数量少得多的情况下实现了最先进的性能。

Compressive imaging using coded apertures (CA) is a powerful technique that can be used to recover depth, light fields, hyperspectral images and other quantities from a single snapshot. The performance of compressive imaging systems based on CAs mostly depends on two factors: the properties of the mask's attenuation pattern, that we refer to as "codification", and the computational techniques used to recover the quantity of interest from the coded snapshot. In this work, we introduce the idea of using time-varying CAs synchronized with spatially varying pixel shutters. We divide the exposure of a sensor into sub-exposures at the beginning of which the CA mask changes and at which the sensor's pixels are simultaneously and individually switched "on" or "off". This is a practically appealing codification as it does not introduce additional optical components other than the already present CA but uses a change in the pixel shutter that can be easily realized electronically. We show that our proposed time-multiplexed coded aperture (TMCA) can be optimized end to end and induces better coded snapshots enabling superior reconstructions in two different applications: compressive light field imaging and hyperspectral imaging. We demonstrate both in simulation and with real captures (taken with prototypes we built) that this codification outperforms the state-of-the-art compressive imaging systems by a large margin in those applications.

使用编码孔径 (CA) 的压缩成像是一种强大的技术，可用于从单个快照恢复深度、光场、高光谱图像和其他数量。基于CAs的压缩成像系统的性能主要取决于两个因素：我们称之为“编码”的掩模衰减模式的特性，以及用于从编码快照中恢复感兴趣数量的计算技术。在这项工作中，我们介绍了使用与空间变化像素快门同步的时变CAs的思想。我们将传感器的曝光分为子曝光，在子曝光开始时CA遮罩会发生变化，在子曝光开始时，传感器的像素会同时单独地“打开”或“关闭”。这实际上是一个很有吸引力的编码，因为它没有引入除现有CA之外的其他光学组件，而是使用了像素快门的变化，可以轻松地通过电子方式实现。我们表明，我们提出的时间复用编码孔径 (TMCA) 可以端到端进行优化，并产生更好的编码快照，从而在压缩光场成像和高光谱成像两种不同的应用中实现更好的重建。我们通过模拟和真实捕获（使用我们构建的原型）证明，这种编码在这些应用中大大优于最先进的压缩成像系统。

We propose a hierarchical graph neural network (GNN) model that learns how to cluster a set of images into an unknown number of identities using a training set of images annotated with labels belonging to a disjoint set of identities. Our hierarchical GNN uses a novel approach to merge connected components predicted at each level of the hierarchy to form a new graph at the next level. Unlike fully unsupervised hierarchical clustering, the choice of grouping and complexity criteria stems naturally from supervision in the training set. The resulting method, Hi-LANDER, achieves an average of 49% improvement in F-score and 7% increase in Normalized Mutual Information (NMI) relative to current GNN-based clustering algorithms. Additionally, state-of-the-art GNN-based methods rely on separate models to predict linkage probabilities and node densities as intermediate steps of the clustering process. In contrast, our unified framework achieves a three-fold decrease in computational cost. Our training and inference code are released.

我们提出了一个层次图神经网络 (GNN) 模型，该模型学习如何使用带有不相交身份集标签的图像训练集将一组图像聚类为未知数量的身份。我们的分层GNN使用一种新的方法来合并在层次结构的每一层预测的连接组件，以在下一层形成一个新的图。与完全无监督的分层聚类不同，分组和复杂性标准的选择自然来自于训练集中的监督。与当前基于GNN的聚类算法相比，Hi-LANDER方法的F分数平均提高49%，归一化互信息 (NMI) 平均提高7%。此外，最先进的基于GNN的方法依赖于单独的模型来预测连

锁概率和节点密度，作为聚类过程的中间步骤。相比之下，我们的统一框架实现了计算成本的三倍降低。我们的训练和推理代码发布了。

Real-world low-light images suffer from two main degradations, namely, inevitable noise and poor visibility. Since the noise exhibits different levels, its estimation has been implemented in recent works when enhancing low-light images from raw Bayer space. When it comes to sRGB color space, the noise estimation becomes more complicated due to the effect of the image processing pipeline. Nevertheless, most existing enhancing algorithms in sRGB space only focus on the low visibility problem or suppress the noise under a hypothetical noise level, leading them impractical due to the lack of robustness. To address this issue, we propose an adaptive unfolding total variation network (UTVNet), which approximates the noise level from the real sRGB low-light image by learning the balancing parameter in the model-based denoising method with total variation regularization. Meanwhile, we learn the noise level map by unrolling the corresponding minimization process for providing the inferences of smoothness and fidelity constraints. Guided by the noise level map, our UTVNet can recover finer details and is more capable to suppress noise in real captured low-light scenes. Extensive experiments on real-world low-light images clearly demonstrate the superior performance of UTVNet over state-of-the-art methods.

现实世界中的微光图像存在两种主要退化，即不可避免的噪声和低可见度。由于噪声表现出不同的水平，在最近的工作中，当从原始拜耳空间增强弱光图像时，已经实现了对噪声的估计。对于sRGB颜色空间，由于图像处理管道的影响，噪声估计变得更加复杂。然而，现有的sRGB空间增强算法大多只针对低可见度问题或在假设的噪声水平下抑制噪声，由于缺乏鲁棒性而不切实际。为了解决这个问题，我们提出了一种自适应展开全变分网络 (UTVNet) ，该网络通过学习基于模型的全变分正则化去噪方法中的平衡参数来逼近真实sRGB微光图像的噪声级。同时，我们通过展开相应的最小化过程来学习噪声级映射，以提供平滑度和保真度约束的推论。在噪声级图的指导下，我们的UTVNet可以恢复更精细的细节，并且更能够在实际捕获的微光场景中抑制噪声。在现实世界的微光图像上进行的大量实验清楚地表明，UTVNet的性能优于最先进的方法。

we show that relation modeling between visual elements matters in cropping view recommendation. Cropping view recommendation addresses the problem of image recomposition conditioned on the composition quality and the ranking of views (cropped sub-regions). This task is challenging because the visual difference is subtle when a visual element is reserved or removed. Existing methods represent visual elements by extracting region-based convolutional features inside and outside the cropping view boundaries, without probing a fundamental question: why some visual elements are of interest or of discard? In this work, we observe that the relation between different visual elements significantly affects their relative positions to the desired cropping view, and such relation can be characterized by the attraction inside/outside the cropping view boundaries and the repulsion across the boundaries. By instantiating a transformer-based solution that represents visual elements as visual words and that models the dependencies between visual words, we report not only state of-the-art performance on public benchmarks, but also interesting visualizations that depict the attraction and repulsion between visual elements, which may shed light on what makes for effective cropping view recommendation.

我们表明，在裁剪视图推荐中，视觉元素之间的关系建模很重要。裁剪视图推荐解决了基于构图质量和视图排序（裁剪子区域）的图像重新排序问题。这项任务很有挑战性，因为当保留或删除某个视觉元素时，视觉差异很细微。现有的方法通过在裁剪视图边界内外提取基于区域的卷积特征来表示视觉元素，而没有探究一个基本问题：为什么某些视觉元素感兴趣或被丢弃？在这项工作中，我们观察到不同视觉元素之间的关系显著影响它们相对于所需裁剪视图的相对位置，并且这种关系可以通过裁剪视图边界内外的吸引和边界之间的排斥来表征。通过实例化一个基于转换器的解决方案，该解决方案将视觉元素表

示为视觉词，并对视觉词之间的依赖关系进行建模，我们不仅报告了公共基准上的最新表现，还报告了描述视觉元素之间吸引和排斥的有趣的视觉效果，这可能有助于了解如何进行有效的裁剪视图推荐。

We present a novel framework for mesh reconstruction from unstructured point clouds by taking advantage of the learned visibility of the 3D points in the virtual views and traditional graph-cut based mesh generation. Specifically, we first propose a three-step network that explicitly employs depth completion for visibility prediction. Then the visibility information of multiple views is aggregated to generate a 3D mesh model by solving an optimization problem considering visibility in which a novel adaptive visibility weighting term in surface determination is also introduced to suppress line of sight with a large incident angle. Compared to other learning-based approaches, our pipeline only exercises the learning on a 2D binary classification task, i.e., points visible or not in a view, which is much more generalizable and practically more efficient and capable to deal with a large number of points. Experiments demonstrate that our method with favorable transferability and robustness, and achieve competing performances w.r.t. state-of-the-art learning-based approaches on small complex objects and outperforms on large indoor and outdoor scenes. Code is available at <https://github.com/GDAOSU/vis2mesh>.

我们提出了一种新的非结构化点云网格重建框架，利用虚拟视图中三维点的可见性和传统的基于图割的网格生成。具体地说，我们首先提出了一个三步网络，该网络明确采用深度补全进行可见性预测。然后，通过解决考虑可见性的优化问题，将多个视图的可见性信息聚合生成三维网格模型，在曲面确定中引入了一种新的自适应可见性权重项，以抑制具有大入射角的视线。与其他基于学习的方法相比，我们的管道仅在2D二元分类任务上进行学习，即在视图中可见或不可见的点，这更具普遍性，实际上更有效，能够处理大量点。实验表明，我们的方法具有良好的可转移性和鲁棒性，并且在小型复杂对象上实现了与当前最先进的基于学习的方法相媲美的性能，并且在大型室内和室外场景上的性能优于传统的基于学习的方法。代码可在<https://github.com/GDAOSU/vis2mesh>。

Adversarial training (AT) has become the de-facto standard to obtain models robust against adversarial examples. However, AT exhibits severe robust overfitting: cross-entropy loss on adversarial examples, so-called robust loss, decreases continuously on training examples, while eventually increasing on test examples. In practice, this leads to poor robust generalization, i.e., adversarial robustness does not generalize well to new examples. In this paper, we study the relationship between robust generalization and flatness of the robust loss landscape in weight space, i.e., whether robust loss changes significantly when perturbing weights. To this end, we propose average- and worst-case metrics to measure flatness in the robust loss landscape and show a correlation between good robust generalization and flatness. For example, throughout training, flatness reduces significantly during overfitting such that early stopping effectively finds flatter minima in the robust loss landscape. Similarly, AT variants achieving higher adversarial robustness also correspond to flatter minima. This holds for many popular choices, e.g., AT-AWP, TRADES, MART, AT with self-supervision or additional unlabeled examples, as well as simple regularization techniques, e.g., AutoAugment, weight decay or label noise. For fair comparison across these approaches, our flatness measures are specifically designed to be scale-invariant and we conduct extensive experiments to validate our findings.

对抗性训练 (AT) 已成为获取对抗性示例的健壮模型的事实标准。然而，AT表现出严重的稳健过度拟合：对抗性示例上的交叉熵损失，即所谓的稳健损失，在训练示例上不断减少，而在测试示例上最终增加。在实践中，这会导致较差的鲁棒性推广，即对抗性鲁棒性不能很好地推广到新示例。在本文中，我们研究了鲁棒泛化与权重空间中鲁棒损失图的平坦性之间的关系，即当扰动权重时，鲁棒损失是否显著变化。为此，我们提出了平均和最坏情况下的度量标准来衡量稳健损失情况下的平坦度，并展示了良好

稳健泛化与平坦度之间的相关性。例如，在整个训练过程中，平坦度在过度拟合过程中显著降低，因此提前停止可以有效地在稳健的损失环境中找到更平坦的最小值。同样，获得更高对抗鲁棒性的AT变体也对应于更平坦的最小值。这适用于许多流行的选择，例如AT-AWP、贸易、MART、具有自我监督或附加未标记示例的AT，以及简单的正则化技术，例如自动增强、重量衰减或标签噪声。为了在这些方法中进行公平比较，我们的平面度测量专门设计为尺度不变，我们进行了大量实验以验证我们的发现。

Most video super-resolution methods focus on restoring high-resolution video frames from low-resolution videos without taking into account compression. However, most videos on the web or mobile devices are compressed, and the compression can be severe when the bandwidth is limited. In this paper, we propose a new compression-informed video super-resolution model to restore high-resolution content without introducing artifacts caused by compression. The proposed model consists of three modules for video super-resolution: bi-directional recurrent warping, detail-preserving flow estimation, and Laplacian enhancement. All these three modules are used to deal with compression properties such as the location of the intra-frames in the input and smoothness in the output frames. For thorough performance evaluation, we conducted extensive experiments on standard datasets with a wide range of compression rates, covering many real video use cases. We showed that our method not only recovers high-resolution content on uncompressed frames from the widely-used benchmark datasets, but also achieves state-of-the-art performance in super-resolving compressed videos based on numerous quantitative metrics. We also evaluated the proposed method by simulating streaming from YouTube to demonstrate its effectiveness and robustness. The source codes and trained models are available at <https://github.com/google-research/google-research/tree/master/comisr>.

大多数视频超分辨率方法侧重于从低分辨率视频恢复高分辨率视频帧，而不考虑压缩。但是，web或移动设备上的大多数视频都是压缩的，当带宽有限时，压缩可能会很严重。在本文中，我们提出了一种新的基于压缩的视频超分辨率模型来恢复高分辨率内容，而不引入压缩引起的伪影。该模型由三个视频超分辨率模块组成：双向循环扭曲、细节保持流估计和拉普拉斯增强。所有这三个模块都用于处理压缩特性，例如输入帧中帧内的位置和输出帧中的平滑度。为了进行全面的性能评估，我们在标准数据集上进行了广泛的实验，具有广泛的压缩率，涵盖了许多真实的视频用例。我们表明，我们的方法不仅可以从广泛使用的基准数据集中恢复未压缩帧上的高分辨率内容，而且还可以在基于大量量化指标的超分辨率压缩视频中实现最先进的性能。我们还通过模拟YouTube上的流媒体来评估所提出的方法，以证明其有效性和鲁棒性。源代码和经过培训的模型可在<https://github.com/google-research/google-research/tree/master/comisr>。

Mixed-precision networks allow for a variable bit-width quantization for every layer in the network. A major limitation of existing work is that the bit-width for each layer must be predefined during training time. This allows little flexibility if the characteristics of the device on which the network is deployed change during runtime. In this work, we propose Bit-Mixer, the very first method to train a meta-quantized network where during test time any layer can change its bit-width without affecting at all the overall network's ability for highly accurate inference. To this end, we make 2 key contributions: (a) Transitional Batch-Norms, and (b) a 3-stage optimization process which is shown capable of training such a network. We show that our method can result in mixed precision networks that exhibit the desirable flexibility properties for on-device deployment without compromising accuracy. Code will be made available.

混合精度网络允许对网络中的每一层进行可变比特宽度量化。现有工作的一个主要限制是必须在训练期间预定义每个层的位宽度。如果部署网络的设备的特性在运行时发生变化，那么这就没有多少灵活性。在这项工作中，我们提出了位混合器，这是训练元量化网络的第一种方法，在测试期间，任何层都可以改变其位宽度，而不会影响整个网络进行高精度推理的能力。为此，我们做出了两个关键贡献：(a) 过

渡批次规范，和 (b) 三阶段优化过程，该过程被证明能够训练这样的网络。我们表明，我们的方法可以产生混合精度网络，在不影响精度的情况下，在设备部署中表现出所需的灵活性。代码将可用。

The application of light field data in salient object detection is becoming increasingly popular in recent years. The difficulty lies in how to effectively fuse the features within the focal stack and how to cooperate them with the feature of the all-focus image. Previous methods usually fuse focal stack features via convolution or ConvLSTM, which are both less effective and ill-posed. In this paper, we model the information fusion within focal stack via graph networks. They introduce powerful context propagation from neighbouring nodes and also avoid ill-posed implementations. On the one hand, we construct local graph connections thus avoiding prohibitive computational costs of traditional graph networks. On the other hand, instead of processing the two kinds of data separately, we build a novel dual graph model to guide the focal stack fusion process using all-focus patterns. To handle the second difficulty, previous methods usually implement one-shot fusion for focal stack and all-focus features, hence lacking a thorough exploration of their supplements. We introduce a reciprocative guidance scheme and enable mutual guidance between these two kinds of information at multiple steps. As such, both kinds of features can be enhanced iteratively, finally benefiting the saliency prediction. Extensive experimental results show that the proposed models are all beneficial and we achieve significantly better results than state-of-the-art methods.

近年来，光场数据在显著目标检测中的应用越来越广泛。难点在于如何有效地融合焦堆中的特征，以及如何将它们与全聚焦图像的特征相结合。以往的方法通常通过卷积或ConvLSTM融合焦堆特征，这两种方法都效率较低且不适用。在本文中，我们通过图网络建立了焦堆内的信息融合模型。它们引入了来自相邻节点的强大上下文传播，还避免了不适用实现。一方面，我们构造局部图连接，从而避免了传统图网络难以承受的计算开销。另一方面，我们建立了一种新的双图模型，以指导使用所有聚焦模式的聚焦叠加融合过程，而不是单独处理这两种数据。为了解决第二个难题，以前的方法通常对焦点堆栈和所有焦点特征进行一次融合，因此缺乏对其补充的深入研究。我们引入了一种交互引导方案，并在多个步骤中实现这两种信息之间的相互引导。因此，这两种特征都可以迭代增强，最终有利于显著性预测。大量的实验结果表明，所提出的模型都是有益的，并且我们取得了比最先进的方法更好的结果。

Interpreting the decision logic behind effective deep convolutional neural networks (CNN) on images complements the success of deep learning models. However, the existing methods can only interpret some specific decision logic on individual or a small number of images. To facilitate human understandability and generalization ability, it is important to develop representative interpretations that interpret common decision logics of a CNN on a large group of similar images, which reveal the common semantics data contributes to many closely related predictions. In this paper, we develop a novel unsupervised approach to produce a highly representative interpretation for a large number of similar images. We formulate the problem of finding representative interpretations as a co-clustering problem, and convert it into a submodular cost submodular cover problem based on a sample of the linear decision boundaries of a CNN. We also present a visualization and similarity ranking method. Our extensive experiments demonstrate the excellent performance of our method.

在图像上解释有效的深度卷积神经网络 (CNN) 背后的决策逻辑是对深度学习模型成功的补充。然而，现有的方法只能对单个或少量图像解释某些特定的决策逻辑。为了促进人类的可理解性和泛化能力，重要的是开发具有代表性的解释，在大量相似图像上解释CNN的共同决策逻辑，这揭示了共同语义数据有助于许多密切相关的预测。在本文中，我们开发了一种新的无监督方法，为大量相似的图像生成具有高度代表性的解释。我们将寻找代表性解释的问题描述为一个共聚类问题，并将其转化为基于CNN线性决策边界样本的子模代价子模覆盖问题。我们还提出了一种可视化和相似性排序方法。我们的大量实验证明了我们的方法的优异性能。

Recently, some approaches are proposed to harness deep convolutional networks to facilitate superpixel segmentation. The common practice is to first evenly divide the image into a pre-defined number of grids and then learn to associate each pixel with its surrounding grids. However, simply applying a series of convolution operations with limited receptive fields can only implicitly perceive the relations between the pixel and its surrounding grids. Consequently, existing methods often fail to provide an effective context when inferring the association map. To remedy this issue, we propose a novel Association Implantation (AI) module to enable the network to explicitly capture the relations between the pixel and its surrounding grids. The proposed AI module directly implants the features of grid cells to the surrounding of its corresponding central pixel, and conducts convolution on the padded window to adaptively transfer knowledge between them. With such an implantation operation, the network could explicitly harvest the pixel-grid level context, which is more in line with the target of superpixel segmentation comparing to the pixel-wise relation. Furthermore, to pursue better boundary precision, we design a boundary-perceiving loss to help the network discriminate the pixels around boundaries in hidden feature level, which could benefit the subsequent inferring modules to accurately identify more boundary pixels. Extensive experiments on BSDS500 and NYUV2 datasets show that our method could not only achieve state-of-the-art performance but maintain satisfactory inference efficiency. Code and pre-trained model are available at <https://github.com/wangyxxjtu/AINet-ICCV2021>.

最近，人们提出了一些利用深度卷积网络进行超像素分割的方法。通常的做法是首先将图像均匀地划分为预定义数量的网格，然后学习将每个像素与其周围的网格相关联。然而，简单地应用一系列具有有限感受野的卷积运算只能隐式地感知像素与其周围网格之间的关系。因此，现有的方法在推断关联映射时往往无法提供有效的上下文。为了解决这个问题，我们提出了一种新的关联植入（AI）模块，使网络能够显式捕获像素与其周围网格之间的关系。该人工智能模块将网格单元的特征直接植入其对应的中心像素周围，并在填充窗口上进行卷积以自适应地在网格单元之间传递知识。通过这种植入操作，网络可以显式地获取像素网格级上下文，与像素级关系相比，更符合超像素分割的目标。此外，为了追求更好的边界精度，我们设计了一种边界感知损失，以帮助网络在隐藏特征层识别边界周围的像素，这有助于后续推理模块准确识别更多的边界像素。在BSDS500和NYUV2数据集上的大量实验表明，我们的方法不仅可以达到最先进的性能，而且可以保持令人满意的推理效率。代码和预先培训的模型可在<https://github.com/wangyxxjtu/AINet-ICCV2021>。

Event cameras are ideally suited to capture HDR visual information without blur but perform poorly on static or slowly changing scenes. Conversely, conventional image sensors measure absolute intensity of slowly changing scenes effectively but do poorly on high dynamic range or quickly changing scenes. In this paper, we present an event-based video reconstruction pipeline for High Dynamic Range (HDR) scenarios. The proposed algorithm includes a frame augmentation pre-processing step that deblurs and temporally interpolates frame data using events. The augmented frame and event data are then fused using a novel asynchronous Kalman filter under a unifying uncertainty model for both sensors. Our experimental results are evaluated on both publicly available datasets with challenging lighting conditions and fast motions and our new dataset with HDR reference. The proposed algorithm outperforms state-of-the-art methods in both absolute intensity error (48% reduction) and image similarity indexes (average 11% improvement).

事件摄影机非常适合捕捉HDR视觉信息而不产生模糊，但在静态或缓慢变化的场景中表现不佳。相反，传统的图像传感器可以有效地测量缓慢变化的场景的绝对强度，但在高动态范围或快速变化的场景中效果不佳。在本文中，我们提出了一种基于事件的高动态范围（HDR）场景视频重建管道。该算法包括一个帧增强预处理步骤，该步骤使用事件对帧数据进行去模糊和时间插值。然后，在两个传感器的统一不确定性模型下，使用一种新的异步卡尔曼滤波器对增强的帧和事件数据进行融合。我们的实验结果在具

有挑战性的光照条件和快速运动的公开数据集上进行了评估，并且我们的新数据集具有HDR参考。该算法在绝对亮度误差（减少48%）和图像相似性指数（平均提高11%）方面均优于现有的方法。

Though recent years have witnessed remarkable progress in single image super-resolution (SISR) tasks with the prosperous development of deep neural networks (DNNs), the deep learning methods are confronted with the computation and memory consumption issues in practice, especially for resource-limited platforms such as mobile devices. To overcome the challenge and facilitate the real-time deployment of SISR tasks on mobile, we combine neural architecture search with pruning search and propose an automatic search framework that derives sparse super-resolution (SR) models with high image quality while satisfying the real-time inference requirement. To decrease the search cost, we leverage the weight sharing strategy by introducing a supernet and decouple the search problem into three stages, including supernet construction, compiler-aware architecture and pruning search, and compiler-aware pruning ratio search. With the proposed framework, we are the first to achieve real-time SR inference (with only tens of milliseconds per frame) for implementing 720p resolution with competitive image quality (in terms of PSNR and SSIM) on mobile platforms (Samsung Galaxy S20).

尽管近年来随着深度神经网络 (DNN) 的蓬勃发展，单图像超分辨率 (SISR) 任务取得了显著进展，但深度学习方法在实际应用中面临着计算和内存消耗问题，特别是对于资源有限的平台，如移动设备。为了克服这一挑战并便于在移动设备上实时部署SISR任务，我们将神经结构搜索与剪枝搜索相结合，提出了一种自动搜索框架，该框架在满足实时推理要求的同时，导出具有高图像质量的稀疏超分辨率 (SR) 模型。为了降低搜索成本，我们通过引入超网来利用权重共享策略，并将搜索问题分解为三个阶段，包括超网构造、编译器感知体系结构和剪枝搜索以及编译器感知剪枝比率搜索。利用所提出的框架，我们是第一个实现实时SR推断（每帧仅数十毫秒）的公司，用于在移动平台（三星Galaxy S20）上实现720p分辨率和具有竞争力的图像质量（PSNR和SSIM）。

The goal of few-shot learning (FSL) is to recognize a set of novel classes with only few labeled samples by exploiting a large set of abundant base class samples. Adopting a meta-learning framework, most recent FSL methods meta-learn a deep feature embedding network, and during inference classify novel class samples using nearest neighbor in the learned high-dimensional embedding space. This means that these methods are prone to the hubness problem, that is, a certain class prototype becomes the nearest neighbor of many test instances regardless which classes they belong to. However, this problem is largely ignored in existing FSL studies. In this work, for the first time we show that many FSL methods indeed suffer from the hubness problem. To mitigate its negative effects, we further propose to employ z-score feature normalization, a simple yet effective transformation, during meta-training. A theoretical analysis is provided on why it helps. Extensive experiments are then conducted to show that with z-score normalization, the performance of many recent FSL methods can be boosted, resulting in new state-of-the-art on three benchmarks.

少数镜头学习 (FSL) 的目标是通过利用大量丰富的基类样本来识别一组只有少量标记样本的新类。最新的FSL方法采用元学习框架，元学习深度特征嵌入网络，并在推理过程中使用学习到的高维嵌入空间中的最近邻对新类样本进行分类。这意味着这些方法容易出现hubness问题，也就是说，某个类原型成为许多测试实例的最近邻，而不管它们属于哪个类。然而，在现有的FSL研究中，这一问题在很大程度上被忽略了。在这项工作中，我们第一次表明，许多FSL方法确实受到Hubness问题的影响。为了减轻其负面影响，我们进一步建议在元训练期间使用z-score特征规范化，这是一种简单但有效的转换。本文从理论上分析了它的作用。大量的实验表明，通过z-score归一化，可以提高许多最近的FSL方法的性能，从而在三个基准上实现新的最新水平。

Video-based person re-identification (re-ID) aims at matching the same person across video clips. Efficiently exploiting multi-scale fine-grained features while building the structural interaction among them is pivotal for its success. In this paper, we propose a hybrid framework, Dense Interaction Learning (DenseIL), that takes the principal advantages of both CNN-based and Attention-based architectures to tackle video-based person re-ID difficulties. DenseIL contains a CNN encoder and a Dense Interaction (DI) decoder. The CNN encoder is responsible for efficiently extracting discriminative spatial features while the DI decoder is designed to densely model spatial-temporal inherent interaction across frames. Different from previous works, we additionally let the DI decoder densely attends to intermediate fine-grained CNN features and that naturally yields multi-grained spatial-temporal representation for each video clip. Moreover, we introduce Spatio-TEmporal Positional Embedding (STEP-Emb) into the DI decoder to investigate the positional relation among the spatial-temporal inputs. Our experiments consistently and significantly outperform all the state-of-the-art methods on multiple standard video-based person re-ID datasets.

基于视频的人员重新识别 (re-ID) 旨在跨视频片段匹配同一个人。有效地利用多尺度细粒度特征，同时构建它们之间的结构交互是其成功的关键。在本文中，我们提出了一个混合框架，密集交互学习 (DenseIL)，它利用了基于CNN和基于注意的体系结构的主要优点来解决基于视频的人物识别困难。DenseIL包含CNN编码器和密集交互 (DI) 解码器。CNN编码器负责有效地提取有区别的空间特征，而DI解码器设计用于对跨帧的时空固有交互进行密集建模。与以前的工作不同，我们还让DI解码器密集地关注中间细粒度CNN特征，这自然会产生每个视频剪辑的多粒度时空表示。此外，我们将时空位置嵌入 (STEP-Emb) 引入到DI解码器中，以研究时空输入之间的位置关系。我们的实验在基于多个标准视频的person-re-ID数据集上始终显著优于所有最先进的方法。

Existing algorithms for explaining the output of image classifiers perform poorly on inputs where the object of interest is partially occluded. We present a novel, black-box algorithm for computing explanations that uses a principled approach based on causal theory. We have implemented the method in the DeepCover tool. We obtain explanations that are much more accurate than those generated by the existing explanation tools on images with occlusions and observe a level of performance comparable to the state of the art when explaining images without occlusions.

现有的用于解释图像分类器输出的算法在感兴趣的对象被部分遮挡的输入上表现不佳。我们提出了一种新的黑盒算法来计算解释，它使用基于因果理论的原则性方法。我们已经在DeepCover工具中实现了该方法。我们获得的解释比现有解释工具在有遮挡的图像上生成的解释准确得多，并且在解释无遮挡的图像时观察到与最新技术相当的性能水平。

It is widely acknowledged that single image super-resolution (SISR) methods would not perform well if the assumed degradation model deviates from those in real images. Although several degradation models take additional factors into consideration, such as blur, they are still not effective enough to cover the diverse degradations of real images. To address this issue, this paper proposes to design a more complex but practical degradation model that consists of randomly shuffled blur, downsampling and noise degradations. Specifically, the blur is approximated by two convolutions with isotropic and anisotropic Gaussian kernels; the downsampling is randomly chosen from nearest, bilinear and bicubic interpolations; the noise is synthesized by adding Gaussian noise with different noise levels, adopting JPEG compression with different quality factors, and generating processed camera sensor noise via reverse-forward camera image signal processing (ISP) pipeline model and RAW image noise model. To verify the effectiveness of the new degradation model, we have trained a deep blind ESRGAN super-resolver and then applied it to super-resolve both synthetic and real images with diverse degradations. The experimental results demonstrate that the new degradation model can help to significantly improve the practicability of deep super-resolvers, thus providing a powerful alternative solution for real SISR applications.

人们普遍认为，如果假定的退化模型与真实图像中的退化模型相背离，单图像超分辨率 (SISR) 方法将无法很好地执行。虽然一些退化模型考虑了其他因素，如模糊，但它们仍然不足以有效地覆盖真实图像的各种退化。为了解决这个问题，本文提出设计一个更复杂但实用的退化模型，该模型包括随机混洗模糊、下采样和噪声退化。具体地说，模糊由两个具有各向同性和各向异性高斯核的卷积来近似；下采样从最近、双线性和双三次插值中随机选择；通过添加不同噪声水平的高斯噪声，采用不同质量因子的JPEG压缩，并通过反向-正向相机图像信号处理 (ISP) 管道模型和原始图像噪声模型生成处理后的相机传感器噪声，合成噪声。为了验证新退化模型的有效性，我们训练了一个深盲ESRGAN超级分解器，然后将其应用于具有不同退化的合成图像和真实图像的超级分解。实验结果表明，新的退化模型有助于显著提高深度超级旋转变压器的实用性，从而为实际SISR应用提供了一个强有力的新方案。

Generalization beyond the training distribution is a core challenge in machine learning. The common practice of mixing and shuffling examples when training neural networks may not be optimal in this regard. We show that partitioning the data into well-chosen, non-i.i.d. subsets treated as multiple training environments can guide the learning of models with better out-of-distribution generalization. We describe a training procedure to capture the patterns that are stable across environments while discarding spurious ones. The method makes a step beyond correlation-based learning: the choice of the partitioning allows injecting information about the task that cannot be otherwise recovered from the joint distribution of the training data. We demonstrate multiple use cases with the task of visual question answering, which is notorious for dataset biases. We obtain significant improvements on VQA-CP, using environments built from prior knowledge, existing meta data, or unsupervised clustering. We also get improvements on GQA using annotations of "equivalent questions", and on multi-dataset training (VQA v2 / Visual Genome) by treating them as distinct environments.

超越训练分布的泛化是机器学习的一个核心挑战。在这方面，训练神经网络时混合和洗牌示例的常见做法可能不是最优的。我们表明，将数据划分为精心选择的、非i.i.d.子集（作为多个训练环境处理）可以以更好的分布外泛化来指导模型的学习。我们描述了一个训练过程来捕获跨环境稳定的模式，同时丢弃虚假的模式。该方法超越了基于相关性的学习：分区的选择允许注入有关任务的信息，否则无法从训练数据的联合分布中恢复这些信息。我们用可视化问答任务演示了多个用例，这种任务因数据集偏差而臭名昭著。我们使用基于先验知识、现有元数据或无监督聚类构建的环境，对VQA-CP进行了重大改进。我

们还使用“等价问题”注释改进了GQA，并将多数据集训练（VQA v2/Visual Genome）作为不同的环境处理。

Understanding the inner workings of deep neural networks (DNNs) is essential to provide trustworthy artificial intelligence techniques for practical applications. Existing studies typically involve linking semantic concepts to units or layers of DNNs, but fail to explain the inference process. In this paper, we introduce neural architecture disentanglement (NAD) to fill the gap. Specifically, NAD learns to disentangle a pre-trained DNN into sub-architectures according to independent tasks, forming information flows that describe the inference processes. We investigate whether, where, and how the disentanglement occurs through experiments conducted with handcrafted and automatically-searched network architectures, on both object-based and scene-based datasets. Based on the experimental results, we present three new findings that provide fresh insights into the inner logic of DNNs. First, DNNs can be divided into sub-architectures for independent tasks. Second, deeper layers do not always correspond to higher semantics. Third, the connection type in a DNN affects how the information flows across layers, leading to different disentanglement behaviors. With NAD, we further explain why DNNs sometimes give wrong predictions. Experimental results show that misclassified images have a high probability of being assigned to task sub-architectures similar to the correct ones. Our code is available at <https://github.com/hujiecpp/NAD>.

了解深层神经网络（DNN）的内部工作原理对于为实际应用提供可靠的人工智能技术至关重要。现有的研究通常涉及将语义概念与DNN的单位或层联系起来，但无法解释推理过程。在本文中，我们引入了神经结构解纠缠（NAD）来填补这一空白。具体而言，NAD学习根据独立任务将预先训练的DNN分解为子架构，形成描述推理过程的信息流。我们通过在基于对象和基于场景的数据集上使用手工制作和自动搜索的网络体系结构进行的实验，来研究这种分离是否发生、发生在何处以及如何发生。基于实验结果，我们提出了三个新的发现，为DNN的内在逻辑提供了新的见解。首先，DNN可以划分为独立任务的子体系结构。第二，更深层次并不总是对应更高的语义。第三，DNN中的连接类型会影响信息如何跨层流动，从而导致不同的解纠缠行为。通过NAD，我们进一步解释了DNN有时给出错误预测的原因。实验结果表明，错误分类的图像很有可能被分配到与正确图像相似的任务子结构中。我们的代码可在<https://github.com/hujiecpp/NAD>。

We present QueryInst, a new perspective for instance segmentation. QueryInst is a multi-stage end-to-end system that treats instances of interest as learnable queries, enabling query based object detectors, e.g., Sparse R-CNN, to have strong instance segmentation performance. The attributes of instances such as categories, bounding boxes, instance masks, and instance association embeddings are represented by queries in a unified manner. In QueryInst, a query is shared by both detection and segmentation via dynamic convolutions and driven by parallelly-supervised multi-stage learning. We conduct extensive experiments on three challenging benchmarks, i.e., COCO, CityScapes, and YouTube-VIS to evaluate the effectiveness of QueryInst in object detection, instance segmentation, and video instance segmentation tasks. For the first time, we demonstrate that a simple end-to-end query based framework can achieve the state-of-the-art performance in various instance-level recognition tasks. Code is available at <https://github.com/hustvl/QueryInst>.

我们提出了QueryList，一种新的实例分割视角。QueryInst是一个多阶段的端到端系统，它将感兴趣的实例视为可学习的查询，使基于查询的对象检测器（例如稀疏R-CNN）具有强大的实例分割性能。实例的属性（如类别、边界框、实例掩码和实例关联嵌入）以统一的方式由查询表示。在QueryInstant中，查询由检测和分割通过动态卷积共享，并由并行监督的多阶段学习驱动。我们在三个具有挑战性的基准上进行了广泛的实验，即COCO、CityScapes和YouTube VIS，以评估QueryInst在目标检测、实例分割

和视频实例分割任务中的有效性。我们首次证明了一个简单的基于端到端查询的框架可以在各种实例级识别任务中实现最先进的性能。代码可在<https://github.com/hustvl/QueryInst>.

Most existing image de-raining networks could only learn fixed mapping rules between paired rainy/clean images on single synthetic dataset and then stay static for lifetime. However, since single synthetic dataset merely provides a partial view for the distribution of rain streaks, the deep models well trained on an individual synthetic dataset tend to overfit on this biased distribution. This leads to the inability of these methods to well generalize to complex and changeable real-world rainy scenes, thus limiting their practical applications. In this paper, we try for the first time to accumulate the de-raining knowledge from multiple synthetic datasets on a single network parameter set to improve the de-raining generalization of deep networks. To achieve this goal, we explore Neural Reorganization (NR) to allow the de-raining network to keep a subtle stability-plasticity trade-off rather than naive stabilization after training phase. Specifically, we design our NR algorithm by borrowing the synaptic consolidation mechanism in the biological brain and knowledge distillation. Equipped with our NR algorithm, the deep model can be trained on a list of synthetic rainy datasets by overcoming catastrophic forgetting, making it a general-version de-raining network. Extensive experimental validation shows that due to the successful accumulation of de-raining knowledge, our proposed method can not only process multiple synthetic datasets consistently, but also achieve state-of-the-art results when dealing with real-world rainy images.

大多数现有的图像去雨网络只能在单个合成数据集上学习成对的雨/干净图像之间的固定映射规则，然后终生保持静态。然而，由于单个合成数据集仅提供雨带分布的局部视图，因此在单个合成数据集上经过良好训练的深层模型往往过度拟合这种有偏分布。这导致这些方法无法很好地推广到复杂多变的真实雨景中，从而限制了它们的实际应用。在本文中，我们首次尝试在单个网络参数集上从多个合成数据集中积累去训练知识，以改进深度网络的去训练泛化。为了实现这一目标，我们探索了神经重组（NR），以使去训练网络在训练阶段后保持微妙的稳定性-可塑性权衡，而不是单纯的稳定。具体来说，我们通过借用生物大脑中的突触巩固机制和知识提取来设计NR算法。通过使用我们的NR算法，通过克服灾难性遗忘，可以在一系列人工降雨数据集上训练deep模型，使其成为一个通用版本的去降雨网络。大量的实验验证表明，由于去雨知识的成功积累，我们提出的方法不仅可以一致地处理多个合成数据集，而且在处理真实世界的雨图像时也可以获得最先进的结果。

We propose a novel approach for probabilistic generative modeling of 3D shapes. Unlike most existing models that learn to deterministically translate a latent vector to a shape, our model, Point-Voxel Diffusion (PVD), is a unified, probabilistic formulation for unconditional shape generation and conditional, multi-modal shape completion. PVD marries denoising diffusion models with the hybrid, point-voxel representation of 3D shapes. It can be viewed as a series of denoising steps, reversing the diffusion process from observed point cloud data to Gaussian noise, and is trained by optimizing a variational lower bound to the (conditional) likelihood function. Experiments demonstrate that PVD is capable of synthesizing high-fidelity shapes, completing partial point clouds, and generating multiple completion results from single-view depth scans of real objects.

我们提出了一种新的三维形状概率生成建模方法。与大多数学习确定地将潜在向量转换为形状的现有模型不同，我们的模型点体素扩散（PVD）是无条件形状生成和条件多模态形状完成的统一概率公式。PVD将去噪扩散模型与3D形状的混合点体素表示相结合。它可以看作是一系列去噪步骤，将观测点云数据的扩散过程逆转为高斯噪声，并通过优化（条件）似然函数的变分下界进行训练。实验表明，PVD能够合成高保真形状，完成部分点云，并从真实对象的单视图深度扫描生成多个完成结果。

Bicubic downscaling is a prevalent technique used to reduce the video storage burden or to accelerate the downstream processing speed. However, the inverse upscaling step is non-trivial, and the downscaled video may also deteriorate the performance of downstream tasks. In this paper, we propose a self-conditioned probabilistic framework for video rescaling to learn the paired downscaling and upscaling procedures simultaneously. During the training, we decrease the entropy of the information lost in the downscaling by maximizing its probability conditioned on the strong spatial-temporal prior information within the downscaled video. After optimization, the downscaled video by our framework preserves more meaningful information, which is beneficial for both the upscaling step and the downstream tasks, e.g., video action recognition task. We further extend the framework to a lossy video compression system, in which a gradient estimator for non-differential industrial lossy codecs is proposed for the end-to-end training of the whole system. Extensive experimental results demonstrate the superiority and effectiveness of our approach on video rescaling, video compression, and efficient action recognition tasks.

双三次降尺度是一种流行的技术，用于减少视频存储负担或加快下游处理速度。然而，反向放大步骤是非常重要的，缩小的视频也可能会降低下游任务的性能。在本文中，我们提出了一个用于视频重缩放的自适应概率框架，以同时学习成对的降尺度和升尺度过程。在训练过程中，我们根据降尺度视频中的强时空先验信息，通过最大化其概率来降低在降尺度过程中丢失的信息的熵。经过优化后，我们的框架缩小的视频保留了更多有意义的信息，这对于放大步骤和下游任务都是有益的，例如视频动作识别任务。我们进一步将该框架扩展到一个有损视频压缩系统，其中提出了一种用于非差分工业有损编解码器的梯度估计器，用于整个系统的端到端训练。大量的实验结果证明了我们的方法在视频重缩放、视频压缩和有效的动作识别任务上的优越性和有效性。

Category-level 6D object pose estimation aims to predict the position and orientation for unseen objects, which plays a pillar role in many scenarios such as robotics and augmented reality. The significant intra-class variation is the bottleneck challenge in this task yet remains unsolved so far. In this paper, we take advantage of category prior to overcome this problem by innovating a structure-guided prior adaptation scheme to accurately estimate 6D pose for individual objects. Different from existing prior-based methods, given one object and its corresponding category prior, we propose to leverage their structure similarity to dynamically adapt the prior to the observed object. The prior adaptation intrinsically associates the adopted prior with different objects, from which we can accurately reconstruct the 3D canonical model of the specific object for pose estimation. To further enhance the structure characteristic of objects, we extract low-rank structure points from the dense object point cloud, therefore more efficiently incorporating sparse structural information during prior adaptation. Extensive experiments on CAMERA25 and REAL275 benchmarks demonstrate significant performance improvement. Project homepage:  
<https://www.cse.cuhk.edu.hk/kaichen/projects/sgpa/sgpa.html>.

类别级6D对象姿势估计旨在预测看不见对象的位置和方向，这在许多场景（如机器人技术和增强现实）中起着支柱作用。显著的类内差异是这项任务的瓶颈挑战，但至今仍未解决。在本文中，我们利用类别先验来克服这个问题，通过创新一种结构引导的先验自适应方案来精确估计单个物体的6D姿势。与现有的基于先验的方法不同，在给定一个对象及其对应的类别先验的情况下，我们提出利用它们的结构相似性来动态地调整先验以适应观测对象。先验自适应将所采用的先验知识与不同的目标进行内在的关联，从而可以准确地重建特定目标的三维规范模型进行姿态估计。为了进一步增强目标的结构特征，我们从稠密的目标点云中提取低阶结构点，从而在先验自适应过程中更有效地融合稀疏结构信息。在CAMERA25和REAL275基准上进行的大量实验表明，性能有了显著提高。项目主页：<https://www.cse.cuhk.edu.hk/kaichen/projects/sgpa/sgpa.html>。

Deep image prior (DIP) serves as a good inductive bias for diverse inverse problems. Among them, denoising is known to be particularly challenging for the DIP due to noise fitting with the requirement of an early stopping. To address the issue, we first analyze the DIP by the notion of effective degrees of freedom (DF) to monitor the optimization progress and propose a principled stopping criterion before fitting to noise without access of a paired ground truth image for Gaussian noise. We also propose the 'stochastic temporal ensemble (STE)' method for incorporating techniques to further improve DIP's performance for denoising. We additionally extend our method to Poisson noise. Our empirical validations show that given a single noisy image, our method denoises the image while preserving rich textual details. Further, our approach outperforms prior arts in LPIPS by large margins with comparable PSNR and SSIM on seven different datasets.

深度图像先验 (DIP) 对于各种反问题来说是一种很好的归纳偏差。其中，众所周知，由于噪声拟合要求提前停止，因此对DIP进行去噪特别具有挑战性。为了解决这个问题，我们首先通过有效自由度 (DF) 的概念来分析DIP，以监控优化过程，并在拟合噪声之前提出一个原则性的停止准则，而无需访问高斯噪声的成对地面真值图像。我们还提出了“随机时间集合 (STE) ”方法，用于结合技术，以进一步提高DIP 的去噪性能。我们还将我们的方法扩展到泊松噪声。我们的实验验证表明，给定一幅带噪图像，我们的方法在预处理丰富的文本细节的同时对图像进行去噪。此外，我们的方法在七种不同数据集上的峰值信噪比 (PSNR) 和SSIM具有可比性，在LPIP方面大大优于现有技术。

The ability to reliably estimate physiological signals from video is a powerful tool in low-cost, pre-clinical health monitoring. In this work we propose a new approach to remote photoplethysmography (rPPG) -- the measurement of blood volume changes from observations of a person's face or skin. Similar to current state-of-the-art methods for rPPG, we apply neural networks to learn deep representations with invariance to nuisance image variation. In contrast to such methods, we employ a fully self-supervised training approach, which has no reliance on expensive ground truth physiological training data. Our proposed method uses contrastive learning with a weak prior over the frequency and temporal smoothness of the target signal of interest. We evaluate our approach on four rPPG datasets, showing that comparable or better results can be achieved compared to recent supervised deep learning methods but without using any annotation. In addition, we incorporate a learned saliency resampling module into both our unsupervised approach and supervised baseline. We show that by allowing the model to learn where to sample the input image, we can reduce the need for hand-engineered features while providing some interpretability into the model's behavior and possible failure modes. We release code for our complete training and evaluation pipeline to encourage reproducible progress in this exciting new direction.

从视频可靠估计生理信号的能力是低成本、临床前健康监测的有力工具。在这项工作中，我们提出了一种远程光体积描记术 (rPPG) 的新方法——通过观察人的面部或皮肤来测量血容量的变化。与目前最先进的rPPG方法类似，我们应用神经网络学习对有害图像变化不变性的深度表示。与这些方法相比，我们采用了一种完全自我监督的训练方法，它不依赖于昂贵的地面真实生理训练数据。我们提出的方法使用对比学习，对感兴趣的目标信号的频率和时间平滑度具有微弱的先验知识。我们在四个rPPG数据集上对我们的方法进行了评估，结果表明，与最近的有监督深度学习方法相比，我们可以获得类似或更好的结果，但不需要使用任何注释。此外，我们在无监督方法和有监督基线中加入了学习显著性重采样模块。我们表明，通过允许模型了解在何处对输入图像进行采样，我们可以减少对手工设计特征的需求，同时为模型的行为和可能的故障模式提供一些可解释性。我们发布了完整培训和评估流程的代码，以鼓励在这一令人兴奋的新方向上取得可复制的进展。

Reversible image conversion (RIC) aims to build a reversible transformation between specific visual content (e.g., short videos) and an embedding image, where the original content can be restored from the embedding when necessary. This work develops Invertible Image Conversion Net (IICNet) as a generic solution to various RIC tasks due to its strong capacity and task-independent design. Unlike previous encoder-decoder based methods, IICNet maintains a highly invertible structure based on invertible neural networks (INNs) to better preserve the information during conversion. We use a relation module and a channel squeeze layer to improve the INN nonlinearity to extract cross-image relations and the network flexibility, respectively. Experimental results demonstrate that IICNet outperforms the specifically-designed methods on existing RIC tasks and can generalize well to various newly-explored tasks. With our generic IICNet, we no longer need to hand-engineer task-specific embedding networks for rapidly occurring visual content. Our source codes are available at: <https://github.com/felixcheng97/IICNet>.

可逆图像转换 (RIC) 旨在构建特定视觉内容 (例如, 短视频) 和嵌入图像之间的可逆转换, 必要时可以从嵌入中恢复原始内容。本工作开发了可逆图像转换网络 (IICNet) , 由于其强大的容量和独立于任务的设计, 它可以作为各种RIC任务的通用解决方案。与以前基于编码器-解码器的方法不同, IICNet保持了基于可逆神经网络 (INN) 的高度可逆结构, 以便在转换过程中更好地保存信息。我们使用一个关系模块和一个通道挤压层来改善INN非线性, 分别提取交叉图像关系和网络灵活性。实验结果表明, IICNet在现有RIC任务上优于专门设计的方法, 并且可以很好地推广到各种新探索的任务。使用我们的通用IICNet, 我们不再需要为快速出现的视觉内容手工设计特定于任务的嵌入网络。我们的源代码可从以下网址获得: <https://github.com/felixcheng97/IICNet>.

Scribble-supervised semantic segmentation has gained much attention recently for its promising performance without high-quality annotations. Due to the lack of supervision, confident and consistent predictions are usually hard to obtain. Typically, people handle these problems by either adopting an auxiliary task with the well-labeled dataset or incorporating a graphical model with additional requirements on scribble annotations. Instead, this work aims to achieve semantic segmentation by scribble annotations directly without extra information and other limitations. Specifically, we propose holistic operations, including minimizing entropy and a network embedded random walk on the neural representation to reduce uncertainty. Given the probabilistic transition matrix of a random walk, we further train the network with self-supervision on its neural eigenspace to impose consistency on predictions between related images. Comprehensive experiments and ablation studies verify the proposed approach, which demonstrates superiority over others; it is even comparable to some full-label supervised ones and works well when scribbles are randomly shrunk or dropped.

Scribble监督语义切分因其在无需高质量注释的情况下具有良好的性能而备受关注。由于缺乏监督, 通常很难获得自信和一致的预测。通常, 人们通过采用带有良好标记的数据集的辅助任务或合并对涂鸦注释有额外要求的图形模型来处理这些问题。相反, 这项工作的目的是通过直接涂鸦注释来实现语义分割, 而不需要额外的信息和其他限制。具体而言, 我们提出了整体操作, 包括最小化熵和在神经表示上嵌入网络的随机行走, 以减少不确定性。在给定随机游动的概率转移矩阵的情况下, 我们进一步对网络进行训练, 并在其神经特征空间上进行自我监督, 以使相关图像之间的预测保持一致。综合实验和烧蚀研究证实了该方法的优越性; 它甚至可以与一些全标签监督的涂鸦相媲美, 并且在随意缩小或删除涂鸦时效果良好。

Recent work has shown that the accuracy of machine learning models can vary substantially when evaluated on a distribution that even slightly differs from that of the training data. As a result, predicting model performance on previously unseen distributions without access to labeled data is an important challenge with implications for increasing the reliability of machine learning models. In the context of distribution shift, distance measures are often used to adapt models and improve their performance on new domains, however accuracy estimation is seldom explored in these investigations. Our investigation determines that common distributional distances such as Frechet distance or Maximum Mean Discrepancy, fail to induce reliable estimates of performance under distribution shift. On the other hand, we find that our proposed difference of confidences (DoC) approach yields successful estimates of a classifier's performance over a variety of shifts and model architectures. Despite its simplicity, we observe that DoC outperforms other methods across synthetic, natural, and adversarial distribution shifts, reducing error by (>46%) on several realistic and challenging datasets such as ImageNet-Vid-Robust and ImageNet-Rendition.

最近的研究表明，机器学习模型的精度在分布上可能会有很大差异，甚至与训练数据的分布略有不同。因此，在不访问标记数据的情况下，预测以前看不见的分布上的模型性能对于提高机器学习模型的可靠性是一个重要的挑战。在分布转移的背景下，通常使用距离度量来调整模型并改进其在新领域的性能，但是在这些研究中很少探讨精度估计。我们的研究确定，常见的分布距离，如Frechet距离或最大平均差异，无法在分布转移情况下得出可靠的性能估计。另一方面，我们发现我们提出的置信度差异（DoC）方法能够成功地估计分类器在各种移位和模型架构下的性能。尽管简单，我们观察到DoC在合成、自然和对抗性分布变化方面优于其他方法，在几个真实且具有挑战性的数据集（如ImageNet Vid Robust和ImageNet Rendition）上减少了误差（>46%）。

Due to the superior ability of global dependency modeling, Transformer and its variants have become the primary choice of many vision-and-language tasks. However, in tasks like Visual Question Answering (VQA) and Referring Expression Comprehension (REC), the multimodal prediction often requires visual information from macro- to micro-views. Therefore, how to dynamically schedule the global and local dependency modeling in Transformer has become an emerging issue. In this paper, we propose an example-dependent routing scheme called TRAnsformer Routing (TRAR) to address this issue. Specifically, in TRAR, each visual Transformer layer is equipped with a routing module with different attention spans. The model can dynamically select the corresponding attentions based on the output of the previous inference step, so as to formulate the optimal routing path for each example. Notably, with careful designs, TRAR can reduce the additional computation and memory overhead to almost negligible. To validate TRAR, we conduct extensive experiments on five benchmark datasets of VQA and REC, and achieve superior performance gains than the standard Transformers and a bunch of state-of-the-art methods.

由于全局依赖建模的优越能力，Transformer及其变体已成为许多视觉和语言任务的主要选择。然而，在视觉问答（VQA）和指称表达理解（REC）等任务中，多模态预测通常需要从宏观到微观的视觉信息。因此，如何动态地调度Transformer中的全局和局部依赖建模成为一个新兴的问题。在本文中，我们提出了一个称为TRAnsformer routing（TRAR）的依赖于示例的路由方案来解决这个问题。具体地说，在TRAR中，每个可视转换器层都配备了具有不同注意广度的路由模块。该模型可以根据前一推理步骤的输出动态选择相应的注意事项，从而为每个示例制定最优路由路径。值得注意的是，通过仔细的设计，TRAR可以将额外的计算和内存开销减少到几乎可以忽略的程度。为了验证TRAR，我们在VQA和REC的五个基准数据集上进行了大量实验，取得了比标准变压器和一系列最先进的方法更高的性能增益。

Solving geometric tasks involving point clouds by using machine learning is a challenging problem. Standard feed-forward neural networks combine linear or, if the bias parameter is included, affine layers and activation functions. Their geometric modeling is limited, which motivated the prior work introducing the multilayer hypersphere perceptron (MLHP). Its constituent part, i.e., the hypersphere neuron, is obtained by applying a conformal embedding of Euclidean space. By virtue of Clifford algebra, it can be implemented as the Cartesian dot product of inputs and weights. If the embedding is applied in a manner consistent with the dimensionality of the input space geometry, the decision surfaces of the model units become combinations of hyperspheres and make the decision-making process geometrically interpretable for humans. Our extension of the MLHP model, the multilayer geometric perceptron (MLGP), and its respective layer units, i.e., geometric neurons, are consistent with the 3D geometry and provide a geometric handle of the learned coefficients. In particular, the geometric neuron activations are isometric in 3D, which is necessary for rotation and translation equivariance. When classifying the 3D Tetris shapes, we quantitatively show that our model requires no activation function in the hidden layers other than the embedding to outperform the vanilla multilayer perceptron. In the presence of noise in the data, our model is also superior to the MLHP.

使用机器学习解决涉及点云的几何任务是一个具有挑战性的问题。标准前馈神经网络结合线性或仿射层和激活函数（如果包含偏差参数）。他们的几何建模是有限的，这激发了先前引入多层超球面感知器（MLHP）的工作。它的组成部分，即超球神经元，是通过应用欧氏空间的共形嵌入而获得的。借助Clifford代数，它可以实现为输入和权重的笛卡尔点积。如果以与输入空间几何体的维度一致的方式应用嵌入，则模型单元的决策面将成为超球体的组合，并使决策过程在几何上可为人类解释。我们对MLHP模型的扩展，多层几何感知器（MLGP）及其相应的层单元，即几何神经元，与3D几何一致，并提供学习系数的几何处理。特别是，几何神经元激活在3D中是等距的，这对于旋转和平移等变是必要的。在对三维俄罗斯方块形状进行分类时，我们定量地表明，我们的模型不需要隐藏层中的激活函数，而只需要嵌入，就可以优于香草多层感知器。在数据中存在噪声的情况下，我们的模型也优于MLHP。

Fine-grained 3D segmentation is an important task in 3D object understanding, especially in applications such as intelligent manufacturing or parts analysis for 3D objects. However, many challenges involved in such problem are yet to be solved, such as i) interpreting the complex structures located in different regions for 3D objects; ii) capturing fine-grained structures with sufficient topology correctness. Current deep learning and graph machine learning methods fail to tackle such challenges and thus provide inferior performance in fine-grained 3D analysis. In this work, methods in topological data analysis are incorporated with geometric deep learning model for the task of fine-grained segmentation for 3D objects. We propose a novel neural network model called Persistent Homology based Graph Convolution Network (PHGCN), which i) integrates persistent homology into graph convolution network to capture multi-scale structural information that can accurately represent complex structures for 3D objects; ii) applies a novel Persistence Diagram Loss that provides sufficient topology correctness for segmentation over the fine-grained structures. Extensive experiments on fine-grained 3D segmentation validate the effectiveness of the proposed PHGCN model and show significant improvements over current state-of-the-art methods.

细粒度三维分割是三维对象理解中的一项重要任务，特别是在智能制造或三维对象零件分析等应用中。然而，这类问题涉及的许多挑战尚未解决，例如i) 为三维对象解释位于不同区域的复杂结构；ii) 捕获具有足够拓扑正确性的细粒度结构。当前的深度学习和图形机器学习方法无法解决这些挑战，因此在细粒度三维分析中的性能较差。在这项工作中，拓扑数据分析方法与几何深度学习模型相结合，用于三维对象的细粒度分割任务。我们提出了一种新的神经网络模型，称为基于持久同调的图卷积网络（PHGCN），它i) 将持久同调集成到图卷积网络中，以捕获能够准确表示三维对象复杂结构的多尺度

结构信息；ii) 应用一种新的持久性图，它为细粒度结构的分段提供了足够的拓扑正确性。对细粒度3D分割的大量实验证了所提出的PHGCN模型的有效性，并显示出相对于当前最先进的方法的显著改进。

Existing video stabilization methods often generate visible distortion or require aggressive cropping of frame boundaries, resulting in smaller field of views. In this work, we present a frame synthesis algorithm to achieve full-frame video stabilization. We first estimate dense warp fields from neighboring frames and then synthesize the stabilized frame by fusing the warped contents. Our core technical novelty lies in the learning-based hybrid-space fusion that alleviates artifacts caused by optical flow inaccuracy and fast-moving objects. We validate the effectiveness of our method on the NUS, selfie, and DeepStab video datasets. Extensive experiment results demonstrate the merits of our approach over prior video stabilization methods.

现有的视频稳定方法通常会产生可见的失真或需要对帧边界进行剧烈的裁剪，从而导致较小的视场。在这项工作中，我们提出了一种帧合成算法来实现全帧视频稳定。我们首先从相邻帧估计密集扭曲场，然后通过融合扭曲内容合成稳定帧。我们的技术创新在于基于学习的混合空间融合，它可以减轻光流不准确和快速移动物体造成的伪影。我们在NUS、selfie和DeepStab视频数据集上验证了我们的方法的有效性。大量的实验结果表明，我们的方法优于以往的视频稳定方法。

Low-latency deep spiking neural networks (SNNs) have become a promising alternative to conventional artificial neural networks (ANNs) because of their potential for increased energy efficiency on event-driven neuromorphic hardware. Neural networks, including SNNs, however, are subject to various adversarial attacks and must be trained to remain resilient against such attacks for many applications. Nevertheless, due to prohibitively high training costs associated with SNNs, analysis, and optimization of deep SNNs under various adversarial attacks have been largely overlooked. In this paper, we first present a detailed analysis of the inherent robustness of low-latency SNNs against popular gradient-based attacks, namely fast gradient sign method (FGSM) and projected gradient descent (PGD). Motivated by this analysis, to harness the model robustness against these attacks we present an SNN training algorithm that uses crafted input noise and incurs no additional training time. To evaluate the merits of our algorithm, we conducted extensive experiments with variants of VGG and ResNet on both CIFAR-10 and CIFAR-100 datasets. Compared to standard trained direct input SNNs, our trained models yield improved classification accuracy of up to 13.7% and 10.1% on FGSM and PGD attack-generated images, respectively, with negligible loss in clean image accuracy. Our models also outperform inherently-robust SNNs trained on rate-coded inputs with improved or similar classification performance on attack-generated images while having up to 25x and 4.6x lower latency and computation energy, respectively.

低潜伏期深脉冲神经网络 (SNN) 已成为传统人工神经网络 (ANN) 的一种很有前途的替代方案，因为它们在事件驱动的神经形态硬件上具有提高能量效率的潜力。然而，包括SNN在内的神经网络会受到各种对抗性攻击，因此必须对其进行训练，以便在许多应用中保持对此类攻击的弹性。然而，由于与SNN相关的高昂训练成本，在各种对抗性攻击下对深层SNN的分析和优化在很大程度上被忽略了。在本文中，我们首先详细分析了低延迟SNN对流行的基于梯度的攻击的固有鲁棒性，即快速梯度符号法 (FGSM) 和投影梯度下降法 (PGD)。受此分析的启发，为了利用模型的鲁棒性抵御这些攻击，我们提出了一种SNN训练算法，该算法使用精心编制的输入噪声，并且不需要额外的训练时间。为了评估我们算法的优点，我们在CIFAR-10和CIFAR-100数据集上对VGG和ResNet的变体进行了广泛的实验。与标准训练的直接输入SNN相比，我们训练的模型在FGSM和PGD攻击生成的图像上的分类精度分别提高了13.7%和10.1%，而干净图像的精度损失可以忽略不计。我们的模型还优于在速率编码输入上训练的固有鲁棒SNN，在攻击生成的图像上具有改进或类似的分类性能，同时延迟和计算能量分别降低25倍和4.6倍。

Nuclear instance segmentation is a challenging task due to a large number of touching and overlapping nuclei in pathological images. Existing methods cannot effectively recognize the accurate boundary owing to neglecting the relationship between pixels (e.g., direction information). In this paper, we propose a novel Centripetal Direction Network (CDNet) for nuclear instance segmentation. Specifically, we define the centripetal direction feature as a class of adjacent directions pointing to the nuclear center to represent the spatial relationship between pixels within the nucleus. These direction features are then used to construct a direction difference map to represent the similarity within instances and the differences between instances. Finally, we propose a direction-guided refinement module, which acts as a plug-and-play module to effectively integrate auxiliary tasks and aggregate the features of different branches. Experiments on MoNuSeg and CPM17 datasets show that CDNet is significantly better than the other methods and achieves the state-of-the-art performance. The code is available at <https://github.com/honglianghe/CDNet>.

核实例分割是一项具有挑战性的任务，因为病理图像中存在大量的接触和重叠核。现有的方法由于忽略了像素之间的关系（如方向信息），无法有效地识别出精确的边界。本文提出了一种用于核实例分割的向心方向网络（CDNet）。具体来说，我们将向心方向特征定义为一类指向核中心的相邻方向，以表示核内像素之间的空间关系。然后使用这些方向特征构建方向差异映射，以表示实例内的相似性和实例之间的差异。最后，我们提出了一个方向引导的细化模块，它作为一个即插即用模块来有效地集成辅助任务并聚合不同分支的特征。在MoNuSeg和CPM17数据集上的实验表明，CDNet明显优于其他方法，并实现了最先进的性能。该守则可于<https://github.com/honglianghe/CDNet>.

Recently, directly detecting 3D objects from 3D point clouds has received increasing attention. To extract object representation from an irregular point cloud, existing methods usually take a point grouping step to assign the points to an object candidate so that a PointNet-like network could be used to derive object features from the grouped points. However, the inaccurate point assignments caused by the hand-crafted grouping scheme decrease the performance of 3D object detection. In this paper, we present a simple yet effective method for directly detecting 3D objects from the 3D point cloud. Instead of grouping local points to each object candidate, our method computes the feature of an object from all the points in the point cloud with the help of an attention mechanism in the Transformers, where the contribution of each point is automatically learned in the network training. With an improved attention stacking scheme, our method fuses object features in different stages and generates more accurate object detection results. With few bells and whistles, the proposed method achieves state-of-the-art 3D object detection performance on two widely used benchmarks, ScanNet V2 and SUN RGB-D.

近年来，直接从三维点云中检测三维物体越来越受到人们的关注。为了从不规则点云中提取对象表示，现有方法通常采取点分组步骤将点分配给对象候选，以便使用类似点网的网络从分组的点派生对象特征。然而，手工制作的分组方案导致的不准确的点指定会降低三维对象检测的性能。在本文中，我们提出了一种简单而有效的方法直接检测三维物体从三维点云。我们的方法不是将局部点分组到每个候选对象，而是借助Transformers中的注意机制从点云中的所有点计算对象的特征，其中每个点的贡献在网络训练中自动学习。通过改进的注意力叠加方案，我们的方法融合了不同阶段的目标特征，并生成更准确的目标检测结果。该方法在scannetv2和sunrgb-D两个广泛使用的基准上实现了最先进的三维目标检测性能。

Face clustering plays an essential role in exploiting massive unlabeled face data. Recently, graph-based face clustering methods are getting popular for their satisfying performances. However, they usually suffer from excessive memory consumption especially on large-scale graphs, and rely on empirical thresholds to determine the connectivities between samples in inference, which restricts their applications in various real-world scenes. To address such problems, in this paper, we explore face clustering from the pairwise angle. Specifically, we formulate the face clustering task as a pairwise relationship classification task, avoiding the memory-consuming learning on large-scale graphs. The classifier can directly determine the relationship between samples and is enhanced by taking advantage of the contextual information. Moreover, to further facilitate the efficiency of our method, we propose a rank-weighted density to guide the selection of pairs sent to the classifier. Experimental results demonstrate that our method achieves state-of-the-art performances on several public clustering benchmarks at the fastest speed and shows a great advantage in comparison with graph-based clustering methods on memory consumption.

人脸聚类在挖掘海量未标记人脸数据中起着至关重要的作用。近年来，基于图的人脸聚类方法以其令人满意的性能得到了广泛的应用。然而，它们通常会遭受过多的内存消耗，尤其是在大规模图上，并且在推理过程中依赖经验阈值来确定样本之间的关联性，这限制了它们在各种实际场景中的应用。为了解决这些问题，本文从两两的角度探讨了人脸聚类问题。具体来说，我们将人脸聚类任务描述为一个成对关系分类任务，避免了大规模图上的内存消耗学习。分类器可以直接确定样本之间的关系，并利用上下文信息进行增强。此外，为了进一步提高我们方法的效率，我们提出了一种秩加权密度来指导发送到分类器的对的选择。实验结果表明，我们的方法以最快的速度在几个公共聚类基准上取得了最先进的性能，并且与基于图的聚类方法相比，在内存消耗方面显示出巨大的优势。

Deep neural networks (DNN) have shown superior performance in a variety of tasks. As they rapidly evolve, their escalating computation and memory demands make it challenging to deploy them on resource-constrained edge devices. Though extensive efficient accelerator designs, from traditional electronics to emerging photonics, have been successfully demonstrated, they are still bottlenecked by expensive memory accesses due to tremendous gaps between the bandwidth/power-latency of electrical memory and computing cores. Previous solutions fail to fully-leverage the ultra-fast computational speed of emerging DNN accelerators to break through the critical memory bound. In this work, we propose a general and unified framework to trade expensive memory transactions with ultra-fast on-chip computations, directly translating to performance improvement. We are the first to jointly explore the intrinsic correlations and bit-level redundancy within DNN kernels and propose a multi-level *in situ* generation mechanism with mixed-precision bases to achieve on-the-fly recovery of high-resolution parameters with minimum hardware overhead. Extensive experiments demonstrate that our proposed joint method can boost the memory efficiency by 10-20x with comparable accuracy over four state-of-the-art designs when benchmarked on ResNet-18/DenseNet-121/MobileNetV2/V3 with various tasks.

深度神经网络 (DNN) 在各种任务中表现出优越的性能。随着它们的快速发展，不断升级的计算和内存需求使得在资源受限的边缘设备上部署它们变得非常困难。尽管从传统电子学到新兴光子学的广泛高效加速器设计已经成功地得到了证明，但由于电存储器和计算核心的带宽/功率/延迟之间存在巨大的差距，它们仍然受到昂贵内存访问的制约。以前的解决方案未能充分利用新兴DNN加速器的超快计算速度来突破临界内存限制。在这项工作中，我们提出了一个通用和统一的框架，用超快的片上计算来交换昂贵的内存事务，直接转化为性能改进。我们是第一个联合探索DNN内核中的内在相关性和位级冗余的人，并提出了一种具有混合精度基础的多级原位生成机制，以最小的硬件开销实现高分辨率参数的动态恢复。大量实验表明，当在ResNet-18/DenseNet-121/MobileNetV2/V3上进行各种任务的基准测试时，我们提出的联合方法可以将内存效率提高10-20倍，与四种最先进的设计相比具有相当的精度。

Language bias is a critical issue in Visual Question Answering (VQA), where models often exploit dataset biases for the final decision without considering the image information. As a result, they suffer from performance drop on out-of-distribution data and inadequate visual explanation. Based on experimental analysis for existing robust VQA methods, we stress the language bias in VQA that comes from two aspects, i.e., distribution bias and shortcut bias. We further propose a new de-bias framework, Greedy Gradient Ensemble (GGE), which combines multiple biased models for unbiased base model learning. With the greedy strategy, GGE forces the biased models to over-fit the biased data distribution in priority, thus makes the base model pay more attention to examples that are hard to solve by biased models. The experiments demonstrate that our method makes better use of visual information and achieves state-of-the-art performance on diagnosing dataset VQA-CP without using extra annotations.

语言偏差是视觉问答 (VQA) 中的一个关键问题，模型通常利用数据集偏差进行最终决策，而不考虑图像信息。因此，他们在分布数据之外的情况下会出现性能下降，并且视觉解释不充分。在对现有稳健 VQA 方法进行实验分析的基础上，着重分析了 VQA 中的语言偏差，即分布偏差和快捷偏差。我们进一步提出了一个新的去偏框架，贪婪梯度集成 (GGE)，它结合了多个有偏模型进行无偏基础模型学习。通过贪婪策略，GGE 强制有偏模型优先过度拟合有偏数据分布，从而使基础模型更加关注有偏模型难以求解的实例。实验表明，我们的方法能够更好地利用视觉信息，在不使用额外注释的情况下，在诊断数据集 VQA-CP 时达到了最先进的性能。

High-level understanding of stories in video such as movies and TV shows from raw data is extremely challenging. Modern video question answering (VideoQA) systems often use additional human-made sources like plot synopses, scripts, video descriptions or knowledge bases. In this work, we present a new approach to understand the whole story without such external sources. The secret lies in the dialog: unlike any prior work, we treat dialog as a noisy source to be converted into text description via dialog summarization, much like recent methods treat video. The input of each modality is encoded by transformers independently, and a simple fusion method combines all modalities, using soft temporal attention for localization over long inputs. Our model outperforms the state of the art on the KnowIT VQA dataset by a large margin, without using question-specific human annotation or human-made plot summaries. It even outperforms human evaluators who have never watched any whole episode before. Code is available at <https://engindeniz.github.io/dialogsummary-videoqa>

从原始数据高度理解电影和电视节目等视频中的故事非常具有挑战性。现代视频答疑 (VideoQA) 系统通常使用额外的人工来源，如剧情简介、脚本、视频描述或知识库。在这项工作中，我们提出了一种新的方法来理解整个故事，而不需要这些外部来源。秘密在于对话：与以前的任何工作不同，我们将对话视为噪声源，通过对话摘要转换为文本描述，就像最近处理视频的方法一样。每个模态的输入由变压器独立编码，一种简单的融合方法将所有模态结合起来，使用软时间注意在长输入上进行定位。我们的模型在 KnowIT VQA 数据集上的性能大大超过了最新技术，没有使用特定问题的人工注释或人工绘制的情节摘要。它甚至比以前从未看过整集的人类评估者表现得更好。代码可在<https://engindeniz.github.io/dialogsummary-videoqa>

Fashion is intertwined with external cultural factors, but identifying these links remains a manual process limited to only the most salient phenomena. We propose a data-driven approach to identify specific cultural factors affecting the clothes people wear. Using large-scale datasets of news articles and vintage photos spanning a century, we present a multi-modal statistical model to detect influence relationships between happenings in the world and people's choice of clothing. Furthermore, on two image datasets we apply our model to improve the concrete vision tasks of visual style forecasting and photo timestamping. Our work is a first step towards a computational, scalable, and easily refreshable approach to link culture to clothing.

时尚与外部文化因素交织在一起，但识别这些联系仍然是一个手工过程，仅限于最显著的现象。我们提出了一种数据驱动的方法来识别影响人们穿着的特定文化因素。利用跨越一个世纪的大规模新闻文章和经典照片数据集，我们提出了一个多模态统计模型，以检测世界上发生的事件与人们的服装选择之间的关系。此外，在两个图像数据集上，我们应用我们的模型来改进视觉风格预测和照片时间戳的具体视觉任务。我们的工作是朝着将文化与服装联系起来的计算性、可伸缩性和易于更新的方法迈出的第一步。

Model training and evaluation are two main time-consuming processes during neural architecture search (NAS). Although weight-sharing based methods have been proposed to reduce the number of trained networks, these methods still need to train the supernet for hundreds of epochs and evaluate thousands of subnets to find the optimal network architecture. In this paper, we propose NAS with Batch Normalization (BN), which we refer to as BN-NAS, to accelerate both the evaluation and training process. For fast evaluation, we propose a novel BN-based indicator that predicts subnet performance at a very early training stage. We further improve the training efficiency by only training the BN parameters during the supernet training. This is based on our observation that training the whole supernet is not necessary while training only BN parameters accelerates network convergence for network architecture search. Extensive experiments show that our method can significantly shorten the time of training supernet by more than 10 times and evaluating subnets by more than 600,000 times without losing accuracy.

模型训练和评估是神经结构搜索 (NAS) 中两个主要的耗时过程。虽然已经提出了基于权重共享的方法来减少训练网络的数量，但这些方法仍然需要训练数百个时代的超级网，并评估数千个子网以找到最佳的网络结构。在本文中，我们提出了带有批量规范化 (BN) 的NAS，我们称之为BN-NAS，以加速评估和培训过程。为了快速评估，我们提出了一种新的基于BN的指标，它可以在非常早期的训练阶段预测子网性能。在supernet训练过程中，我们只训练BN参数，进一步提高了训练效率。这是基于我们的观察，即训练整个超网是不必要的，而只训练BN参数可以加速网络收敛以进行网络架构搜索。大量实验表明，该方法可以在不损失精度的前提下，将训练超网的时间缩短10倍以上，评估子网的时间缩短60万倍以上。

Attention networks perform well on diverse computer vision tasks. The core idea is that the signal of interest is stronger in some pixels ("foreground"), and by selectively focusing computation on these pixels, networks can extract subtle information buried in noise and other sources of corruption. Our paper is based on one key observation: in many real-world applications, many sources of corruption, such as illumination and motion, are often shared between the "foreground" and the "background" pixels. Can we utilize this to our advantage? We propose the utility of inverse attention networks, which focus on extracting information about these shared sources of corruption. We show that this helps to effectively suppress shared covariates and amplify signal information, resulting in improved performance. We illustrate this on the task of camera-based physiological measurement where the signal of interest is weak and global illumination variations and motion act as significant shared sources of corruption. We perform experiments on three datasets and show that our approach of inverse attention produces state-of-the-art results, increasing the signal-to-noise ratio by up to 5.8 dB, reducing heart rate and breathing rate estimation errors by as much as 30 %, recovering subtle waveform dynamics, and generalizing from RGB to NIR videos without retraining.

注意力网络在不同的计算机视觉任务中表现良好。其核心思想是，感兴趣的信号在某些像素（“前景”）中更强，通过选择性地将计算集中在这些像素上，网络可以提取隐藏在噪声和其他腐败源中的微妙信息。我们的论文基于一个关键观察：在许多实际应用中，“前景”和“背景”像素之间通常共享许多损坏源，如照明和运动。我们能利用这个优势吗？我们提出了反向注意网络的效用，它专注于提取关于这些共享腐败来源的信息。我们表明，这有助于有效抑制共享协变量并放大信号信息，从而提高性能。我们在基于摄像头的生理测量任务中说明了这一点，其中感兴趣的信号很弱，并且全局照明变化和运动作为重要的共享腐败源。我们在三个数据集上进行了实验，结果表明，我们的反向注意方法产生了最先进的结果，使信噪比提高了5.8 dB，心率和呼吸频率估计误差降低了30%，恢复了微妙的波形动力学，从RGB推广到NIR视频，无需再培训。

Automatic polyp segmentation from colonoscopy images is an essential step in computer aided diagnosis for colorectal cancer. Most of polyp segmentation methods reported in recent years are based on fully supervised deep learning. However, annotation for polyp images by physicians during the diagnosis is time-consuming and costly. In this paper, we present a novel semi-supervised polyp segmentation via collaborative and adversarial learning of focused and dispersive representations learning model, where focused and dispersive extraction module are used to deal with the diversity of location and shape of polyps. In addition, confidence maps produced by a discriminator in an adversarial training framework shows the effectiveness of leveraging unlabeled data and improving the performance of segmentation network. Consistent regularization is further employed to optimize the segmentation networks to strengthen the representation of the outputs of focused and dispersive extraction module. We also propose an auxiliary adversarial learning method to better leverage unlabeled examples to further improve semantic segmentation accuracy. We conduct extensive experiments on two famous polyp datasets: Kvasir-SEG and CVC-Clinic DB. Experimental results demonstrate the effectiveness of the proposed model, consistently outperforming state-of-the-art semi-supervised segmentation models based on adversarial training and even some advanced fully supervised models. Codes will be released upon publication.

从结肠镜图像中自动分割息肉是计算机辅助诊断结直肠癌的重要步骤。近年来报道的大多数息肉分割方法都是基于全监督深度学习的。然而，医生在诊断过程中对息肉图像进行注释既耗时又昂贵。在本文中，我们提出了一种新的基于聚焦和分散表征的协作和对抗学习的半监督息肉分割学习模型，其中聚焦和分散提取模块用于处理息肉位置和形状的多样性。此外，在对抗性训练框架中由鉴别器生成的置信图显示了利用未标记数据和改进分割网络性能的有效性。一致正则化进一步用于优化分割网络，以增强聚

焦和分散提取模块输出的表示。我们还提出了一种辅助的对抗性学习方法，以更好地利用未标记的示例进一步提高语义分割的准确性。我们在两个著名的息肉数据集：Kvasir SEG和CVC Clinic DB上进行了广泛的实验。实验结果证明了该模型的有效性，始终优于基于对抗训练的最新半监督分割模型，甚至一些先进的全监督分割模型。代码将在发布后发布。

Event cameras are bio-inspired sensors that respond to brightness changes asynchronously and output in the form of event streams instead of frame-based images. They own outstanding advantages compared with traditional cameras: higher temporal resolution, higher dynamic range, and lower power consumption. However, the spatial resolution of existing event cameras is insufficient and challenging to be enhanced at the hardware level while maintaining the asynchronous philosophy of circuit design. Therefore, it is imperative to explore the algorithm of event stream super-resolution, which is a non-trivial task due to the sparsity and strong spatio-temporal correlation of the events from an event camera. In this paper, we propose an end-to-end framework based on spiking neural network for event stream super-resolution, which can generate high-resolution (HR) event stream from the input low-resolution (LR) event stream. A spatiotemporal constraint learning mechanism is proposed to learn the spatial and temporal distributions of the event stream simultaneously. We validate our method on four large-scale datasets and the results show that our method achieves state-of-the-art performance. The satisfying results on two downstream applications, i.e. object classification and image reconstruction, further demonstrate the usability of our method. To prove the application potential of our method, we deploy it on a mobile platform. The high-quality HR event stream generated by our real-time system demonstrates the effectiveness and efficiency of our method.

事件摄影机是一种仿生传感器，可异步响应亮度变化，并以事件流的形式输出，而不是基于帧的图像。与传统相机相比，它们具有更高的时间分辨率、更高的动态范围和更低的功耗。然而，现有事件摄像机的空间分辨率不足，在保持电路设计的异步原理的同时，在硬件层面上提高分辨率具有挑战性。因此，探索事件流超分辨率算法势在必行，这是一项非常重要的任务，因为事件摄像机中的事件具有稀疏性和强时空相关性。本文提出了一种基于尖峰神经网络的事件流超分辨率端到端框架，该框架可以从输入的低分辨率（LR）事件流生成高分辨率（HR）事件流。提出了一种时空约束学习机制来同时学习事件流的时空分布。我们在四个大规模数据集上验证了我们的方法，结果表明我们的方法达到了最先进的性能。在两个下游应用，即目标分类和图像重建上的令人满意的结果进一步证明了我们方法的可用性。为了证明我们的方法的应用潜力，我们将其部署在移动平台上。我们的实时系统生成的高质量HR事件流证明了我们方法的有效性和效率。

Recent methods for long-tailed instance segmentation still struggle on rare object classes with few training data. We propose a simple yet effective method, Feature Augmentation and Sampling Adaptation (FASA), that addresses the data scarcity issue by augmenting the feature space especially for rare classes. Both the Feature Augmentation (FA) and feature sampling components are adaptive to the actual training status -- FA is informed by the feature mean and variance of observed real samples from past iterations, and we sample the generated virtual features in a loss-adapted manner to avoid over-fitting. FASA does not require any elaborate loss design, and removes the need for inter-class transfer learning that often involves large cost and manually-defined head/tail class groups. We show FASA is a fast, generic method that can be easily plugged into standard or long-tailed segmentation frameworks, with consistent performance gains and little added cost. FASA is also applicable to other tasks like long-tailed classification with state-of-the-art performance.

最近的长尾实例分割方法在训练数据较少的稀有对象类上仍然存在困难。我们提出了一种简单而有效的方法，即特征增强和采样自适应（FASA），该方法通过增加特征空间（尤其是稀有类）来解决数据稀缺问题。特征增强（FA）和特征采样组件都能适应实际的训练状态——FA由过去迭代中观察到的真实样本的特征均值和方差来通知，我们以损失自适应的方式对生成的虚拟特征进行采样，以避免过拟合。FASA不需要任何复杂的损失设计，并且消除了通常涉及大量成本和手动定义的头/尾类组的类间转移学习的需要。我们证明了FASA是一种快速、通用的方法，可以很容易地插入标准或长尾分割框架，具有一致的性能增益和很少的附加成本。FASA还适用于其他任务，如具有最先进性能的长尾分类。

Deep learning for image and video compression has demonstrated promising results both as a standalone technology and a hybrid combination with existing codecs. However, these systems still come with high computational costs. Deep learning models are typically applied directly in pixel space, making them expensive when resolutions become large. In this work, we propose an online-trained upsampler to augment an existing codec. The upsampler is a small neural network trained on an isolated group of frames. Its parameters are signalled to the decoder. This hybrid solution has a small scope of only 10s or 100s of frames and allows for a low complexity both on the encoding and the decoding side. Our algorithm works in offline and in zero-latency settings. Our evaluation employs the popular x265 codec on several high-resolution datasets ranging from Full HD to 8K. We demonstrate rate savings between 8.6% and 27.5% and provide ablation studies to show the impact of our design decisions. In comparison to similar works, our approach performs favourably.

图像和视频压缩的深入学习已经证明了作为一种独立技术和与现有编解码器的混合组合的有希望的结果。然而，这些系统的计算成本仍然很高。深度学习模型通常直接应用于像素空间，当分辨率变大时，这使得它们变得昂贵。在这项工作中，我们提出了一个在线培训的上采样，以扩大现有的编解码器。上采样器是在一组孤立的帧上训练的一个小型神经网络。其参数通过信号发送给解码器。该混合解决方案具有仅10s或100s帧的小范围，并且允许编码和解码侧的低复杂度。我们的算法在离线和零延迟设置下工作。我们的评估在从全高清到8K的多个高分辨率数据集上采用了流行的x265编解码器。我们证明了在8.6%和27.5%之间的费率节约，并提供了烧蚀研究，以显示我们的设计决策的影响。与类似的工作相比，我们的方法表现良好。

Compared with traditional methods, the deep learning-based multi-focus image fusion methods can effectively improve the performance of image fusion tasks. However, the existing deep learning-based methods encounter a common issue of a large number of parameters, which leads to the deep learning models with high time complexity and low fusion efficiency. To address this issue, we propose a novel discrete Tchebichef moment-based Deep neural network, termed as DTMNet, for multi-focus image fusion. The proposed DTMNet is an end-to-end deep neural network with only one convolutional layer and three fully connected layers. The convolutional layer is fixed with DTM coefficients (DTMConv) to extract high/low-frequency information without learning parameters effectively. The three fully connected layers have learnable parameters for feature classification. Therefore, the proposed DTMNet for multi-focus image fusion has a small number of parameters (0.01M paras vs. 4.93M paras of regular CNN) and high computational efficiency (0.32s vs. 79.09s by regular CNN to fuse an image). In addition, a large-scale multi-focus image dataset is synthesized for training and verifying the deep learning model. Experimental results on three public datasets demonstrate that the proposed method is competitive with or even outperforms the state-of-the-art multi-focus image fusion methods in terms of subjective visual perception and objective evaluation metrics.

与传统方法相比，基于深度学习的多聚焦图像融合方法能够有效地提高图像融合任务的性能。然而，现有的基于深度学习的方法普遍存在参数较多的问题，导致深度学习模型时间复杂度高、融合效率低。为了解决这个问题，我们提出了一种新的基于离散切比雪夫矩的深度神经网络，称为DTMNet，用于多聚焦图像融合。所提出的DTMNet是一个端到端的深度神经网络，只有一个卷积层和三个完全连接层。卷积层用DTM系数(DTMConv)固定，在不学习参数的情况下有效地提取高频/低频信息。三个完全连接的层具有可学习的特征分类参数。因此，建议的用于多聚焦图像融合的DTMNet具有少量参数(0.01M段与常规CNN的4.93M段)和高计算效率(常规CNN融合图像的时间为0.32s与79.09s)。此外，还合成了一个大规模多聚焦图像数据集，用于训练和验证深度学习模型。在三个公共数据集上的实验结果表明，该方法在主观视觉感知和客观评价指标方面与最新的多聚焦图像融合方法具有竞争性，甚至优于现有的多聚焦图像融合方法。

Egocentric video recognition is a challenging task that requires to identify both the actor's motion and the active object that the actor interacts with. Recognizing the active object is particularly hard due to the cluttered background with distracting objects, the frequent field of view changes, severe occlusion, etc. To improve the active object classification, most existing methods use object detectors or human gaze information, which are computationally expensive or require labor-intensive annotations. To avoid these additional costs, we propose an end-to-end Interactive Prototype Learning (IPL) framework to learn better active object representations by leveraging the motion cues from the actor. First, we introduce a set of verb prototypes to disentangle active object features from distracting object features. Each prototype corresponds to a primary motion pattern of an egocentric action, offering a distinctive supervision signal for active object feature learning. Second, we design two interactive operations to enable the extraction of active object features, i.e., noun-to-verb assignment and verb-to-noun selection. These operations are parameter-efficient and can learn judicious location-aware features on top of 3D CNN backbones. We demonstrate that the IPL framework can generalize to different backbones and outperform the state-of-the-art on three large-scale egocentric video datasets, i.e., EPIC-KITCHENS-55, EPIC-KITCHENS-100 and EGTEA.

以自我为中心的视频识别是一项具有挑战性的任务，需要识别演员的动作和演员与之交互的活动对象。由于背景杂乱、目标分散、视野变化频繁、遮挡严重等原因，识别活动目标尤其困难。为了改进活动目标分类，大多数现有方法使用目标检测器或人类注视信息，它们的计算成本很高，或者需要劳动密集型注释。为了避免这些额外的成本，我们提出了一个端到端的交互式原型学习(IPL)框架，通过利用演员的运动线索来学习更好的活动对象表示。首先，我们引入了一组动词原型来分离主动宾语特征和分散宾语特征。每个原型对应于自我中心动作的主要运动模式，为主动对象特征学习提供独特的监督信号。其次，我们设计了两个交互操作来实现主动对象特征的提取，即名词到动词的赋值和动词到名词的选择。这些操作具有参数效率，可以在3D CNN主干上学习明智的位置感知功能。我们证明了IPL框架可以推广到不同的主干网，并且在三个大规模以自我为中心的视频数据集上，即EPIC-KITCHENS-55、EPIC-KITCHENS-100和EGTEA上的表现优于最新技术。

Motion blur is one of the major challenges remaining for visual odometry methods. In low-light conditions where longer exposure times are necessary, motion blur can appear even for relatively slow camera motions. In this paper we present a novel hybrid visual odometry pipeline with direct approach that explicitly models and estimates the camera's local trajectory within exposure time. This allows us to actively compensate for any motion blur that occurs due to the camera motion. In addition, we also contribute a novel benchmarking dataset for motion blur aware visual odometry. In experiments we show that by directly modeling the image formation process we are able to improve robustness of the visual odometry, while keeping comparable accuracy as that for images without motion blur.

运动模糊是视觉里程计方法仍然面临的主要挑战之一。在需要较长曝光时间的弱光条件下，即使相机运动相对较慢，也可能出现运动模糊。在本文中，我们提出了一种新的混合视觉里程管道与直接的方法，明确建模和估计相机的局部轨迹在曝光时间。这使我们能够主动补偿由于相机运动而产生的任何运动模糊。此外，我们还为运动模糊感知视觉里程测量提供了一个新的基准数据集。在实验中，我们表明，通过直接建模图像形成过程，我们能够提高视觉里程计的鲁棒性，同时保持与没有运动模糊的图像相当的精度。

Geometry-aware modules are widely applied in recent deep learning architectures for scene representation and rendering. However, these modules require intrinsic camera information that might not be obtained accurately. In this paper, we propose a Spatial Transformation Routing (STR) mechanism to model the spatial properties without applying any geometric prior. The STR mechanism treats the spatial transformation as the message passing process, and the relation between the view poses and the routing weights is modeled by an end-to-end trainable neural network. Besides, an Occupancy Concept Mapping (OCM) framework is proposed to provide explainable rationals for scene-fusion processes. We conducted experiments on several datasets and show that the proposed STR mechanism improves the performance of the Generative Query Network (GQN). The visualization results reveal that the routing process can pass the observed information from one location of some view to the associated location in the other view, which demonstrates the advantage of the proposed model in terms of spatial cognition.

几何感知模块广泛应用于最近的场景表示和渲染深度学习体系结构中。但是，这些模块需要可能无法准确获取的固有摄像机信息。在本文中，我们提出了一种空间转换路由（STR）机制来模拟空间属性，而不应用任何几何先验知识。STR机制将空间变换视为消息传递过程，视图姿态和路由权重之间的关系由端到端可训练神经网络建模。此外，提出了一种占用概念映射（OCM）框架，为场景融合过程提供了可解释的合理性。我们在几个数据集上进行了实验，结果表明所提出的STR机制提高了生成查询网络（GQN）的性能。可视化结果表明，路由过程可以将观察到的信息从某个视图的一个位置传递到另一个视图中的相关位置，这表明了该模型在空间认知方面的优势。

For several emerging technologies such as augmented reality, autonomous driving and robotics, visual localization is a critical component. Directly regressing camera pose/3D scene coordinates from the input image using deep neural networks has shown great potential. However, such methods assume a stationary data distribution with all scenes simultaneously available during training. In this paper, we approach the problem of visual localization in a continual learning setup -- whereby the model is trained on scenes in an incremental manner. Our results show that similar to the classification domain, non-stationary data induces catastrophic forgetting in deep networks for visual localization. To address this issue, a strong baseline based on storing and replaying images from a fixed buffer is proposed. Furthermore, we propose a new sampling method based on coverage score (Buff-CS) that adapts the existing sampling strategies in the buffering process to the problem of visual localization. Results demonstrate consistent improvements over standard buffering methods on two challenging datasets -- 7Scenes, 12Scenes, and also 19Scenes by combining the former scenes.

对于一些新兴技术，如增强现实技术、自动驾驶技术和机器人技术，视觉定位是一个关键组成部分。使用深度神经网络从输入图像直接回归相机姿态/三维场景坐标显示了巨大的潜力。然而，这种方法假设在训练期间所有场景同时可用的情况下数据分布是平稳的。在本文中，我们探讨了在连续学习环境中的视觉定位问题——即以增量方式在场景中训练模型。我们的结果表明，与分类域类似，非平稳数据在视觉定位的深层网络中会导致灾难性遗忘。为了解决这个问题，提出了一种基于从固定缓冲区存储和重放图像的强基线。此外，我们还提出了一种新的基于覆盖分数的采样方法（Buff-CS），该方法将缓冲过程中现有的采样策略应用于视觉定位问题。结果表明，在两个具有挑战性的数据集（7个场景、12个场景和19个场景）上，通过组合前一个场景，标准缓冲方法得到了一致的改进。

Diagnosing diseases from medical radiographs and writing reports requires professional knowledge and is time-consuming. To address this, automatic medical report generation approaches have recently gained interest. However, identifying diseases as well as correctly predicting their corresponding sizes, locations and other medical description patterns, which is essential for generating high-quality reports, is challenging. Although previous methods focused on producing readable reports, how to accurately detect and describe findings that match with the query X-Ray has not been successfully addressed. In this paper, we propose a multi-modality semantic attention model to integrate visual features, predicted key finding embeddings, as well as clinical features, and progressively decode reports with visual-textual semantic consistency. First, multi-modality features are extracted and attended with the hidden states from the sentence decoder, to encode enriched context vectors for better decoding a report. These modalities include regional visual features of scans, semantic word embeddings of the top-K findings predicted with high probabilities, and clinical features of indications. Second, the progressive report decoder consists of a sentence decoder and a word decoder, where we propose image-sentence matching and description accuracy losses to constrain the visual-textual semantic consistency. Extensive experiments on the public MIMIC-CXR and IU X-Ray datasets show that our model achieves consistent improvements over the state-of-the-art methods.

通过医学射线照片和撰写报告诊断疾病需要专业知识，而且耗时。为了解决这个问题，自动医疗报告生成方法最近引起了人们的兴趣。然而，识别疾病以及正确预测其相应的大小、位置和其他医学描述模式，这对于生成高质量报告至关重要，是一项挑战。尽管以前的方法侧重于生成可读的报告，但如何准确地检测和描述与查询X射线相匹配的结果尚未成功解决。在本文中，我们提出了一个多模态语义注意模型，以整合视觉特征、预测关键发现嵌入以及临床特征，并逐步解码具有视觉文本语义一致性的报告。首先，从句子解码器中提取多模态特征并处理隐藏状态，对丰富的上下文向量进行编码，以便更好地解码报告。这些模式包括扫描的区域视觉特征、高概率预测的top-K发现的语义词嵌入以及适应症的临床特征。其次，渐进式报表解码器由句子解码器和单词解码器组成，其中我们提出了图像句子匹配和描述精度损失来约束视觉文本语义一致性。在公共MIMIC-CXR和IU X射线数据集上的大量实验表明，我们的模型比最先进的方法取得了一致的改进。

Stereo depth estimation relies on optimal correspondence matching between pixels on epipolar lines in the left and right images to infer depth. In this work, we revisit the problem from a sequence-to-sequence correspondence perspective to replace cost volume construction with dense pixel matching using position information and attention. This approach, named STereo TRansformer (STTR), has several advantages: It 1) relaxes the limitation of a fixed disparity range, 2) identifies occluded regions and provides confidence estimates, and 3) imposes uniqueness constraints during the matching process. We report promising results on both synthetic and real-world datasets and demonstrate that STTR generalizes across different domains, even without fine-tuning.

立体深度估计依赖于左右图像中极线上像素之间的最佳对应匹配来推断深度。在这项工作中，我们从序列到序列对应的角度重新审视这个问题，以使用位置信息和注意力的密集像素匹配替换成本-体积构造。这种称为立体变换器 (STTR) 的方法有几个优点：1) 放宽了固定视差范围的限制；2) 识别遮挡区域并提供置信度估计；3) 在匹配过程中施加唯一性约束。我们在合成数据集和真实数据集上都报告了有希望的结果，并证明STTR可以在不同的领域中推广，即使没有微调。

Modern cameras are equipped with a wide array of sensors that enable recording the geospatial context of an image. Taking advantage of this, we explore depth estimation under the assumption that the camera is geocalibrated, a problem we refer to as geo-enabled depth estimation. Our key insight is that if capture location is known, the corresponding overhead viewpoint offers a valuable resource for understanding the scale of the scene. We propose an end-to-end architecture for depth estimation that uses geospatial context to infer a synthetic ground-level depth map from a co-located overhead image, then fuses it inside of an encoder/decoder style segmentation network. To support evaluation of our methods, we extend a recently released dataset with overhead imagery and corresponding height maps. Results demonstrate that integrating geospatial context significantly reduces error compared to baselines, both at close ranges and when evaluating at much larger distances than existing benchmarks consider.

现代相机配备了大量传感器，能够记录图像的地理空间背景。利用这一点，我们在假设相机经过地理校准的情况下探索深度估计，我们称之为地理启用深度估计问题。我们的主要见解是，如果捕获位置已知，则相应的头顶视点为理解场景的规模提供了宝贵的资源。我们提出了一种用于深度估计的端到端架构，该架构使用地理空间上下文从共同定位的头顶图像推断合成的地面深度图，然后将其融合到编码器/解码器样式的分割网络中。为了支持对我们方法的评估，我们扩展了一个最近发布的数据集，其中包含头顶图像和相应的高度图。结果表明，集成地理空间上下文显著地降低了与基线相比的误差，无论是在近距离，还是在比现有基准考虑更大的距离时评估。

We present a simple yet highly generalizable method for explaining interacting parts within a neural network's reasoning process. First, we design an algorithm based on cross derivatives for computing statistical interaction effects between individual features, which is generalized to both 2-way and higher-order (3-way or more) interactions. We present results side by side with a weight-based attribution technique, corroborating that cross derivatives are a superior metric for both 2-way and higher-order interaction detection. Moreover, we extend the use of cross derivatives as an explanatory device in neural networks to the computer vision setting by expanding Grad-CAM, a popular gradient-based explanatory tool for CNNs, to the higher order. While Grad-CAM can only explain the importance of individual objects in images, our method, which we call Taylor-CAM, can explain a neural network's relational reasoning across multiple objects. We show the success of our explanations both qualitatively and quantitatively, including with a user study. We will release all code as a tool package to facilitate explainable deep learning.

我们提出了一种简单但高度概括的方法来解释神经网络推理过程中的交互部分。首先，我们设计了一种基于交叉导数的算法来计算单个特征之间的统计交互效应，该算法被推广到双向和高阶（三向或更多）交互。我们使用基于权重的归因技术并排给出了结果，证实了交叉导数是双向和高阶交互检测的一个优越指标。此外，我们将交叉导数作为神经网络中的解释工具扩展到计算机视觉设置，将基于梯度的CNN解释工具Grad CAM扩展到更高阶。Grad CAM只能解释图像中单个对象的重要性，而我们称之为Taylor CAM的方法可以解释神经网络跨多个对象的关系推理。我们从定性和定量两方面展示了我们的解释的成功，包括用户研究。我们将把所有代码作为一个工具包发布，以促进可解释的深入学习。

Compact convolutional neural networks (CNNs) have witnessed exceptional improvements in performance in recent years. However, they still fail to provide the same predictive power as CNNs with a large number of parameters. The diverse and even abundant features captured by the layers is an important characteristic of these successful CNNs. However, differences in this characteristic between large CNNs and their compact counterparts have rarely been investigated. In compact CNNs, due to the limited number of parameters, abundant features are unlikely to be obtained, and feature diversity becomes an essential characteristic. Diverse features present in the activation maps derived from a data point during model inference may indicate the presence of a set of unique descriptors necessary to distinguish between objects of different classes. In contrast, data points with low feature diversity may not provide a sufficient amount of unique descriptors to make a valid prediction; we refer to them as random predictions. Random predictions can negatively impact the optimization process and harm the final performance. This paper proposes addressing the problem raised by random predictions by reshaping the standard cross-entropy to make it biased toward data points with a limited number of unique descriptive features. Our novel Bias Loss focuses the training on a set of valuable data points and prevents the vast number of samples with poor learning features from misleading the optimization process. Furthermore, to show the importance of diversity, we present a family of SkipblockNet models whose architectures are brought to boost the number of unique descriptors in the last layers. Experiments conducted on benchmark datasets demonstrate the superiority of the proposed loss function over the cross-entropy loss. Moreover, our SkipblockNet-M can achieve 1% higher classification accuracy than MobileNetV3 Large with similar computational cost on the ImageNet ILSVRC-2012 classification dataset. The code is available on the link - [https://github.com/lusinlu/biasloss\\_skipblocknet](https://github.com/lusinlu/biasloss_skipblocknet).

近年来，紧凑卷积神经网络（CNN）在性能上有了显著的提高。然而，在大量参数下，它们仍然无法提供与CNN相同的预测能力。层捕获的多样性甚至丰富的特征是这些成功CNN的一个重要特征。然而，大型CNN与紧凑型CNN在这一特性上的差异很少被研究。在紧凑型CNN中，由于参数数量有限，不可能获得丰富的特征，特征多样性成为一个基本特征。在模型推理期间从数据点导出的激活图中存在的不同特征可能表明存在一组唯一描述符，这些描述符是区分不同类别对象所必需的。相比之下，具有低特征多样性的数据点可能无法提供足够数量的唯一描述符来进行有效预测；我们称之为随机预测。随机预测会对优化过程产生负面影响，并损害最终性能。本文提出通过重塑标准交叉熵，使其偏向于具有有限数量独特描述特征的数据点，来解决随机预测提出的问题。我们新颖的偏差损失将训练集中在一组有价值的数据点上，并防止大量学习特性差的样本误导优化过程。此外，为了说明多样性的重要性，我们提出了一系列SkipblockNet模型，其体系结构用于增加最后一层中唯一描述符的数量。在基准数据集上进行的实验表明，所提出的损失函数优于交叉熵损失函数。此外，在ImageNet ILSVRC-2012分类数据集上，我们的SkipblockNet-M可以实现比MobileNetV3高1%的分类精度，且计算成本类似。该代码可在以下链接中找到：[https://github.com/lusinlu/biasloss\\_skipblocknet](https://github.com/lusinlu/biasloss_skipblocknet)。

State-of-the-art approaches for visually-guided audio source separation typically assume sources that have characteristic sounds, such as musical instruments. These approaches often ignore the visual context of these sound sources or avoid modeling object interactions that may be useful to better characterize the sources, especially when the same object class may produce varied sounds from distinct interactions. To address this challenging problem, we propose Audio Visual Scene Graph Segmenter (AVSGS), a novel deep learning model that embeds the visual structure of the scene as a graph and segments this graph into subgraphs, each subgraph being associated with a unique sound obtained by co-segmenting the audio spectrogram. At its core, AVSGS uses a recursive neural network that emits mutually-orthogonal sub-graph embeddings of the visual graph using multi-head attention. These embeddings are used for conditioning an audio encoder-decoder towards source separation. Our pipeline is trained end-to-end via a self-supervised task consisting of separating audio sources using the visual graph from artificially mixed sounds. In this paper, we also introduce an ""in the wild" video dataset for sound source separation that contains multiple non-musical sources, which we call Audio Separation in the wild (ASIW). This dataset is adapted from the AudioCaps dataset, and provides a challenging, natural, and daily-life setting for source separation. Thorough experiments on the proposed ASIW and the standard MUSIC datasets demonstrate state-of-the-art sound separation performance of our method against recent prior approaches.

视觉引导音频源分离的最新方法通常假定源具有特征声音，例如乐器。这些方法通常会忽略这些声源的视觉环境，或避免对可能有助于更好地描述声源的对象交互进行建模，尤其是当同一对象类可能会从不同的交互中产生不同的声音时。为了解决这个具有挑战性的问题，我们提出了视听场景图分割器 (AVSGS)，这是一种新的深度学习模型，它将场景的视觉结构嵌入到一个图中，并将该图分割成子图，每个子图与通过共同分割声谱图获得的唯一声音相关联。在其核心部分，AVSGS使用递归神经网络，该网络使用多头注意发出视觉图形的相互正交子图嵌入。这些嵌入件用于调节音频编码器-解码器以实现源分离。我们的管道通过一个自我监督的任务进行端到端的训练，该任务包括使用视觉图形从人工混合的声音中分离音频源。在本文中，我们还引入了一个用于声源分离的“野生”视频数据集，其中包含多个非音乐源，我们称之为野生音频分离 (ASIW)。此数据集改编自AudioCaps数据集，为源分离提供了一个具有挑战性、自然和日常生活的设置。在所提出的ASIW和标准音乐数据集上的彻底实验证明了我们的方法相对于最近的先前方法具有最先进的声音分离性能。

Explainability for machine learning models has gained considerable attention within the research community given the importance of deploying more reliable machine-learning systems. In computer vision applications, generative counterfactual methods indicate how to perturb a model's input to change its prediction, providing details about the model's decision-making. Current methods tend to generate trivial counterfactuals about a model's decisions, as they often suggest to exaggerate or remove the presence of the attribute being classified. For the machine learning practitioner, these types of counterfactuals offer little value, since they provide no new information about undesired model or data biases. In this work, we identify the problem of trivial counterfactual generation and we propose DiVE to alleviate it. DiVE learns a perturbation in a disentangled latent space that is constrained using a diversity-enforcing loss to uncover multiple valuable explanations about the model's prediction. Further, we introduce a mechanism to prevent the model from producing trivial explanations. Experiments on CelebA and Symbols demonstrate that our model improves the success rate of producing high-quality valuable explanations when compared to previous state-of-the-art methods.

鉴于部署更可靠的机器学习系统的重要性，机器学习模型的可解释性在研究界得到了相当大的关注。在计算机视觉应用中，生成反事实方法指示如何扰动模型的输入以改变其预测，从而提供有关模型决策的详细信息。当前的方法倾向于生成关于模型决策的琐碎的反事实，因为它们经常建议夸大或删除被分类属性的存在。对于机器学习实践者来说，这些类型的反事实没有什么价值，因为它们没有提供关于不希望的模型或数据偏差的新信息。在这项工作中，我们确定了琐碎的反事实生成问题，并提出了缓解这一问题的方法。DiVE学习在分离的潜在空间中的扰动，该空间使用多样性强制损失进行约束，以揭示关于模型预测的多个有价值的解释。此外，我们引入了一种机制来防止模型产生琐碎的解释。在CelebA和Symbols上的实验表明，与以前最先进的方法相比，我们的模型提高了生成高质量有价值解释的成功率。

We present RePOSE, a fast iterative refinement method for 6D object pose estimation. Prior methods perform refinement by feeding zoomed-in input and rendered RGB images into a CNN and directly regressing an update of a refined pose. Their runtime is slow due to the computational cost of CNN, which is especially prominent in multiple-object pose refinement. To overcome this problem, RePOSE leverages image rendering for fast feature extraction using a 3D model with a learnable texture. We call this deep texture rendering, which uses a shallow multi-layer perceptron to directly regress a view-invariant image representation of an object. Furthermore, we utilize differentiable Levenberg-Marquard (LM) optimization to refine a pose fast and accurately by minimizing the feature-metric error between the input and rendered image representations without the need of zooming in. These image representations are trained such that differentiable LM optimization converges within few iterations. Consequently, RePOSE runs at 92 FPS and achieves state-of-the-art accuracy of 51.6% on the Occlusion LineMOD dataset - a 4.1% absolute improvement over the prior art, and comparable result on the YCB-Video dataset with a much faster runtime. The code is available at <https://github.com/sh8/repose>.

我们提出了RePOSE，一种用于6D物体姿态估计的快速迭代细化方法。先前的方法通过将放大的输入和渲染的RGB图像输入CNN并直接回归优化姿势的更新来执行优化。由于CNN的计算成本，它们的运行速度很慢，这在多对象姿态优化中尤为突出。为了克服这个问题，RESTE利用图像渲染，使用具有可学习纹理的3D模型进行快速特征提取。我们称之为深度纹理渲染，它使用浅层多层感知器直接回归对象的视图不变图像表示。此外，我们利用可微Levenberg-Marquard (LM) 优化，通过最小化输入和渲染图像表示之间的特征度量误差，快速准确地优化姿势，而无需放大。这些图像表示经过训练，使得可微LM优化在几次迭代内收敛。因此，Resosel以92 FPS的速度运行，在Occlusion LineMOD数据集上达到了51.6%的最新精度-比现有技术绝对提高了4.1%，在YCB视频数据集上的运行速度更快，结果也相当。该守则可于<https://github.com/sh8/repose>。

Successful active speaker detection requires a three-stage pipeline: (i) audio-visual encoding for all speakers in the clip, (ii) inter-speaker relation modeling between a reference speaker and the background speakers within each frame, and (iii) temporal modeling for the reference speaker. Each stage of this pipeline plays an important role for the final performance of the created architecture. Based on a series of controlled experiments, this work presents several practical guidelines for audio-visual active speaker detection. Correspondingly, we present a new architecture called ASDNet, which achieves a new state-of-the-art on the AVA-ActiveSpeaker dataset with a mAP of 93.5% outperforming the second best with a large margin of 4.7%. Our code and pretrained models are publicly available.

成功的主动说话人检测需要三个阶段的管道：(i) 剪辑中所有说话人的视听编码，(ii) 每个帧中参考说话人和背景说话人之间的说话人间关系建模，以及(iii) 参考说话人的时间建模。该管道的每个阶段都对所创建体系结构的最终性能起着重要作用。基于一系列的控制实验，本文提出了一些实用的音频-视频主动说话人检测准则。相应地，我们提出了一种称为ASDNet的新体系结构，它在AVA-ActiveSpeaker

数据集上实现了一种新的最新技术，其mAP值为93.5%，远远超过了第二好的4.7%。我们的代码和预训练模型是公开的。

Recent methods for visual question answering rely on large-scale annotated datasets. Manual annotation of questions and answers for videos, however, is tedious, expensive and prevents scalability. In this work, we propose to avoid manual annotation and generate a large-scale training dataset for video question answering making use of automatic cross-modal supervision. We leverage a question generation transformer trained on text data and use it to generate question-answer pairs from transcribed video narrations. Given narrated videos, we then automatically generate the HowToVQA69M dataset with 69M video-question-answer triplets. To handle the open vocabulary of diverse answers in this dataset, we propose a training procedure based on a contrastive loss between a video-question multi-modal transformer and an answer transformer. We introduce the zero-shot VideoQA task and show excellent results, in particular for rare answers. Furthermore, we demonstrate our method to significantly outperform the state of the art on MSRVTT-QA, MSVD-QA, ActivityNet-QA and How2QA. Finally, for a detailed evaluation we introduce iVQA, a new VideoQA dataset with reduced language biases and high-quality redundant manual annotations.

最近的可视化问答方法依赖于大规模的带注释的数据集。然而，视频问答的手动注释繁琐、昂贵，并且妨碍了可扩展性。在这项工作中，我们建议避免手动注释，并利用自动跨模式监控生成大规模视频问答训练数据集。我们利用一个在文本数据上训练的问题生成转换器，并使用它从转录的视频叙述中生成问题-答案对。给定叙述视频，然后我们自动生成HowToVQA69M数据集，其中包含69M个视频问答三元组。为了处理这个数据集中不同答案的开放词汇表，我们提出了一个基于视频问题多模态变换器和答案变换器之间对比损失的训练过程。我们介绍了零镜头视频质量保证任务，并展示了出色的结果，特别是对于罕见的答案。此外，我们还证明了我们的方法在MSRVTT-QA、MSVD-QA、ActivityNet QA和How2QA方面显著优于最新技术。最后，为了进行详细的评估，我们引入了iVQA，这是一个新的VideoQA数据集，减少了语言偏见，并提供了高质量的冗余手动注释。

As billions of personal data being shared through social media and network, the data privacy and security have drawn an increasing attention. Several attempts have been made to alleviate the leakage of identity information from face photos, with the aid of, e.g., image obfuscation techniques. However, most of the present results are either perceptually unsatisfactory or ineffective against face recognition systems. Our goal in this paper is to develop a technique that can encrypt the personal photos such that they can protect users from unauthorized face recognition systems but remain visually identical to the original version for human beings. To achieve this, we propose a targeted identity-protection iterative method (TIP-IM) to generate adversarial identity masks which can be overlaid on facial images, such that the original identities can be concealed without sacrificing the visual quality. Extensive experiments demonstrate that TIP-IM provides 95%+ protection success rate against various state-of-the-art face recognition models under practical open-set test scenarios. Besides, we also show the practical and effective applicability of our method on a commercial API service.

随着数十亿个人数据通过社交媒体和网络共享，数据隐私和安全问题越来越受到关注。借助图像模糊技术等手段，已经做出了几次尝试，以缓解面部照片中身份信息的泄露。然而，目前的大多数结果要么在感知上不令人满意，要么对人脸识别系统无效。我们在这篇论文中的目标是开发一种技术，可以加密个人照片，这样他们可以保护用户免受未经授权的人脸识别系统的攻击，但在视觉上与人类的原始版本保持一致。为了实现这一点，我们提出了一种有针对性的身份保护迭代方法 (TIP-IM)，以生成可覆盖在人脸图像上的敌对身份掩码，从而在不牺牲视觉质量的情况下隐藏原始身份。大量实验表明，TIP-IM在

实际开放集测试场景下，针对各种最先进的人脸识别模型提供了95%以上的保护成功率。此外，我们还展示了我们的方法在商业API服务上的实用性和有效性。

We propose UniT, a Unified Transformer model to simultaneously learn the most prominent tasks across different domains, ranging from object detection to natural language understanding and multimodal reasoning. Based on the transformer encoder-decoder architecture, our UniT model encodes each input modality with an encoder and makes predictions on each task with a shared decoder over the encoded input representations, followed by task-specific output heads. The entire model is jointly trained end-to-end with losses from each task. Compared to previous efforts on multi-task learning with transformers, we share the same model parameters across all tasks instead of separately fine-tuning task-specific models and handle a much higher variety of tasks across different domains. In our experiments, we learn 7 tasks jointly over 8 datasets, achieving strong performance on each task with significantly fewer parameters. Our code is available in MMF at <https://mmf.sh>.

我们提出了UniT，一个统一的转换器模型，用于同时学习不同领域中最突出的任务，从目标检测到自然语言理解和多模态推理。基于transformer编码器-解码器架构，我们的单元模型使用编码器对每个输入模态进行编码，并使用编码输入表示的共享解码器对每个任务进行预测，然后是特定于任务的输出头。整个模型是端到端联合训练的，每个任务都有损失。与以前使用transformers进行多任务学习的工作相比，我们在所有任务中共享相同的模型参数，而不是单独微调特定于任务的模型，并在不同领域处理更多种类的任务。在我们的实验中，我们在8个数据集上共同学习了7项任务，在每个任务上都取得了很好的性能，而且参数明显较少。我们的代码以MMF格式提供，网址为<https://mmf.sh>。

Despite exciting progress in pre-training for visual-linguistic (VL) representations, very few aspire to a small VL model. In this paper, we study knowledge distillation (KD) to effectively compress a transformer-based large VL model into a small VL model. The major challenge arises from the inconsistent regional visual tokens extracted from different detectors of Teacher and Student, resulting in the misalignment of hidden representations and attention distributions. To address the problem, we retrain and adapt the Teacher by using the same region proposals from Student's detector while the features are from Teacher's own object detector. With aligned network inputs, the adapted Teacher is capable of transferring the knowledge through the intermediate representations. Specifically, we use the mean square error loss to mimic the attention distribution inside the transformer block and present a token-wise noise contrastive loss to align the hidden state by contrasting with negative representations stored in a sample queue. To this end, we show that our proposed distillation significantly improves the performance of small VL models on image captioning and visual question answering tasks. It reaches 120.8 in CIDEr score on COCO captioning, an improvement of 5.1 over its non-distilled counterpart; and an accuracy of 69.8 on VQA 2.0, a 0.8 gain from the baseline. Our extensive experiments and ablations confirm the effectiveness of VL distillation in both pre-training and fine-tuning stages.

尽管在视觉语言（VL）表征的预训练方面取得了令人兴奋的进展，但很少有人渴望使用小型VL模型。在本文中，我们研究了知识蒸馏（KD），以有效地将基于变压器的大型VL模型压缩为小型VL模型。主要的挑战来自于从教师和学生不同的检测器中提取的不一致的区域视觉标记，导致隐藏表征和注意力分布的错位。为了解决这个问题，我们使用来自学生检测器的相同区域建议对教师进行再培训和调整，而特征来自教师自己的对象检测器。通过对齐网络输入，适应的教师能够通过中间表示传递知识。具体地说，我们使用均方误差损失来模拟转换器块内的注意分布，并通过与存储在样本队列中的否定表示进行对比来呈现令牌噪声对比损失来对齐隐藏状态。为此，我们表明，我们提出的蒸馏显著提高了小型VL模型在图像字幕和视觉问答任务上的性能。它在苹果酒中的COCO字幕得分达到120.8分，比未蒸馏的同类产品

提高了5.1分；VQA 2.0的准确度为69.8，比基线增加0.8。我们的大量实验和烧蚀证实了VL蒸馏在预训练和微调阶段的有效性。

Appearance and motion are two important sources of information in video object segmentation (VOS). Previous methods mainly focus on using simplex solutions, lowering the upper bound of feature collaboration among and across these two cues. In this paper, we study a novel framework, termed the FSNet (Full-duplex Strategy Network), which designs a relational cross-attention module (RCAM) to achieve the bidirectional message propagation across embedding subspaces. Furthermore, the bidirectional purification module (BPM) is introduced to update the inconsistent features between the spatial-temporal embeddings, effectively improving the model robustness. By considering the mutual restraint within the full-duplex strategy, our FSNet performs the cross-modal feature-passing (i.e., transmission and receiving) simultaneously before the fusion and decoding stage, making it robust to various challenging scenarios (e.g., motion blur, occlusion) in VOS. Extensive experiments on five popular benchmarks (i.e., DAVIS16, FBMS, MCL, SegTrack-V2, and DAVSOD19) show that our FSNet outperforms other state-of-the-arts for both the VOS and video salient object detection tasks.

在视频对象分割 (VOS) 中，外观和运动是两个重要的信息源。以前的方法主要侧重于使用单纯形解，降低这两个线索之间和之间特征协作的上限。在本文中，我们研究了一种新的框架，称为FSNet（全双工策略网络），它设计了一个关系交叉注意模块 (RCAM)，以实现嵌入子空间中的双向消息传播。此外，引入双向净化模块 (BPM) 来更新时空嵌入之间的不一致特征，有效地提高了模型的鲁棒性。通过考虑全双工策略中的相互约束，我们的FSNet在融合和解码阶段之前同时执行跨模式特征传递（即，传输和接收），使其对VOS中的各种挑战场景（例如，运动模糊、遮挡）具有鲁棒性。在五个流行基准（即DAVIS16、FBMS、MCL、SegTrack-V2和DAVSOD19）上进行的大量实验表明，我们的FSNet在VOS和视频显著目标检测任务方面都优于其他最先进的技术。

Many computer vision problems face difficulties when imaging through turbulent refractive media (e.g., air and water) due to the refraction and scattering of light. These effects cause geometric distortion that requires either handcrafted physical priors or supervised learning methods to remove. In this paper, we present a novel unsupervised network to recover the latent distortion-free image. The key idea is to model non-rigid distortions as deformable grids. Our network consists of a grid deformer that estimates the distortion field and an image generator that outputs the distortion-free image. By leveraging the positional encoding operator, we can simplify the network structure while maintaining fine spatial details in the recovered images. Our method doesn't need to be trained on labeled data and has good transferability across various turbulent image datasets with different types of distortions. Extensive experiments on both simulated and real-captured turbulent images demonstrate that our method can remove both air and water distortions without much customization.

由于光的折射和散射，当通过湍流折射介质（如空气和水）成像时，许多计算机视觉问题面临困难。这些影响会导致几何失真，需要手工制作的物理先验或监督学习方法来消除。在本文中，我们提出了一种新的无监督网络来恢复潜在的无失真图像。关键思想是将非刚性变形建模为可变形网格。我们的网络由一个网格变形器和一个输出无失真图像的图像生成器组成，网格变形器用于估计失真场。通过利用位置编码算子，我们可以简化网络结构，同时在恢复的图像中保持良好的空间细节。我们的方法不需要对标记数据进行训练，并且在具有不同畸变类型的各种湍流图像数据集之间具有良好的可传递性。在模拟和真实拍摄的湍流图像上进行的大量实验表明，我们的方法可以消除空气和水的畸变，而无需太多定制。

In this paper we propose BlockCopy, a scheme that accelerates pretrained frame-based CNNs to process video more efficiently, compared to standard frame-by-frame processing. To this end, a lightweight policy network determines important regions in an image, and operations are applied on selected regions only, using custom block-sparse convolutions. Features of non-selected regions are simply copied from the preceding frame, reducing the number of computations and latency. The execution policy is trained using reinforcement learning in an online fashion without requiring ground truth annotations. Our universal framework is demonstrated on dense prediction tasks such as pedestrian detection, instance segmentation and semantic segmentation, using both state of the art (Center and Scale Predictor, MGAN, SwiftNet) and standard baseline networks (Mask-RCNN, DeepLabV3+). BlockCopy achieves significant FLOPS savings and inference speedup with minimal impact on accuracy.

在本文中，我们提出了BlockCopy方案，与标准逐帧处理方案相比，该方案加速了基于预训练帧的CNN以更高效地处理视频。为此，轻量级策略网络确定图像中的重要区域，并使用自定义块稀疏卷积仅对选定区域应用操作。非选定区域的特征仅从前一帧复制，减少了计算量和延迟。执行策略使用在线强化学习进行培训，无需地面真相注释。我们的通用框架在密集预测任务（如行人检测、实例分割和语义分割）上得到了演示，同时使用了最新技术（中心和尺度预测器、MGAN、SwiftNet）和标准基线网络（Mask RCNN、DeepLabV3+）。BlockCopy在对准确性影响最小的情况下实现了显著的触发器节省和推理加速。

We consider class incremental learning (CIL) problem, in which a learning agent continuously learns new classes from incrementally arriving training data batches and aims to predict well on all the classes learned so far. The main challenge of the problem is the catastrophic forgetting, and for the exemplar-memory based CIL methods, it is generally known that the forgetting is commonly caused by the classification score bias that is injected due to the data imbalance between the new classes and the old classes (in the exemplar-memory). While several methods have been proposed to correct such score bias by some additional post-processing, e.g., score re-scaling or balanced fine-tuning, no systematic analysis on the root cause of such bias has been done. To that end, we analyze that computing the softmax probabilities by combining the output scores for all old and new classes could be the main cause of the bias. Then, we propose a new CIL method, dubbed as Separated Softmax for Incremental Learning (SS-IL), that consists of separated softmax (SS) output layer combined with task-wise knowledge distillation (TKD) to resolve such bias. Throughout our extensive experimental results on several large-scale CIL benchmark datasets, we show our SS-IL achieves strong state-of-the-art accuracy through attaining much more balanced prediction scores across old and new classes, without any additional post-processing.

我们考虑类增量学习 (CIL) 问题，其中学习代理不断地从递增到达的训练数据批次中学习新的类并且旨在预测迄今为止所学的所有类。该问题的主要挑战是灾难性遗忘，对于基于范例记忆的CIL方法，一般都知道遗忘通常是由分类分数偏差引起的，该偏差是由于新类和旧类（在范例记忆中）之间的数据不平衡引起的。虽然已经提出了几种方法来通过一些额外的后处理来纠正这种分数偏差，例如，分数重新调整或平衡微调，但尚未对这种偏差的根本原因进行系统分析。为此，我们分析通过合并所有新旧类别的输出分数来计算softmax概率可能是产生偏差的主要原因。然后，我们提出了一种新的CIL方法，称为增量学习分离Softmax (SS-IL)，该方法由分离Softmax (SS) 输出层与任务智能知识提取 (TKD) 相结合来解决这种偏差。通过我们在几个大规模CIL基准数据集上的大量实验结果，我们表明，SS-IL通过在新旧类中获得更平衡的预测分数，而无需任何额外的后处理，从而实现了强大的最先进的准确性。

In this paper, we investigate video summarization in the supervised setting. Since video summarization is subjective to the preference of the end-user, the design of a unique model is limited. In this work, we propose a model that provides personalized video summaries by conditioning the summarization process with predefined categorical user labels referred to as preferences. The underlying method is based on multiple pairwise rankers (called Multi-ranker), where the rankers are trained jointly to provide local summaries as well as a global summarization of a given video. In order to demonstrate the relevance and applications of our method in contrast with a classical global summarizer, we conduct experiments on multiple benchmark datasets, notably through a user study and comparisons with the state-of-art methods in the global video summarization task.

在本文中，我们研究了监督环境下的视频摘要。由于视频摘要受最终用户偏好的影响，因此独特模型的设计受到限制。在这项工作中，我们提出了一个模型，通过使用预定义的分类用户标签（称为首选项）来调节摘要过程，从而提供个性化的视频摘要。基本方法是基于多个成对ranker（称为Multi-ranker），其中ranker被联合训练以提供给定视频的局部摘要以及全局摘要。为了证明我们的方法与经典的全局摘要器相比的相关性和应用，我们在多个基准数据集上进行了实验，特别是通过用户研究以及与全局视频摘要任务中的最新方法的比较。

Acquisition of training data for the standard semantic segmentation is expensive if requiring that each pixel is labeled. Yet, current methods significantly deteriorate in weakly supervised settings, e.g. where a fraction of pixels is labeled or when only image-level tags are available. It has been shown that regularized losses---originally developed for unsupervised low-level segmentation and representing geometric priors on pixel labels---can considerably improve the quality of weakly supervised training. However, many common priors require optimization stronger than gradient descent. Thus, such regularizers have limited applicability in deep learning. We propose a new robust trust region approach for regularized losses improving the state-of-the-art results. Our approach can be seen as a higher-order generalization of the classic chain rule. It allows neural network optimization to use strong low-level solvers for the corresponding regularizers, including discrete ones.

如果需要对每个像素进行标记，则为标准语义分割获取训练数据的成本很高。然而，当前的方法在弱监督环境下会显著恶化，例如，标记了一小部分像素或者只有图像级标记可用。研究表明，正则化损失——最初用于无监督低层分割，并在像素标签上表示几何先验——可以显著提高弱监督训练的质量。然而，许多常见的先验知识需要比梯度下降更强的优化。因此，此类正则化器在深度学习中的适用性有限。我们提出了一种新的鲁棒信赖域方法，用于正则化损失，从而改进了最新的结果。我们的方法可视为经典链式规则的高阶推广。它允许神经网络优化为相应的正则化器（包括离散正则化器）使用强低级解算器。

We present imGHUM, the first holistic generative model of 3D human shape and articulated pose, represented as a signed distance function. In contrast to prior work, we model the full human body implicitly as a function zero-level-set and without the use of an explicit template mesh. We propose a novel network architecture and a learning paradigm, which make it possible to learn a detailed implicit generative model of human pose, shape, and semantics, on par with state-of-the-art mesh-based models. Our model features desired detail for human models, such as articulated pose including hand motion and facial expressions, a broad spectrum of shape variations, and can be queried at arbitrary resolutions and spatial locations. Additionally, our model has attached spatial semantics making it straightforward to establish correspondences between different shape instances, thus enabling applications that are difficult to tackle using classical implicit representations. In extensive experiments, we demonstrate the model accuracy and its applicability to current research problems.

我们提出了imGHUM，第一个三维人体形状和关节姿势的整体生成模型，表示为符号距离函数。与之前的工作不同，我们将整个人体隐式建模为一个函数零水平集，并且不使用显式模板网格。我们提出了一种新的网络架构和学习范式，这使得有可能学习一个详细的隐式生成模型的人体姿势，形状和语义，与先进的基于网格的模型。我们的模型具有人体模型所需的细节，例如关节姿势，包括手部运动和面部表情，广泛的形状变化，并且可以在任意分辨率和空间位置进行查询。此外，我们的模型还附加了空间语义，使得在不同形状实例之间建立对应关系变得简单，从而支持难以使用经典隐式表示的应用程序。在大量的实验中，我们证明了模型的准确性及其对当前研究问题的适用性。

Deep convolutional neural networks (CNNs) have pushed forward the frontier of super-resolution (SR) research. However, current CNN models exhibit a major flaw: they are biased towards learning low-frequency signals. This bias becomes more problematic for the image SR task which targets reconstructing all fine details and image textures. To tackle this challenge, we propose to improve the learning of high-frequency features both locally and globally and introduce two novel architectural units to existing SR models. Specifically, we propose a dynamic high-pass filtering (HPF) module that locally applies adaptive filter weights for each spatial location and channel group to preserve high-frequency signals. We also propose a matrix multi-spectral channel attention (MMCA) module that predicts the attention map of features decomposed in the frequency domain. This module operates in a global context to adaptively recalibrate feature responses at different frequencies. Extensive qualitative and quantitative results demonstrate that our proposed modules achieve better accuracy and visual improvements against state-of-the-art methods on several benchmark datasets.

深卷积神经网络 (CNN) 推动了超分辨率 (SR) 研究的前沿。然而，当前的CNN模型显示出一个主要缺陷：它们倾向于学习低频信号。对于目标为重建所有精细细节和图像纹理的图像SR任务来说，这种偏差变得更加困难。为了应对这一挑战，我们建议改进本地和全球高频特性的学习，并在现有SR模型中引入两种新的体系结构单元。具体而言，我们提出了一种动态高通滤波 (HPF) 模块，该模块对每个空间位置和信道组局部应用自适应滤波器权重，以保留高频信号。我们还提出了一个矩阵多光谱通道注意 (MMCA) 模块，用于预测频域分解特征的注意图。该模块在全局环境下运行，以自适应地重新校准不同频率下的特征响应。大量的定性和定量结果表明，我们提出的模块在几个基准数据集上比最先进的方法实现了更好的准确性和视觉改善。

The generalization capability of deep neural networks has been substantially improved by applying a wide spectrum of regularization methods, e.g., restricting function space, injecting randomness during training, augmenting data, etc. In this work, we propose a simple yet effective regularization method named progressive self-knowledge distillation (PS-KD), which progressively distills a model's own knowledge to soften hard targets (i.e., one-hot vectors) during training. Hence, it can be interpreted within a framework of knowledge distillation as a student becomes a teacher itself. Specifically, targets are adjusted adaptively by combining the ground-truth and past predictions from the model itself. We show that PS-KD provides an effect of hard example mining by rescaling gradients according to difficulty in classifying examples. The proposed method is applicable to any supervised learning tasks with hard targets and can be easily combined with existing regularization methods to further enhance the generalization performance. Furthermore, it is confirmed that PS-KD achieves not only better accuracy, but also provides high quality of confidence estimates in terms of calibration as well as ordinal ranking. Extensive experimental results on three different tasks, image classification, object detection, and machine translation, demonstrate that our method consistently improves the performance of the state-of-the-art baselines.

通过应用广泛的正则化方法，如限制函数空间、在训练期间注入随机性、增加数据等，深度神经网络的泛化能力得到了显著提高，我们提出了一种简单而有效的正则化方法，称为渐进式自我知识提取（PS-KD），该方法在训练过程中逐步提取模型自身的知识以软化硬目标（即一个热向量）。因此，当学生成为教师时，可以在知识提炼的框架内对其进行解释。具体而言，通过结合地面真实值和来自模型本身过去预测自适应调整目标。我们证明了PS-KD通过根据示例分类的难度重新调整梯度来提供硬示例挖掘的效果。该方法适用于任何具有硬目标的有监督学习任务，并且可以很容易地与现有的正则化方法相结合，以进一步提高泛化性能。此外，还证实了PS-KD不仅具有更好的准确性，而且在校准和顺序排序方面提供了高质量的置信度估计。在图像分类、目标检测和机器翻译三种不同任务上的大量实验结果表明，我们的方法持续改进了最先进的基线的性能。

Training a single deep blind model to handle different quality factors for JPEG image artifacts removal has been attracting considerable attention due to its convenience for practical usage. However, existing deep blind methods usually directly reconstruct the image without predicting the quality factor, thus lacking the flexibility to control the output as the non-blind methods. To remedy this problem, in this paper, we propose a flexible blind convolutional neural network, namely FBCNN, that can predict the adjustable quality factor to control the trade-off between artifacts removal and details preservation. Specifically, FBCNN decouples the quality factor from the JPEG image via a decoupler module and then embeds the predicted quality factor into the subsequent reconstructor module through a quality factor attention block for flexible control. Besides, we find existing methods are prone to fail on non-aligned double JPEG images even with only a one-pixel shift, and we thus propose a double JPEG degradation model to augment the training data. Extensive experiments on single JPEG images, more general double JPEG images, and real-world JPEG images demonstrate that our proposed FBCNN achieves favorable performance against state-of-the-art methods in terms of both quantitative metrics and visual quality.

训练一个单一的深盲模型来处理不同的质量因子以去除JPEG图像伪影，由于其便于实际使用，已经引起了广泛的关注。然而，现有的深盲方法通常直接重建图像而不预测图像的质量因子，因此缺乏与非盲方法一样的输出控制灵活性。为了解决这个问题，在本文中，我们提出了一种灵活的盲卷积神经网络，即FBCNN，它可以预测可调的质量因子来控制伪影去除和细节保留之间的权衡。具体而言，FBCNN通过解耦器模块将质量因子从JPEG图像中解耦，然后通过质量因子注意块将预测的质量因子嵌入后续重建器模块，以实现灵活控制。此外，我们发现现有的方法在非对齐双JPEG图像上很容易失败，即使只有一个像素偏移，因此我们提出了双JPEG退化模型来增加训练数据。对单个JPEG图像、更一般的双JPEG图像和真

实JPEG图像进行的大量实验表明，我们提出的FBCNN在量化指标和视觉质量方面都达到了与最先进方法相比的良好性能。

This paper introduces HPNet, a novel deep-learning approach for segmenting a 3D shape represented as a point cloud into primitive patches. The key to deep primitive segmentation is learning a feature representation that can separate points of different primitives. Unlike utilizing a single feature representation, HPNet leverages hybrid representations that combine one learned semantic descriptor, two spectral descriptors derived from predicted geometric parameters, as well as an adjacency matrix that encodes sharp edges. Moreover, instead of merely concatenating the descriptors, HPNet optimally combines hybrid representations by learning combination weights. This weighting module builds on the entropy of input features. The output primitive segmentation is obtained from a mean-shift clustering module. Experimental results on benchmark datasets ANSI and ABCParts show that HPNet leads to significant performance gains from baseline approaches.

本文介绍了HPNet，一种新的深度学习方法，用于将点云表示的三维形状分割为原始面片。深度基元分割的关键是学习能够分离不同基元点的特征表示。与使用单一特征表示不同，HPNet利用混合表示，将一个学习的语义描述符、两个从预测的几何参数衍生的光谱描述符以及对锐边进行编码的邻接矩阵组合在一起。此外，HPNet通过学习组合权重来优化混合表示，而不仅仅是连接描述符。该加权模块基于输入特征的熵。输出基元分割由均值漂移聚类模块获得。在基准数据集ANSI和ABCPart上的实验结果表明，HPNet比基准方法有显著的性能提升。

We contribute to approximate algorithms for the quadratic assignment problem also known as graph matching. Inspired by the success of the fusion moves technique developed for multilabel discrete Markov random fields, we investigate its applicability to graph matching. In particular, we show how fusion moves can be efficiently combined with the dedicated state-of-the-art dual methods that have recently shown superior results in computer vision and bio-imaging applications. As our empirical evaluation on a wide variety of graph matching datasets suggests, fusion moves significantly improve performance of these methods in terms of speed and quality of the obtained solutions. Our method sets a new state-of-the-art with a notable margin with respect to its competitors.

我们致力于二次分配问题（也称为图匹配）的近似算法。受为多标签离散马尔可夫随机场开发的融合移动技术的成功启发，我们研究了其在图匹配中的适用性。特别是，我们展示了如何将融合移动与专用的最先进的双重方法有效地结合起来，这些方法最近在计算机视觉和生物成像应用中显示了优异的结果。正如我们对各种图形匹配数据集的经验评估所表明的那样，融合移动显著提高了这些方法在获得的解决方案的速度和质量方面的性能。我们的方法创造了一种新的技术水平，与竞争对手相比有显著的差距。

This paper targets at fast video moment retrieval (fast VMR), aiming to localize the target moment efficiently and accurately as queried by a given natural language sentence. We argue that most existing VMR approaches can be divided into three modules namely video encoder, text encoder, and cross-modal interaction module, where the last module is the test-time computational bottleneck. To tackle this issue, we replace the cross-modal interaction module with a cross-modal common space, in which moment-query alignment is learned and efficient moment search can be performed. For the sake of robustness in the learned space, we propose a fine-grained semantic distillation framework to transfer knowledge from additional semantic structures. Specifically, we build a semantic role tree that decomposes a query sentence into different phrases (subtrees). A hierarchical semantic-guided attention module is designed to perform message propagation across the whole tree and yield discriminative features. Finally, the important and discriminative semantics are transferred to the common space by a matching-score distillation process. Extensive experimental results on three popular VMR benchmarks demonstrate that our proposed method enjoys the merits of high speed and significant performance.

本文以快速视频矩检索 (fast-VMR) 为研究对象，目的是在给定的自然语言句子中高效、准确地定位目标矩。我们认为，大多数现有的VMR方法可分为三个模块，即视频编码器、文本编码器和跨模式交互模块，其中最后一个模块是测试时间计算瓶颈。为了解决这个问题，我们将跨模态交互模块替换为跨模态公共空间，在该公共空间中学习短查询对齐，并可以执行有效的矩搜索。为了在学习空间中保持鲁棒性，我们提出了一个细粒度语义提取框架来从附加语义结构中转移知识。具体来说，我们构建了一个语义角色树，将查询语句分解为不同的短语（子树）。设计了一个分层语义引导注意模块，用于在整个树上进行信息传播，并产生区分特征。最后，通过匹配分数提取过程将重要语义和区分语义转移到公共空间。在三个流行的VMR基准上的大量实验结果表明，我们提出的方法具有速度快、性能显著的优点。

To learn distinguishable patterns, most of recent works in vehicle re-identification (ReID) struggled to redevelop official benchmarks to provide various supervisions, which requires prohibitive human labors. In this paper, we seek to achieve the similar goal but do not involve more human efforts. To this end, we introduce a novel framework, which successfully encodes both geometric local features and global representations to distinguish vehicle instances, optimized only by the supervision from official ID labels. Specifically, given our insight that objects in ReID share similar geometric characteristics, we propose to borrow self-supervised representation learning to facilitate geometric features discovery. To condense these features, we introduce an interpretable attention module, with the core of local maxima aggregation instead of fully automatic learning, whose mechanism is completely understandable and whose response map is physically reasonable. To the best of our knowledge, we are the first that perform self-supervised learning to discover geometric features. We conduct comprehensive experiments on three most popular datasets for vehicle ReID, i.e., VeRi-776, CityFlow-ReID, and VehicleID. We report our state-of-the-art (SOTA) performances and promising visualization results. We also show the excellent scalability of our approach on other ReID related tasks, i.e., person ReID and multi-target multi-camera (MTMC) vehicle tracking.

为了了解可区分的模式，最近在车辆重新识别（ReID）方面的大多数工作都在努力重新制定官方基准，以提供各种监督，这需要大量人力。在本文中，我们寻求实现类似的目标，但不涉及更多的人力努力。为此，我们引入了一个新的框架，该框架成功地编码了几何局部特征和全局表示，以区分车辆实例，仅通过官方ID标签的监督进行优化。具体地说，鉴于ReID中的对象具有相似的几何特征，我们建议借用自监督表示学习来促进几何特征发现。为了浓缩这些特征，我们引入了一个可解释的注意模块，其核心是局部极大值聚合，而不是全自动学习，其机制是完全可理解的，其响应图是物理上合理的。据我们所知，我们是第一个执行自我监督学习以发现几何特征的公司。我们在三个最流行的车辆ReID数据集上进行了综合实验，即VeRi-776、CityFlow ReID和VehicleID。我们报告了我们最先进的（SOTA）性能和有

希望的可视化结果。我们还展示了我们的方法在其他ReID相关任务上的出色可扩展性，即人员ReID和多目标多摄像机 (MTMC) 车辆跟踪。

Motion, as the most distinct phenomenon in a video to involve the changes over time, has been unique and critical to the development of video representation learning. In this paper, we ask the question: how important is the motion particularly for self-supervised video representation learning. To this end, we compose a duet of exploiting the motion for data augmentation and feature learning in the regime of contrastive learning. Specifically, we present a Motion-focused Contrastive Learning (MCL) method that regards such duet as the foundation. On one hand, MCL capitalizes on optical flow of each frame in a video to temporally and spatially sample the tubelets (i.e., sequences of associated frame patches across time) as data augmentations. On the other hand, MCL further aligns gradient maps of the convolutional layers to optical flow maps from spatial, temporal and spatio-temporal perspectives, in order to ground motion information in feature learning. Extensive experiments conducted on R(2+1)D backbone demonstrate the effectiveness of our MCL. On UCF101, the linear classifier trained on the representations learnt by MCL achieves 81.91% top-1 accuracy, outperforming ImageNet supervised pre-training by 6.78%. On Kinetics-400, MCL achieves 66.62% top-1 accuracy under the linear protocol.

运动，作为视频中最明显的涉及随时间变化的现象，对视频表征学习的发展具有独特性和关键性。在本文中，我们提出了一个问题：运动对于自监督视频表示学习有多重要。为此，我们在对比学习的基础上，将运动应用于数据增强和特征学习。具体而言，我们提出了一个运动聚焦对比学习（MCL）方法，认为这样的二重奏为基础。一方面，MCL利用视频中每一帧的光流在时间和空间上采样小管（即，跨时间的相关帧块序列）作为数据增强。另一方面，MCL进一步从空间、时间和时空角度将卷积层的梯度图与光流图对齐，以便在特征学习中获得地面运动信息。在R (2+1) D主干上进行的大量实验证明了我们的MCL的有效性。在UCF101上，根据MCL学习的表示训练的线性分类器达到81.91%的top-1精度，比ImageNet监督的预训练高出6.78%。在Kinetics-400上，MCL在线性协议下达到了66.62%的top-1精度。

Video captioning is an important vision task and has been intensively studied in the computer vision community. Existing methods that utilize the fine-grained spatial information have achieved significant improvements, however, they either rely on costly external object detectors or do not sufficiently model the spatial/temporal relations. In this paper, we aim at designing a spatial information extraction and aggregation method for video captioning without the need of external object detectors. For this purpose, we propose a Recurrent Region Attention module to better extract diverse spatial features, and by employing Motion-Guided Cross-frame Message Passing, our model is aware of the temporal structure and able to establish high-order relations among the diverse regions across frames. They jointly encourage information communication and produce compact and powerful video representations. Furthermore, an Adjusted Temporal Graph Decoder is proposed to flexibly update video features and model high-order temporal relations during decoding. Experimental results on three benchmark datasets: MSVD, MSR-VTT, and VATEX demonstrate that our proposed method can outperform state-of-the-art methods.

视频字幕是一项重要的视觉任务，在计算机视觉领域得到了广泛的研究。利用细粒度空间信息的现有方法已经取得了显著的改进，但是，它们要么依赖于昂贵的外部对象检测器，要么没有充分地建模空间/时间关系。本文旨在设计一种无需外部目标检测器的视频字幕空间信息提取和聚合方法。为此，我们提出了一个重复区域注意模块来更好地提取不同的空间特征，并且通过使用运动引导的跨帧消息传递，我们的模型能够感知时间结构，并且能够在跨帧的不同区域之间建立高阶关系。它们共同鼓励信息交流，并制作紧凑而强大的视频表示。此外，还提出了一种调整后的时态图解码器，以灵活地更新视频特征，并

在解码过程中建立高阶时态关系模型。在三个基准数据集（MSVD、MSR-VTT和VATEX）上的实验结果表明，我们提出的方法优于最先进的方法。

Conventional face super-resolution methods usually assume testing low-resolution (LR) images lie in the same domain as the training ones. Due to different lighting conditions and imaging hardware, domain gaps between training and testing images inevitably occur in many real-world scenarios. Neglecting those domain gaps would lead to inferior face super-resolution (FSR) performance. However, how to transfer a trained FSR model to a target domain efficiently and effectively has not been investigated. To tackle this problem, we develop a Domain-Aware Pyramid-based Face Super-Resolution network, named DAP-FSR network. Our DAP-FSR is the first attempt to super-resolve LR faces from a target domain by exploiting only a pair of high-resolution (HR) and LR exemplar in the target domain. To be specific, our DAP-FSR firstly employs its encoder to extract the multi-scale latent representations of the input LR face. Considering only one target domain example is available, we propose to augment the target domain data by mixing the latent representations of the target domain face and source domain ones and then feed the mixed representations to the decoder of our DAP-FSR. The decoder will generate new face images resembling the target domain image style. The generated HR faces in turn are used to optimize our decoder to reduce the domain gap. By iteratively updating the latent representations and our decoder, our DAP-FSR will be adapted to the target domain, thus achieving authentic and high-quality upsampled HR faces. Extensive experiments on three benchmarks validate the effectiveness and superior performance of our DAP-FSR compared to the state-of-the-art methods.

传统的人脸超分辨率方法通常假设检测低分辨率 (LR) 图像与训练图像位于同一区域。由于不同的照明条件和成像硬件，在许多真实场景中不可避免地会出现训练图像和测试图像之间的领域差异。忽略这些域间隙将导致较差的面部超分辨率 (FSR) 性能。然而，如何将训练好的FSR模型有效地转移到目标域还没有被研究。为了解决这个问题，我们开发了一个基于域感知金字塔的人脸超分辨率网络，名为DAP-FSR网络。我们的DAP-FSR是通过仅利用目标域中的一对高分辨率 (HR) 和LR示例，首次尝试从目标域超分辨LR面。具体来说，我们的DAP-FSR首先使用其编码器来提取输入LR面的多尺度潜在表示。考虑到只有一个目标域示例可用，我们建议通过混合目标域面和源域面的潜在表示来扩充目标域数据，然后将混合表示馈送到我们的DAP-FSR解码器。解码器将生成与目标域图像样式相似的新人脸图像。生成的HR面依次用于优化我们的解码器，以减少域间隙。通过迭代更新潜在表示和解码器，我们的DAP-FSR将适应目标域，从而获得真实和高质量的上采样HR人脸。在三个基准上的大量实验验证了我们的DAP-FSR与最先进的方法相比的有效性和优越性能。

Instance segmentation in 3D scenes is fundamental in many applications of scene understanding. It is yet challenging due to the compound factors of data irregularity and uncertainty in the numbers of instances. State-of-the-art methods largely rely on a general pipeline that first learns point-wise features discriminative at semantic and instance levels, followed by a separate step of point grouping for proposing object instances. While promising, they have the shortcomings that (1) the second step is not supervised by the main objective of instance segmentation, and (2) their point-wise feature learning and grouping are less effective to deal with data irregularities, possibly resulting in fragmented segmentations. To address these issues, we propose in this work an end-to-end solution of Semantic Superpoint Tree Network (SSTNet) for proposing object instances from scene points. Key in SSTNet is an intermediate, semantic superpoint tree (SST), which is constructed based on the learned semantic features of superpoints, and which will be traversed and split at intermediate tree nodes for proposals of object instances. We also design in SSTNet a refinement module, termed CliqueNet, to prune superpoints that may be wrongly grouped into instance proposals. Experiments on the benchmarks of ScanNet and S3DIS show the efficacy of our proposed method. At the time of submission, SSTNet ranks top on the ScanNet (V2) leaderboard, with 2% higher of mAP than the second best method.

在场景理解的许多应用中，三维场景中的实例分割是基础。由于数据的不规则性和实例数量的不确定性等复合因素，这仍然是一个挑战。最先进的方法在很大程度上依赖于一个通用的管道，该管道首先学习语义和实例级别的点特征，然后是一个单独的点分组步骤，用于提出对象实例。尽管前景看好，但它们存在以下缺点：（1）第二步不受实例分割主要目标的监督，（2）它们的逐点特征学习和分组在处理数据不规则性方面效率较低，可能导致分段。为了解决这些问题，我们在这项工作中提出了一种端到端的解决方案，即语义超点树网络（SSTNet），用于从场景点提出对象实例。SSTNet中的关键是一个中间语义超点树（SST），它是基于学习到的超点语义特征构建的，并将在中间树节点上遍历和拆分，以提出对象实例。我们还在SSTNet中设计了一个称为CliqueNet的细化模块，用于修剪可能被错误地分组到实例建议中的超级点。在ScanNet和S3DIS上的实验表明了该方法的有效性。提交时，SSTNet在ScanNet (V2) 排行榜上排名第一，mAP比第二好的方法高2%。

Current semantic segmentation methods focus only on mining "local" context, i.e., dependencies between pixels within individual images, by context-aggregation modules (e.g., dilated convolution, neural attention) or structure-aware optimization criteria (e.g., IoU-like loss). However, they ignore "global" context of the training data, i.e., rich semantic relations between pixels across different images. Inspired by recent advance in unsupervised contrastive representation learning, we propose a pixel-wise contrastive algorithm for semantic segmentation in the fully supervised setting. The core idea is to enforce pixel embeddings belonging to a same semantic class to be more similar than embeddings from different classes. It raises a pixel-wise metric learning paradigm for semantic segmentation, by explicitly exploring the structures of labeled pixels, which were rarely explored before. Our method can be effortlessly incorporated into existing segmentation frameworks without extra overhead during testing. We experimentally show that, with famous segmentation models (i.e., DeepLabV3, HRNet, OCR) and backbones (i.e., ResNet, HRNet), our method brings performance improvements across diverse datasets (i.e., Cityscapes, PASCAL-Context, COCO-Stuff, CamVid). We expect this work will encourage our community to rethink the current de facto training paradigm in semantic segmentation.

当前的语义分割方法只关注通过上下文聚合模块（例如，扩展卷积、神经注意）或结构感知优化标准（例如，IoU样丢失）挖掘“局部”上下文，即单个图像中像素之间的依赖关系。然而，它们忽略了训练数据的“全局”上下文，即不同图像中像素之间丰富的语义关系。受无监督对比表征学习最新进展的启发，我们提出了一种在全监督环境下进行语义分割的像素级对比算法。其核心思想是使属于同一语义类的像

素嵌入比来自不同类的嵌入更为相似。它提出了一种基于像素的语义分割度量学习范式，通过明确探索以前很少探索的标记像素的结构。我们的方法可以轻松地融入到现有的分割框架中，而不会在测试过程中产生额外的开销。我们的实验表明，利用著名的分割模型（即DeepLabV3、HRNet、OCR）和主干网（即ResNet、HRNet），我们的方法可以在不同的数据集（即城市景观、PASCAL上下文、COCO Stuff、CamVid）上提高性能。我们期望这项工作将鼓励我们的社区重新思考当前语义切分中事实上的训练范式。

Recently, deep Convolutional Neural Networks (CNNs) can achieve human-level performance in edge detection with the rich and abstract edge representation capacities. However, the high performance of CNN based edge detection is achieved with a large pretrained CNN backbone, which is memory and energy consuming. In addition, it is surprising that the previous wisdom from the traditional edge detectors, such as Canny, Sobel, and LBP are rarely investigated in the rapid-developing deep learning era. To address these issues, we propose a simple, lightweight yet effective architecture named Pixel Difference Network (PiDiNet) for efficient edge detection. Extensive experiments on BSDS500, NYUD, and Multicue are provided to demonstrate its effectiveness, and its high training and inference efficiency. Surprisingly, when training from scratch with only the BSDS500 and VOC datasets, PiDiNet can surpass the recorded result of human perception (0.807 vs. 0.803 in ODS F-measure) on the BSDS500 dataset with 100 FPS and less than 1M parameters. A faster version of PiDiNet with less than 0.1M parameters can still achieve comparable performance among state of the arts with 200 FPS. Results on the NYUD and Multicue datasets show similar observations. The codes are available at <https://github.com/zhuoinoulu/pidinet>.

近年来，深度卷积神经网络（CNN）以其丰富的、抽象的边缘表示能力，在边缘检测方面达到了人的水平。然而，基于CNN的边缘检测的高性能是通过一个大的预训练CNN主干来实现的，该主干占用大量内存和能量。此外，令人惊讶的是，在快速发展的深度学习时代，传统边缘检测器（如Canny、Sobel和LBP）以前的智慧很少被研究。为了解决这些问题，我们提出了一种简单、轻量级但有效的架构，称为像素差分网络（PiDiNet），用于有效的边缘检测。在BSDS500、NYUD和Multicue上进行了大量的实验，以证明其有效性、高训练和推理效率。令人惊讶的是，当仅使用BSDS500和VOC数据集从头开始训练时，PiDiNet可以超过BSDS500数据集上记录的人类感知结果（在ODS F-measure中为0.807对0.803），速度为100 FPS，参数小于1M。参数小于0.1M的更快版本的PiDiNet仍然可以以200 FPS的速度在最先进的技术中实现相当的性能。NYUD和Multicue数据集的结果显示了类似的观察结果。代码可以从以下网址获得：<https://github.com/zhuoinoulu/pidinet>。

Recently, some works found an interesting phenomenon that adversarially robust classifiers can generate good images comparable to generative models. We investigate this phenomenon from an energy perspective and provide a novel explanation. We reformulate adversarial example generation, adversarial training, and image generation in terms of an energy function. We find that adversarial training contributes to obtaining an energy function that is flat and has low energy around the real data, which is the key for generative capability. Based on our new understanding, we further propose a better adversarial training method, Joint Energy Adversarial Training (JEAT), which can generate high-quality images and achieve new state-of-the-art robustness under a wide range of attacks. The Inception Score of the images (CIFAR-10) generated by JEAT is 8.80, much better than original robust classifiers (7.50). In particular, we achieve new state-of-the-art robustness on CIFAR-10 (from 57.20% to 62.04%) and CIFAR-100 (from 30.03% to 30.18%) without extra training data.

最近，一些研究发现了一个有趣的现象，即逆向鲁棒分类器可以生成与生成模型相当的好图像。我们从能量的角度研究了这一现象，并给出了一个新颖的解释。我们根据能量函数重新构造了对抗性示例生成、对抗性训练和图像生成。我们发现，对抗性训练有助于获得围绕真实数据的平坦且低能量的能量函数，这是生成能力的关键。基于我们的新理解，我们进一步提出了一种更好的对抗训练方法，联合能量对抗训练 (JEAT)，它可以生成高质量的图像，并在各种攻击下实现新的最先进的鲁棒性。JEAT生成的图像 (CIFAR-10) 的初始分数为 8.80，远好于原始稳健分类器 (7.50)。特别是，在没有额外训练数据的情况下，我们实现了对 CIFAR-10 (从 57.20% 到 62.04%) 和 CIFAR-100 (从 30.03% 到 30.18%) 的最新鲁棒性。

We present a novel framework to learn to convert the per-pixel photometric information at each view into spatially distinctive and view-invariant low-level features, which can be plugged into existing multi-view stereo pipeline for enhanced 3D reconstruction. Both the illumination conditions during acquisition and the subsequent per-pixel feature transform can be jointly optimized in a differentiable fashion. Our framework automatically adapts to and makes efficient use of the geometric information available in different forms of input data. High-quality 3D reconstructions of a variety of challenging objects are demonstrated on the data captured with an illumination multiplexing device, as well as a point light. Our results compare favorably with state-of-the-art techniques.

我们提出了一个新的框架来学习如何将每个视图中的每像素光度信息转换为空间独特且视图不变的低层特征，这些特征可以插入现有的多视图立体管道中进行增强的三维重建。采集期间的照明条件和随后的每像素特征变换都可以以可微的方式联合优化。我们的框架自动适应并有效利用不同形式输入数据中的几何信息。通过照明多路复用设备以及点光源捕获的数据，演示了各种挑战性对象的高质量三维重建。我们的结果与最先进的技术相比是令人满意的。

Light field images contain both angular and spatial information of captured light rays. The rich information of light fields enables straightforward disparity recovery capability but demands high computational cost as well. In this paper, we design a lightweight disparity estimation model with physical-based multi-disparity-scale cost volume aggregation for fast disparity estimation. By introducing a sub-network of edge guidance, we significantly improve the recovery of geometric details near edges and improve the overall performance. We test the proposed model extensively on both synthetic and real-captured datasets, which provide both densely and sparsely sampled light fields. Finally, we significantly reduce computation cost and GPU memory consumption, while achieving comparable performance with state-of-the-art disparity estimation methods for light fields. Our source code is available at <https://github.com/zcong17huang/FastLFnet>.

光场图像包含捕获光线的角度和空间信息。丰富的光场信息使视差恢复能力变得简单，但同时也需要较高的计算成本。在本文中，我们设计了一个轻量级视差估计模型，该模型采用基于物理的多视差尺度代价-体积聚合来快速估计视差。通过引入边缘引导子网络，我们显著提高了边缘附近几何细节的恢复，提高了整体性能。我们在合成数据集和真实捕获数据集上对所提出的模型进行了广泛的测试，这些数据集提供了密集和稀疏采样的光场。最后，我们显著降低了计算成本和 GPU 内存消耗，同时实现了与最先进的光场视差估计方法相当的性能。我们的源代码可在<https://github.com/zcong17huang/FastLFnet>。

The simultaneous recognition of multiple objects in one image remains a challenging task, spanning multiple events in the recognition field such as various object scales, inconsistent appearances, and confused inter-class relationships. Recent research efforts mainly resort to the statistic label co-occurrences and linguistic word embedding to enhance the unclear semantics. Different from these researches, in this paper, we propose a novel Transformer-based Dual Relation learning framework, constructing complementary relationships by exploring two aspects of correlation, i.e., structural relation graph and semantic relation graph. The structural relation graph aims to capture long-range correlations from object context, by developing a cross-scale transformer-based architecture. The semantic graph dynamically models the semantic meanings of image objects with explicit semantic-aware constraints. In addition, we also incorporate the learnt structural relationship into the semantic graph, constructing a joint relation graph for robust representations. With the collaborative learning of these two effective relation graphs, our approach achieves new state-of-the-art on two popular multi-label recognition benchmarks, i.e. MS-COCO and VOC 2007 dataset.

同时识别一幅图像中的多个对象仍然是一项具有挑战性的任务，它跨越了识别领域中的多个事件，例如不同的对象比例、不一致的外观和混乱的类间关系。近年来的研究主要借助于统计标记共现和语言词嵌入来增强语义不清。与这些研究不同，本文提出了一种新的基于变换器的双关系学习框架，通过探索结构关系图和语义关系图两个方面的相关性来构建互补关系。结构关系图旨在通过开发基于跨尺度变换器的体系结构，从对象上下文捕获长期相关性。语义图通过明确的语义感知约束动态地建模图像对象的语义。此外，我们还将学到的结构关系整合到语义图中，构建了鲁棒表示的联合关系图。通过对这两个有效关系图的协作学习，我们的方法在两个流行的多标签识别基准（即MS-COCO和VOC 2007数据集）上实现了新的技术水平。

We introduce a parallel version of the Quadratic Pseudo-Boolean Optimization (QPBO) algorithm for solving binary optimization tasks, such as image segmentation. The original QPBO implementation by Kolmogorov and Rother relies on the Boykov-Kolmogorov (BK) maxflow/mincut algorithm and performs well for many image analysis tasks. However, the serial nature of their QPBO algorithm results in poor utilization of modern hardware. By redesigning the QPBO algorithm to work with parallel maxflow/mincut algorithms, we significantly reduce solve time of large optimization tasks. We compare our parallel QPBO implementation to other state-of-the-art solvers and benchmark them on two large segmentation tasks and a substantial set of small segmentation tasks. The results show that our parallel QPBO algorithm is over 20 times faster than the serial QPBO algorithm on the large tasks and over three times faster for the majority of the small tasks. Although we focus on image segmentation, our algorithm is generic and can be used for any QPBO problem. Our implementation and experimental results are available at DOI: 10.5281/zenodo.5201620

我们介绍了二次伪布尔优化 (QPBO) 算法的并行版本，用于解决二值优化任务，如图像分割。Kolmogorov和Rother最初的QPBO实现依赖于Boykov-Kolmogorov (BK) maxflow/mincut算法，并在许多图像分析任务中表现良好。然而，其QPBO算法的串行性导致现代硬件的利用率低下。通过将QPBO算法重新设计为与并行maxflow/mincut算法协同工作，我们显著减少了大型优化任务的求解时间。我们将我们的并行QPBO实现与其他最先进的解算器进行比较，并在两个大型分段任务和大量小型分段任务上对它们进行基准测试。结果表明，我们的并行QPBO算法在大任务上比串行QPBO算法快20倍以上，在大多数小任务上比串行QPBO算法快3倍以上。虽然我们专注于图像分割，但我们的算法是通用的，可以用于任何QPBO问题。我们的实现和实验结果可在DOI:10.5281/zenodo上获得。5201620

Human-oriented image captioning with both high diversity and accuracy is a challenging task in vision+language modeling. The reinforcement learning (RL) based frameworks promote the accuracy of image captioning, yet seriously hurt the diversity. In contrast, other methods based on variational auto-encoder (VAE) or generative adversarial network (GAN) can produce diverse yet less accurate captions. In this work, we devote our attention to promote the diversity of RL-based image captioning. To be specific, we devise a partial off-policy learning scheme to balance accuracy and diversity. First, we keep the model exposed to varied candidate captions by sampling from the initial state before RL launched. Second, a novel criterion named max-CIDEr is proposed to serve as the reward for promoting diversity. We combine the above-mentioned off-policy strategy with the on-policy one to moderate the exploration effect, further balancing the diversity and accuracy for human-like image captioning. Experiments show that our method locates the closest to human performance in the diversity-accuracy space, and achieves the highest Pearson correlation as 0.337 with human performance.

在视觉+语言建模中，具有高度多样性和准确性的面向人的图像字幕是一项具有挑战性的任务。基于强化学习 (RL) 的框架提高了图像字幕的准确性，但严重损害了多样性。相比之下，其他基于可变自动编码器 (VAE) 或生成性对抗网络 (GAN) 的方法可以产生多样但不太准确的字幕。在这项工作中，我们致力于促进基于RL的图像字幕的多样性。具体而言，我们设计了一个部分非政策学习计划，以平衡准确性和多样性。首先，我们通过从RL启动前的初始状态采样，使模型暴露于各种候选标题下。第二，提出了一个新的标准max-CIDEr作为促进多样性的奖励。我们将上述非策略策略与策略策略相结合，以缓和探索效果，进一步平衡类人图像字幕的多样性和准确性。实验表明，我们的方法在多样性精度空间中最接近人的绩效，并且与人的绩效的Pearson相关性最高，为0.337。

Instance segmentation methods require large datasets with expensive and thus limited instance-level mask labels. Partially supervised instance segmentation aims to improve mask prediction with limited mask labels by utilizing the more abundant weak box labels. In this work, we show that a class agnostic mask head, commonly used in partially supervised instance segmentation, has difficulties learning a general concept of foreground for the weakly annotated classes using box supervision only. To resolve this problem, we introduce an object mask prior (OMP) that provides the mask head with the general concept of foreground implicitly learned by the box classification head under the supervision of all classes. This helps the class agnostic mask head to focus on the primary object in a region of interest (RoI) and improves generalization to the weakly annotated classes. We test our approach on the COCO dataset using different splits of strongly and weakly supervised classes. Our approach significantly improves over the Mask R-CNN baseline and obtains competitive performance with the state-of-the-art, while offering a much simpler architecture.

实例分割方法需要具有昂贵且有限的实例级掩码标签的大型数据集。部分监督实例分割的目的是利用更丰富的弱盒标签，提高有限掩码标签的掩码预测能力。在这项工作中，我们证明了一个类不可知的掩码头，通常用于部分监督的实例分割，对于仅使用框监督的弱注释类，很难学习前景的一般概念。为了解决这个问题，我们引入了一种对象掩码先验 (OMP)，它为掩码头提供了由盒分类头在所有类的监督下隐式学习的前景的一般概念。这有助于类不可知遮罩头部关注感兴趣区域 (RoI) 中的主要对象，并改进对弱注释类的泛化。我们使用强监督类和弱监督类的不同拆分在COCO数据集上测试我们的方法。我们的方法大大改进了Mask R-CNN基线，获得了与最新技术相媲美的性能，同时提供了更简单的体系结构。

Casual photography is often performed in uncontrolled lighting that can result in low quality images and degrade the performance of downstream processing. We consider the problem of estimating surface normal and reflectance maps of scenes depicting people despite these conditions by supplementing the available visible illumination with a single near infrared (NIR) light source and camera, a so-called "dark flash image". Our method takes as input a single color image captured under arbitrary visible lighting and a single dark flash image captured under controlled front-lit NIR lighting at the same viewpoint, and computes a normal map, a diffuse albedo map, and a specular intensity map of the scene. Since ground truth normal and reflectance maps of faces are difficult to capture, we propose a novel training technique that combines information from two readily available and complementary sources: a stereo depth signal and photometric shading cues. We evaluate our method over a range of subjects and lighting conditions and describe two applications: optimizing stereo geometry and filling the shadows in an image.

随意摄影通常在不受控制的照明下进行，这可能会导致低质量的图像并降低下游处理的性能。我们考虑的问题估计表面正常和反射地图的场景描绘人，尽管这些条件补充现有的可见光照明与一个单一的近红外 (NIR) 光源和相机，所谓的“暗闪光图像”。我们的方法将在任意可见光照明下捕获的单色图像和在同一视点的受控前照NIR照明下捕获的单色闪光图像作为输入，并计算场景的法线贴图、漫反射反照率贴图和镜面反射强度贴图。由于人脸的地真实法线和反射贴图难以捕获，我们提出了一种新的训练技术，该技术结合了来自两个现成的互补源的信息：立体深度信号和光度着色线索。我们在一系列主题和照明条件下评估了我们的方法，并描述了两个应用：优化立体几何和填充图像中的阴影。

Reconstructing dynamic, time-varying scenes with computed tomography (4D-CT) is a challenging and ill-posed problem common to industrial and medical settings. Existing 4D-CT reconstructions are designed for sparse sampling schemes that require fast CT scanners to capture multiple, rapid revolutions around the scene in order to generate high quality results. However, if the scene is moving too fast, then the sampling occurs along a limited view and is difficult to reconstruct due to spatiotemporal ambiguities. In this work, we design a reconstruction pipeline using implicit neural representations coupled with a novel parametric motion field warping to perform limited view 4D-CT reconstruction of rapidly deforming scenes. Importantly, we utilize a differentiable analysis-by-synthesis approach to compare with captured x-ray sinogram data in a self-supervised fashion. Thus, our resulting optimization method requires no training data to reconstruct the scene. We demonstrate that our proposed system robustly reconstructs scenes containing deformable and periodic motion and validate against state-of-the-art baselines. Further, we demonstrate an ability to reconstruct continuous spatiotemporal representations of our scenes and upsample them to arbitrary volumes and frame rates post-optimization. This research opens a new avenue for implicit neural representations in computed tomography reconstruction in general. Code is available at <https://github.com/awreed/DynamicCTReconstruction>.

使用计算机断层扫描 (4D-CT) 重建动态、时变场景是工业和医疗环境中常见的一个具有挑战性且不适当的问题。现有的4D-CT重建是为稀疏采样方案设计的，需要快速CT扫描仪捕捉场景周围的多次快速旋转，以生成高质量的结果。但是，如果场景移动过快，则采样将沿有限的视图进行，并且由于时空模糊性而难以重建。在这项工作中，我们设计了一个重建管道，使用隐式神经表示和一种新的参数运动场扭曲来执行快速变形场景的有限视图4D-CT重建。重要的是，我们利用综合微分分析方法，以自我监督的方式与捕获的x射线正弦图数据进行比较。因此，我们得到的优化方法不需要训练数据来重建场景。我们证明了我们提出的系统能够稳健地重建包含可变形和周期运动的场景，并根据最先进的基线进行验证。此外，我们还展示了重建场景连续时空表示的能力，并在优化后将其采样到任意体积和帧速率。这项研

究为CT重建中的隐式神经表征开辟了一条新的途径。代码可在<https://github.com/awreed/DynamicCTReconstruction>。

Recent advances in 3D perception have shown impressive progress in understanding geometric structures of 3D shapes and even scenes. Inspired by these advances in geometric understanding, we aim to imbue image-based perception with representations learned under geometric constraints. We introduce an approach to learn view-invariant, geometry-aware representations for network pre-training, based on multi-view RGB-D data, that can then be effectively transferred to downstream 2D tasks. We propose to employ contrastive learning under both multi-view image constraints and image-geometry constraints to encode 3D priors into learned 2D representations. This results not only in improvement over 2D-only representation learning on the image-based tasks of semantic segmentation, instance segmentation, and object detection on real-world indoor datasets, but moreover, provides significant improvement in the low data regime. We show a significant improvement of 6.0% on semantic segmentation on full data as well as 11.9% on 20% data against our baselines on ScanNet.

3D感知的最新进展表明，在理解3D形状甚至场景的几何结构方面取得了令人印象深刻的进展。受几何理解的这些进展的启发，我们的目标是将几何约束下学习的表征融入基于图像的感知。我们介绍了一种基于多视图RGB-D数据学习视图不变、几何感知的网络预训练表示的方法，该方法可以有效地传输到下游2D任务。我们建议在多视图图像约束和图像几何约束下使用对比学习将3D先验编码为学习的2D表示。这不仅改善了基于图像的语义分割、实例分割和真实室内数据集目标检测任务的二维表示学习，而且还显著改善了低数据区域。我们在ScanNet上的基线上显示，完整数据的语义分割显著提高了6.0%，20%数据的语义分割显著提高了11.9%。

Accurate camera pose estimation or global camera re-localization is a core component in Structure-from-Motion (SfM) and SLAM systems. Given pair-wise relative camera poses, pose-graph optimization (PGO) involves solving for an optimized set of globally-consistent absolute camera poses. In this work, we propose a novel PGO scheme fueled by graph neural networks (GNN), namely PoGO-Net, to conduct the absolute camera pose regression leveraging multiple rotation averaging (MRA). Specifically, PoGO-Net takes a noisy view-graph as the input, where the nodes and edges are designed to encode the geometric constraints and local graph consistency. Besides, we address the outlier edge removal by exploiting an implicit edge-dropping scheme where the noisy or corrupted edges are effectively filtered out with parameterized networks. Furthermore, we introduce a joint loss function embedding MRA formulation such that the robust inference is capable of achieving real-time performances even for large-scale scenes. Our proposed network is trained end-to-end on public benchmarks, outperforming state-of-the-art approaches in extensive experiments that demonstrate the efficiency and robustness of our proposed network.

精确的摄像机姿态估计或全局摄像机重新定位是运动结构 (SfM) 和SLAM系统的核心组成部分。给定成对的相对相机姿势，姿势图优化 (PGO) 涉及求解一组全局一致的绝对相机姿势。在这项工作中，我们提出了一种新的基于图神经网络 (GNN) 的PGO方案，即PoGO网络，利用多重旋转平均 (MRA) 进行绝对相机姿态回归。具体来说，PoGO网络以一个有噪声的视图图作为输入，其中节点和边被设计为对几何约束和局部图一致性进行编码。此外，我们利用隐式边缘删除方案解决了孤立点边缘去除问题，该方案通过参数化网络有效地过滤掉噪声或损坏的边缘。此外，我们引入了一种嵌入MRA公式的联合损失函数，使得鲁棒推理即使在大规模场景中也能够实现实时性能。我们提出的网络是在公共基准上进行端到端训练的，在广泛的实验中表现优于最先进的方法，证明了我们提出的网络的效率和鲁棒性。

Federated Learning (FL) aims to establish a shared model across decentralized clients under the privacy-preserving constraint. Despite certain success, it is still challenging for FL to deal with non-IID (non-independent and identical distribution) client data, which is a general scenario in real-world FL tasks. It has been demonstrated that the performance of FL will be reduced greatly under the non-IID scenario, since the discrepant data distributions will induce optimization inconsistency and feature divergence issues. Besides, naively minimizing an aggregate loss function in this scenario may have negative impacts on some clients and thus deteriorate their personal model performance. To address these issues, we propose a Unified Feature learning and Optimization objectives alignment method (FedUFO) for non-IID FL. In particular, an adversary module is proposed to reduce the divergence on feature representation among different clients, and two consensus losses are proposed to reduce the inconsistency on optimization objectives from two perspectives. Extensive experiments demonstrate that our FedUFO can outperform the state-of-the-art approaches, including the competitive one data-sharing method. Besides, FedUFO can enable more reasonable and balanced model performance among different clients.

联邦学习 (FL) 的目标是在隐私保护约束下建立一个跨分散客户端的共享模型。尽管取得了一定的成功，但对于FL来说，处理非IID (非独立且相同的分布) 客户机数据仍然是一个挑战，这是现实世界FL任务中的一个常见场景。已经证明，在非IID场景下，FL的性能将大大降低，因为不同的数据分布将导致优化不一致和特征差异问题。此外，在这种情况下，天真地最小化总损失函数可能会对一些客户产生负面影响，从而恶化他们的个人模型性能。为了解决这些问题，我们针对非IID FL提出了一种统一的特征学习和优化目标对齐方法 (FedUFO)。特别是，提出了一个对抗性模块，以减少不同客户之间在特征表示上的差异，并从两个角度提出了两种共识损失，以减少优化目标的不一致性。大量的实验表明，我们的FedUFO可以超越最先进的方法，包括竞争性的数据共享方法。此外，FedUFO可以使不同客户之间的模型性能更加合理和平衡。

Animals have evolved highly functional visual systems to understand motion, assisting perception even under complex environments. In this paper, we work towards developing a computer vision system able to segment objects by exploiting motion cues, i.e. motion segmentation. To achieve this, we introduce a simple variant of the Transformer to segment optical flow frames into primary objects and the background, which can be trained in a self-supervised manner, i.e. without using any manual annotations. Despite using only optical flow, and no appearance information, as input, our approach achieves superior results compared to previous state-of-the-art self-supervised methods on public benchmarks (DAVIS2016, SegTrackv2, FBMS59), while being an order of magnitude faster. On a challenging camouflage dataset (MoCA), we significantly outperform other self-supervised approaches, and are competitive with the top supervised approach, highlighting the importance of motion cues and the potential bias towards appearance in existing video segmentation models.

动物已经进化出功能强大的视觉系统来理解运动，即使在复杂的环境下也能帮助感知。在本文中，我们致力于开发一个计算机视觉系统，能够利用运动线索分割对象，即运动分割。为了实现这一点，我们引入了一种简单的变换器，将光流帧分割成主要对象和背景，可以以自我监督的方式进行训练，即不使用任何手动注释。尽管只使用光流，不使用外观信息作为输入，但我们的方法在公共基准测试 (DAVIS2016、SegTrackv2、FBMS59) 上取得了比以前最先进的自我监督方法更好的结果，同时速度快了一个数量级。在具有挑战性的伪装数据集 (MoCA) 上，我们的表现明显优于其他自监督方法，并与顶级监督方法竞争，突出了运动线索的重要性以及现有视频分割模型中对外观的潜在偏见。

Document unwarping attempts to undo the physical deformation of the paper and recover a 'flatbed' scanned document-image for downstream tasks such as OCR. Current state-of-the-art relies on global unwarping of the document which is not robust to local deformation changes. Moreover, a global unwarping often produces spurious warping artifacts in less warped regions to compensate for severe warps present in other parts of the document. In this paper, we propose the first end-to-end trainable piece-wise unwarping method that predicts local deformation fields and stitches them together with global information to obtain an improved unwarping. The proposed piece-wise formulation results in 4% improvement in terms of multi-scale structural similarity (MS-SSIM) and shows better performance in terms of OCR metrics, character error rate (CER) and word error rate (WER) compared to the state-of-the-art.

文档取消旋转尝试取消纸张的物理变形，并恢复“平板”扫描文档图像，以执行后续任务，如OCR。当前最先进的技术依赖于文档的全局解卷，这对局部变形变化不太可靠。此外，全局取消扭曲通常会在扭曲较少的区域产生虚假扭曲瑕疵，以补偿文档其他部分中存在的严重扭曲。在本文中，我们提出了第一种端到端可训练的分段解爬方法，该方法预测局部变形场，并将其与全局信息缝合在一起以获得改进的解爬。提出的分段公式在多尺度结构相似性（MS-SSIM）方面提高了4%，并且与最新技术相比，在OCR度量、字符错误率（CER）和字错误率（WER）方面表现出更好的性能。

We derive computed tomography (CT) of a time-varying volumetric scattering object, using a small number of moving cameras. We focus on passive tomography of dynamic clouds, as clouds have a major effect on the Earth's climate. State of the art scattering CT assumes a static object. Existing 4D CT methods rely on a linear image formation model and often on significant priors. In this paper, the angular and temporal sampling rates needed for a proper recovery are discussed. Spatiotemporal CT is achieved using gradient-based optimization, which accounts for the correlation time of the dynamic object content. We demonstrate this in physics-based simulations and on experimental real-world data.

我们使用少量的移动摄像机导出时变体积散射对象的计算机断层扫描（CT）。我们关注动态云的被动层析成像，因为云对地球气候有重大影响。最先进的散射CT假定为静态对象。现有的4D CT方法依赖于线性图像形成模型，并且通常依赖于重要的先验知识。本文讨论了适当恢复所需的角度和时间采样率。时空CT使用基于梯度的优化实现，该优化考虑了动态对象内容的相关时间。我们在基于物理的模拟和真实世界的实验数据中证明了这一点。

Editing raster text is a promising but challenging task. We propose to apply text vectorization for the task of raster text editing in display media, such as posters, web pages, or advertisements. In our approach, instead of applying image transformation or generation in the raster domain, we learn a text vectorization model to parse all the rendering parameters including text, location, size, font, style, effects, and hidden background, then utilize those parameters for reconstruction and any editing task. Our text vectorization takes advantage of differentiable text rendering to accurately reproduce the input raster text in a resolution-free parametric format. We show in the experiments that our approach can successfully parse text, styling, and background information in the unified model, and produces artifact-free text editing compared to a raster baseline.

编辑光栅文本是一项有前途但具有挑战性的任务。我们建议将文本矢量化应用于显示媒体（如海报、网页或广告）中的光栅文本编辑任务。在我们的方法中，我们不在光栅域中应用图像变换或生成，而是学习文本矢量化模型来解析所有渲染参数，包括文本、位置、大小、字体、样式、效果和隐藏背景，然后利用这些参数进行重建和任何编辑任务。我们的文本矢量化利用可差分文本渲染，以无分辨率参数化格式精确再现输入光栅文本。实验表明，与光栅基线相比，我们的方法可以成功地解析统一模型中的文本、样式和背景信息，并生成无伪影的文本编辑。

Generating high-fidelity talking head video by fitting with the input audio sequence is a challenging problem that receives considerable attentions recently. In this paper, we address this problem with the aid of neural scene representation networks. Our method is completely different from existing methods that rely on intermediate representations like 2D landmarks or 3D face models to bridge the gap between audio input and video output. Specifically, the feature of input audio signal is directly fed into a conditional implicit function to generate a dynamic neural radiance field, from which a high-fidelity talking-head video corresponding to the audio signal is synthesized using volume rendering. Another advantage of our framework is that not only the head (with hair) region is synthesized as previous methods did, but also the upper body is generated via two individual neural radiance fields. Experimental results demonstrate that our novel framework can (1) produce high-fidelity and natural results, and (2) support free adjustment of audio signals, viewing directions, and background images. Code is available at <https://github.com/YudongGuo/AD-NeRF>.

利用输入音频序列生成高保真的说话人头部视频是一个具有挑战性的问题，近年来受到了广泛关注。在本文中，我们将借助神经场景表示网络来解决这个问题。我们的方法与现有方法完全不同，现有方法依赖于中间表示，如二维地标或三维人脸模型，以弥合音频输入和视频输出之间的差距。具体地说，将输入音频信号的特征直接输入到一个条件隐函数中，生成一个动态的神经辐射场，从该辐射场出发，使用体绘制技术合成与音频信号对应的高保真说话人头部视频。我们的框架的另一个优点是，不仅像以前的方法那样合成头部（带头发）区域，而且通过两个单独的神经辐射场生成上半身。实验结果表明，我们的新框架可以（1）产生高保真和自然的结果，（2）支持音频信号、观看方向和背景图像的自由调整。代码可在<https://github.com/YudongGuo/AD-NeRF>。

Different from traditional video cameras, event cameras capture asynchronous events stream in which each event encodes pixel location, trigger time, and the polarity of the brightness changes. In this paper, we introduce a novel graph-based framework for event cameras, namely SlideGCN. Unlike some recent graph-based methods that use groups of events as input, our approach can efficiently process data event-by-event, unlock the low latency nature of events data while still maintaining the graph's structure internally. For fast graph construction, we develop a radius search algorithm, which better exploits the partial regular structure of event cloud against k-d tree based generic methods. Experiments show that our method reduces the computational complexity up to 100 times with respect to current graph-based methods while keeping state-of-the-art performance on object recognition. Moreover, we verify the superiority of event-wise processing with our method. When the state becomes stable, we can give a prediction with high confidence, thus making an early recognition.

与传统摄像机不同，事件摄像机捕获异步事件流，其中每个事件对像素位置、触发时间和亮度变化的极性进行编码。在本文中，我们介绍了一种新的基于图形的事件摄影机框架，即SlideGCN。与最近一些使用事件组作为输入的基于图的方法不同，我们的方法可以有效地逐事件处理数据，解锁事件数据的低延迟特性，同时仍然在内部保持图的结构。为了快速构建图，我们开发了一种半径搜索算法，该算法更好地利用了事件云的局部规则结构，而不是基于k-d树的泛型方法。实验表明，与现有的基于图形的方法相比，我们的方法在保持最先进的目标识别性能的同时，将计算复杂度降低了100倍。此外，我们还用我们的方法验证了事件处理的优越性。当状态变得稳定时，我们可以给出一个高置信度的预测，从而进行早期识别。

Learning with noisy labels is an important and challenging task for training accurate deep neural networks. However, some commonly-used loss functions, such as Cross Entropy (CE), always suffer from severe overfitting to noisy labels. Although robust loss functions have been designed, they often encounter underfitting. In this paper, we theoretically prove that any loss will be robust to noisy labels when restricting the output of a network to the set of permutations over any fixed vector. When the fixed vector is one-hot, we only need to constrain the output to be one-hot, which means a discrete image and thus zero gradients almost everywhere. This prohibits gradient-based learning of models. In this work, we introduce two sparse regularization strategies to approximate the one-hot constraint: output sharpening and  $\ell_p$ -norm ( $p \leq 1$ ). Output sharpening directly modifies the output distribution of a network to be sharp by adjusting the "temperature" parameter.  $\ell_p$ -norm plays the role of a regularization term to make the output to be sparse. These two simple strategies guarantee the robustness of arbitrary loss functions while not hindering the fitting ability of networks. Experiments on baseline and real-world datasets demonstrate that the sparse regularization can significantly improve the performance of commonly-used loss functions in the presence of noisy labels, and outperform state-of-the-art methods.

带噪声标签的学习对于训练精确的深层神经网络来说是一项重要且具有挑战性的任务。然而，一些常用的损失函数，如交叉熵（CE），总是遭受严重的过拟合噪声标签。虽然设计了稳健的损失函数，但它们经常会遇到拟合不足的问题。在本文中，我们从理论上证明了当将网络的输出限制为任何固定向量上的置换集时，任何损失对噪声标签都是鲁棒的。当固定向量是一个hot时，我们只需要将输出约束为一个hot，这意味着一个离散的图像，因此几乎所有地方的梯度都为零。这禁止基于梯度的模型学习。在这项工作中，我们引入了两种稀疏正则化策略来逼近一个热约束：输出锐化和 $\ell_p$ -范数（ $p \leq 1$ ）。输出锐化通过调整“温度”参数直接将网络的输出分布修改为锐化。 $\ell_p$ -范数起到正则化项的作用，使输出稀疏。这两种简单的策略保证了任意损失函数的鲁棒性，同时又不影响网络的拟合能力。在基线数据集和真实数据集上的实验表明，稀疏正则化可以显著提高在存在噪声标签的情况下常用损失函数的性能，并且优于现有的方法。

Semantic segmentation is a challenging task in the absence of densely labelled data. Only relying on class activation maps (CAM) with image-level labels provides deficient segmentation supervision. Prior works thus consider pre-trained models to produce coarse saliency maps to guide the generation of pseudo segmentation labels. However, the commonly used off-line heuristic generation process cannot fully exploit the benefits of these coarse saliency maps. Motivated by the significant inter-task correlation, we propose a novel weakly supervised multi-task framework termed as AuxSegNet, to leverage saliency detection and multi-label image classification as auxiliary tasks to improve the primary task of semantic segmentation using only image-level ground-truth labels. Inspired by their similar structured semantics, we also propose to learn a cross-task global pixel-level affinity map from the saliency and segmentation representations. The learned cross-task affinity can be used to refine saliency predictions and propagate CAM maps to provide improved pseudo labels for both tasks. The mutual boost between pseudo label updating and cross-task affinity learning enables iterative improvements on segmentation performance. Extensive experiments demonstrate the effectiveness of the proposed auxiliary learning network structure and the cross-task affinity learning method. The proposed approach achieves state-of-the-art weakly supervised segmentation performance on the challenging PASCAL VOC 2012 and MS COCO benchmarks.

在缺乏密集标记数据的情况下，语义分割是一项具有挑战性的任务。仅依靠带有图像级标签的类激活映射（CAM）提供了不足的分割监督。因此，以前的作品考虑预先训练的模型产生粗显著图，以指导产生伪分割标签。然而，常用的离线启发式生成过程不能充分利用这些粗糙显著性图的优点。基于显著的任务间相关性，我们提出了一种新的弱监督多任务框架AuxSegNet，利用显著性检测和多标签图像分类作为辅助任务，改进了仅使用图像级地面真值标签进行语义分割的主要任务。受其相似结构语义的启发，我们还建议从显著性和分割表示中学习跨任务全局像素级亲和性映射。学习到的跨任务亲和性可用于改进显著性预测和传播CAM映射，从而为两个任务提供改进的伪标签。伪标签更新和跨任务相似性学习之间的相互促进使得分段性能的迭代改进成为可能。大量实验证明了所提出的辅助学习网络结构和跨任务亲和学习方法的有效性。该方法在具有挑战性的PASCAL VOC 2012和MS COCO基准上实现了最先进的弱监督分割性能。

Though 3D object detection from point clouds has achieved rapid progress in recent years, the lack of flexible and high-performance proposal refinement remains a great hurdle for existing state-of-the-art two-stage detectors. Previous works on refining 3D proposals have relied on human-designed components such as keypoints sampling, set abstraction and multi-scale feature fusion to produce powerful 3D object representations. Such methods, however, have limited ability to capture rich contextual dependencies among points. In this paper, we leverage the high-quality region proposal network and a channel-wise Transformer architecture to constitute our two-stage 3D object detection framework (CT3D) with minimal hand-crafted design. The proposed CT3D simultaneously performs proposal-aware embedding and channel-wise context aggregation for the point features within each proposal. Specifically, CT3D uses proposal's keypoints for spatial contextual modelling and learns attention propagation in the encoding module, mapping the proposal to point embeddings. Next, a new channel-wise decoding module enriches the query-key interaction via channel-wise re-weighting to effectively merge multi-level contexts, which contributes to more accurate object predictions. Extensive experiments demonstrate that our CT3D method has superior performance and excellent scalability. Remarkably, CT3D achieves the AP of 81.77% in the moderate car category on the KITTI test 3D detection benchmark, outperforms state-of-the-art 3D detectors.

尽管近年来从点云进行三维目标检测取得了快速的进展，但缺乏灵活和高性能的方案改进仍然是现有最先进的两级检测器的一大障碍。以前关于细化3D方案的工作依赖于人工设计的组件，如关键点采样、集合抽象和多尺度特征融合，以生成强大的3D对象表示。然而，这种方法在捕捉点之间丰富的上下文依赖关系方面能力有限。在本文中，我们利用高质量的区域建议网络和通道转换器架构，以最少的手工设计构建了两阶段的3D对象检测框架（CT3D）。建议的CT3D同时为每个建议中的点特征执行建议感知嵌入和通道方式上下文聚合。具体而言，CT3D使用提案的关键点进行空间上下文建模，并在编码模块中学习注意传播，将提案映射到点嵌入。接下来，一个新的通道解码模块通过通道重加权来丰富查询键交互，从而有效地合并多级上下文，这有助于更精确的对象预测。大量实验表明，我们的CT3D方法具有优越的性能和良好的可扩展性。值得注意的是，在KITTI test 3D检测基准上，CT3D在中等车型类别中的AP为81.77%，优于最先进的3D检测器。

Vector graphic documents present visual elements in a resolution free, compact format and are often seen in creative applications. In this work, we attempt to learn a generative model of vector graphic documents. We define vector graphic documents by a multi-modal set of attributes associated to a canvas and a sequence of visual elements such as shapes, images, or texts, and train variational auto-encoders to learn the representation of the documents. We collect a new dataset of design templates from an online service that features complete document structure including occluded elements. In experiments, we show that our model, named CanvasVAE, constitutes a strong baseline for generative modeling of vector graphic documents.

矢量图形文档以无分辨率、紧凑的格式呈现视觉元素，并且经常出现在创造性应用程序中。在这项工作中，我们试图学习矢量图形文档的生成模型。我们通过与画布和一系列视觉元素（如形状、图像或文本）相关联的多模态属性集来定义矢量图形文档，并训练可变自动编码器来学习文档的表示。我们从一个在线服务中收集了一个新的设计模板数据集，该服务具有完整的文档结构，包括隐藏的元素。在实验中，我们证明了我们的模型CanvasVAE为矢量图形文档的生成建模提供了强大的基础。

Point cloud registration is a key task in many computational fields. Previous correspondence matching based methods require the inputs to have distinctive geometric structures to fit a 3D rigid transformation according to point-wise sparse feature matches. However, the accuracy of transformation heavily relies on the quality of extracted features, which are prone to errors with respect to partiality and noise. In addition, they can not utilize the geometric knowledge of all the overlapping regions. On the other hand, previous global feature based approaches can utilize the entire point cloud for the registration, however they ignore the negative effect of non-overlapping points when aggregating global features. In this paper, we present OMNet, a global feature based iterative network for partial-to-partial point cloud registration. We learn overlapping masks to reject non-overlapping regions, which converts the partial-to-partial registration to the registration of the same shape. Moreover, the previously used data is sampled only once from the CAD models for each object, resulting in the same point clouds for the source and reference. We propose a more practical manner of data generation where a CAD model is sampled twice for the source and reference, avoiding the previously prevalent over-fitting issue. Experimental results show that our method achieves state-of-the-art performance compared to traditional and deep learning based methods. Code is available at <https://github.com/megvii-research/OMNet>.

点云配准是许多计算领域的一项关键任务。以前的基于对应匹配的方法要求输入具有独特的几何结构，以便根据逐点稀疏特征匹配来拟合三维刚体变换。然而，变换的准确性在很大程度上依赖于提取的特征的质量，这些特征在偏好和噪声方面容易出错。此外，它们不能利用所有重叠区域的几何知识。另一方面，以前的基于全局特征的方法可以利用整个点云进行配准，但是它们在聚集全局特征时忽略了非重叠点的负面影响。在本文中，我们提出了OMNet，一种用于部分到部分点云配准的基于全局特征的迭代网络。我们学习重叠掩模来拒绝非重叠区域，这将部分到部分的配准转换为相同形状的配准。此外，以前使用的数据仅从每个对象的CAD模型中采样一次，从而产生源和参考的相同点云。我们提出了一种更实用的数据生成方法，其中CAD模型为源和参考采样两次，避免了以前普遍存在的过度拟合问题。实验结果表明，与传统的和基于深度学习的方法相比，我们的方法实现了最先进的性能。代码可在<https://github.com/megvii-research/OMNet>。

The unsupervised domain adaptation (UDA) has been widely adopted to alleviate the data scalability issue, while the existing works usually focus on classifying independently discrete labels. However, in many tasks (e.g., medical diagnosis), the labels are discrete and successively distributed. The UDA for ordinal classification requires inducing non-trivial ordinal distribution prior to the latent space. Target for this, the partially ordered set (poset) is defined for constraining the latent vector. Instead of the typically i.i.d. Gaussian latent prior, in this work, a recursively conditional Gaussian (RCG) set is adapted for ordered constraint modeling, which admits a tractable joint distribution prior. Furthermore, we are able to control the density of content vector that violates the poset constraints by a simple "three-sigma rule". We explicitly disentangle the cross-domain images into a shared ordinal prior induced ordinal content space and two separate source/target ordinal-unrelated spaces, and the self-training is worked on the shared space exclusively for ordinal-aware domain alignment. Extensive experiments on UDA medical diagnoses and facial age estimation demonstrate its effectiveness.

无监督域适配 (UDA) 已被广泛采用以缓解数据可伸缩性问题，而现有的工作通常集中于对独立的离散标签进行分类。然而，在许多任务（例如，医疗诊断）中，标签是离散的且连续分布的。序数分类的UDA要求在潜在空间之前诱导非平凡序数分布。为此，定义了偏序集（偏序集）来约束潜在向量。在这项工作中，一个递归条件高斯 (RCG) 集代替了典型的i.i.d.高斯潜在先验，适用于有序约束建模，它允许一个可处理的联合分布先验。此外，我们能够通过一个简单的“三西格玛规则”来控制违反偏序集约束的内容向量的密度。我们明确地将跨域图像分解为一个共享的有序先验诱导有序内容空间和两个独立的源/目标有序无关空间，并在共享空间上进行自我训练，专门用于有序感知域对齐。大量的UDA医学诊断和面部年龄估计实验证明了其有效性。

Monocular 3D object detection has received increasing attention due to the wide application in autonomous driving. Existing works mainly focus on introducing geometry projection to predict depth priors for each object. Despite their impressive progress, these methods neglect the geometry leverage effect of the projection process, which leads to uncontrollable inferences and damage the training efficiency. In this paper, we propose a Geometry Uncertainty Projection Network (GUP Net) to handle these problems, which can guide the model to learn more reliable depth outputs. The overall framework combines the uncertainty inference and the hierarchical task learning to reduce the negative effects of the geometry leverage. Specifically, an Uncertainty Geometry Projection module is proposed to obtain the geometry guided uncertainty of the inferred depth, which can not only benefit the geometry learning but also provide more reliable depth inferences to reduce the uncontrollability caused by the geometry leverage. Besides, to reduce the instability in the training process caused by the geometry leverage effect, we propose a Hierarchical Task Learning strategy to control the overall optimization process. This learning algorithm can monitor the situation of each task through a well designed learning situation indicator and adaptively assign the proper loss weights for different tasks according to their learning situation and the hierarchical structure, which can significantly improve the stability and the efficiency of the training process. Extensive experiments demonstrate the effectiveness of the proposed method. The overall model can infer more reliable depth and location information than existing methods, which achieves the state-of-the-art performance on the KITTI benchmark.

由于单目三维目标检测在自动驾驶中的广泛应用，其研究受到越来越多的关注。现有的工作主要集中在引入几何投影来预测每个物体的深度先验。尽管这些方法取得了令人瞩目的进步，但它们忽略了投影过程的几何杠杆效应，从而导致无法控制的推断并损害了训练效率。在本文中，我们提出了一种几何不确定性投影网络 (GUP网络) 来处理这些问题，它可以指导模型学习更可靠的深度输出。整体框架结合了不确定性推理和分层任务学习，以减少几何杠杆的负面影响。具体地说，提出了一种不确定性几何投影模块来获取几何引导的推断深度不确定性，这不仅有利于几何学习，而且可以提供更可靠的深度推断，以减少几何杠杆引起的不可控性。此外，为了减少几何杠杆效应导致的训练过程的不稳定性，我们提出了一种分层任务学习策略来控制整个优化过程。该学习算法通过一个设计良好的学习状态指示器来监控每个任务的状态，并根据不同任务的学习状态和层次结构自适应地为其分配适当的损失权重，从而显著提高训练过程的稳定性和效率。大量实验证明了该方法的有效性。整体模型可以推断出比现有方法更可靠的深度和位置信息，从而在KITTI基准上实现了最先进的性能。

Scanning transmission electron microscopy (STEM) is a powerful technique in high-resolution atomic imaging of materials. Decreasing scanning time and reducing electron beam exposure with an acceptable signal-to-noise results are two popular research aspects when applying STEM to beam-sensitive materials. Specifically, partially sampling with fixed electron doses is one of the most important solutions, and then the lost information is restored by computational methods. Following successful applications of deep learning in image in-painting, we have developed an encoder-decoder network to reconstruct STEM images in extremely sparse sampling case. In our model, we combine both local pixel information from convolution operators and global texture features, by applying specific filter operations on frequency domain to acquire initial reconstruction and global structure prior. Our method can effectively restore texture structures and be robust in different sampling ratios with Poisson noise. A comprehensive study demonstrates that our method gains about 50% performance enhancement in comparison with the state-of-art methods. Code is available at <https://github.com/icthrm/Sparse-Sampling-Reconstruction>.

扫描透射电子显微镜 (STEM) 是材料高分辨率原子成像的有力技术。在将STEM应用于束流敏感材料时，缩短扫描时间和减少电子束曝光以获得可接受的信噪比结果是两个流行的研究方向。具体来说，固定电子剂量部分采样是最重要的解决方案之一，然后通过计算方法恢复丢失的信息。随着深度学习在绘画图像中的成功应用，我们开发了一个编码器-解码器网络，用于在极稀疏采样情况下重建STEM图像。在我们的模型中，我们结合了卷积算子的局部像素信息和全局纹理特征，通过在频域上应用特定的滤波操作来获得初始重建和全局结构先验信息。我们的方法可以有效地恢复纹理结构，并且在不同的泊松噪声采样率下具有鲁棒性。一项全面的研究表明，与最先进的方法相比，我们的方法的性能提高了约 50%。代码可在<https://github.com/icthrm/Sparse-Sampling-Reconstruction>.

The core of visual place recognition (VPR) lies in how to identify task-relevant visual cues and embed them into discriminative representations. Focusing on these two points, we propose a novel encoding strategy named Attentional Pyramid Pooling of Salient Visual Residuals (APPSVR). It incorporates three types of attention modules to model the saliency of local features in individual, spatial and cluster dimensions respectively. (1) To inhibit task-irrelevant local features, a semantic-reinforced local weighting scheme is employed for local feature refinement; (2) To leverage the spatial context, an attentional pyramid structure is constructed to adaptively encode regional features according to their relative spatial saliency; (3) To distinguish the different importance of visual clusters to the task, a parametric normalization is proposed to adjust their contribution to image descriptor generation. Experiments demonstrate APPSVR outperforms the existing techniques and achieves a new state-of-the-art performance on VPR benchmark datasets. The visualization shows the saliency map learned in a weakly supervised manner is largely consistent with human cognition.

视觉位置识别 (VPR) 的核心在于如何识别与任务相关的视觉线索并将其嵌入到区分性表征中。针对这两点，我们提出了一种新的编码策略——显著视觉残差的注意金字塔池 (APPSVR)。它包含三种类型的注意模块，分别在个体、空间和集群维度上对局部特征的显著性进行建模。（1）为了抑制与任务无关的局部特征，采用语义增强的局部加权方法对局部特征进行细化；（2）为了利用空间背景，构建了注意金字塔结构，根据区域特征的相对空间显著性自适应编码区域特征；（3）为了区分视觉聚类对任务的不同重要性，提出了一种参数归一化方法来调整它们对图像描述符生成的贡献。实验表明，APPSVR 优于现有技术，并在VPR基准数据集上实现了最新的性能。可视化结果表明，弱监督学习的显著性图与人类认知基本一致。

Point cloud segmentation is a fundamental task in 3D. Despite recent progress on point cloud segmentation with the power of deep networks, current deep learning methods based on the clean label assumptions may fail with noisy labels. Yet, object class labels are often mislabeled in real-world point cloud datasets. In this work, we take the lead in solving this issue by proposing a novel Point Noise-Adaptive Learning (PNAL) framework. Compared to existing noise-robust methods on image tasks, our PNAL is noise-rate blind, to cope with the spatially variant noise rate problem specific to point clouds. Specifically, we propose a novel point-wise confidence selection to obtain reliable labels based on the historical predictions of each point. A novel cluster-wise label correction is proposed with a voting strategy to generate the best possible label taking the neighbor point correlations into consideration. We conduct extensive experiments to demonstrate the effectiveness of PNAL on both synthetic and real-world noisy datasets. In particular, even with 60% symmetric noisy labels, our proposed method produces much better results than its baseline counterpart without PNAL and is comparable to the ideal upper bound trained on a completely clean dataset. Moreover, we fully re-labeled the validation set of a popular but noisy real-world scene dataset ScanNetV2 to make it clean, for rigorous experiment and future research. Our code and data will be released.

点云分割是三维建模中的一项基本任务。尽管最近利用深度网络的力量在点云分割方面取得了进展，但基于清洁标签假设的当前深度学习方法可能会因标签噪声而失败。然而，在现实世界的点云数据集中，对象类标签经常被错误标记。在这项工作中，我们率先提出了一种新的点噪声自适应学习（PNAL）框架来解决这个问题。与现有的抗噪声图像任务方法相比，我们的PNAL是噪声率盲的，能够处理特定于点云的空间变化的噪声率问题。具体地说，我们提出了一种新的逐点置信选择方法，以根据每个点的历史预测获得可靠的标签。提出了一种新的分簇标签校正算法，该算法采用投票策略，在考虑相邻点相关性的情况下生成最佳标签。我们进行了大量实验，以证明PNAL在合成和真实噪声数据集上的有效性。特别是，即使有60%的对称噪声标签，我们提出的方法也比没有PNAL的基线方法产生更好的结果，并且与在完全干净的数据集上训练的理想上界相当。此外，我们完全重新标记了一个流行但嘈杂的真实场景数据集ScanNetV2的验证集，以使其干净，用于严格的实验和未来的研究。我们的代码和数据将被发布。

Goal-conditioned approaches recently have been found very useful to human trajectory prediction, when adequate goal estimates are provided. Yet, goal inference is difficult in itself and often incurs extra learning efforts. We propose to predict pedestrian trajectories via the guidance of goal expertise, which can be obtained with modest expense through a novel goal-search mechanism on already seen training examples. There are three key contributions in our study. First, we devise a framework that exploits the nearest examples for high-quality goal position inquiry. This approach naturally considers multi-modality, physical constraints, compatibility with existing methods and is model-free; it therefore does not require additional learning efforts typical in goal inference. Second, we present an end-to-end trajectory predictor that can efficiently associate goal retrievals to past motion information and dynamically infer possible future trajectories. Third, with these two novel techniques in hand, we conduct a series of experiments on two broadly explored datasets (SDD and ETH/UCY) and show that our approach surpasses previous state-of-the-art performance by notable margins and reduces the need for additional parameters.

最近发现，当提供足够的目标估计时，目标条件方法对于人体轨迹预测非常有用。然而，目标推理本身是困难的，常常需要额外的学习努力。我们建议通过目标专家的指导来预测行人轨迹，这可以通过一种新的目标搜索机制在已经看到的训练示例上以适度的费用获得。我们的研究有三个关键贡献。首先，我们设计了一个框架，利用最近的例子进行高质量的目标位置查询。这种方法自然地考虑了多模态、物理约束、与现有方法的兼容性，并且是无模型的；因此，它不需要额外的学习努力典型的目标推理。其次，我们提出了一种端到端的轨迹预测器，可以有效地将目标检索与过去的运动信息关联起来，并动态

地推断未来可能的轨迹。第三，有了这两种新技术，我们在两个广泛探索的数据集（SDD和ETH/UCY）上进行了一系列实验，结果表明，我们的方法显著优于以前的最先进性能，并减少了对额外参数的需要。

Adversarial training is promising for improving robustness of deep neural networks towards adversarial perturbations, especially on the classification task. The effect of this type of training on semantic segmentation, contrarily, just commences. We make the initial attempt to explore the defense strategy on semantic segmentation by formulating a general adversarial training procedure that can perform decently on both adversarial and clean samples. We propose a dynamic divide-and-conquer adversarial training (DDC-AT) strategy to enhance the defense effect, by setting additional branches in the target model during training, and dealing with pixels with diverse properties towards adversarial perturbation. Our dynamical division mechanism divides pixels into multiple branches automatically. Note all these additional branches can be abandoned during inference and thus leave no extra parameter and computation cost. Extensive experiments with various segmentation models are conducted on PASCAL VOC 2012 and Cityscapes datasets, in which DDC-AT yields satisfying performance under both white- and black-box attack. The code is available at <https://github.com/dvlab-research/Robust-Semantic-Segmentation>.

对抗性训练有望提高深层神经网络对对抗性扰动的鲁棒性，尤其是在分类任务上。相反，这种训练对语义切分的影响才刚刚开始。我们初步尝试通过制定一个通用的对抗性训练程序来探索语义分割的防御策略，该程序可以在对抗性和干净的样本上正常执行。我们提出了一种动态分而治之的对抗性训练（DDC-AT）策略，通过在训练过程中在目标模型中设置额外的分支，以及针对对抗性干扰处理具有不同属性的像素来增强防御效果。我们的动态分割机制自动将像素分割成多个分支。注：所有这些额外的分支都可以在推理过程中被放弃，因此不会留下额外的参数和计算成本。在PASCAL VOC 2012和Cityscapes数据集上对各种分割模型进行了大量实验，其中DDC-AT在白盒和黑盒攻击下都具有令人满意的性能。该守则可于<https://github.com/dvlab-research/Robust-Semantic-Segmentation>。

User data confidentiality protection is becoming a rising challenge in the present deep learning research. Without access to data, conventional data-driven model compression faces a higher risk of performance degradation. Recently, some works propose to generate images from a specific pretrained model to serve as training data. However, the inversion process only utilizes biased feature statistics stored in one model and is from low-dimension to high-dimension. As a consequence, it inevitably encounters the difficulties of generalizability and inexact inversion, which leads to unsatisfactory performance. To address these problems, we propose MixMix based on two simple yet effective techniques: (1) Feature Mixing: utilizes various models to construct a universal feature space for generalized inversion; (2) Data Mixing: mixes the synthesized images and labels to generate exact label information. We prove the effectiveness of MixMix from both theoretical and empirical perspectives. Extensive experiments show that MixMix outperforms existing methods on the mainstream compression tasks, including quantization, knowledge distillation and pruning. Specifically, MixMix achieves up to 4% and 20% accuracy uplift on quantization and pruning, respectively, compared to existing data-free compression work.

在当前的深度学习研究中，用户数据的保密性保护正成为一个日益严峻的挑战。如果无法访问数据，传统的数据驱动模型压缩面临更高的性能降级风险。最近，一些工作建议从特定的预训练模型生成图像作为训练数据。然而，反演过程仅利用存储在一个模型中的有偏特征统计量，并且是从低维到高维的。因此，它不可避免地会遇到推广性和不精确反演的困难，从而导致性能不理想。为了解决这些问题，我们基于两种简单而有效的技术提出了MixMix：（1）特征混合：利用各种模型构造通用特征空间进行广义反演；（2）数据混合：混合合成图像和标签以生成准确的标签信息。我们从理论和实证两方面证明了MixMix的有效性。大量实验表明，MixMix在主流压缩任务（包括量化、知识提取和剪枝）上的性能优

于现有方法。具体来说，与现有的无数据压缩工作相比，MixMix在量化和修剪方面分别实现了高达4%和20%的精度提升。

Computer vision tasks such as object detection and semantic/instance segmentation rely on the painstaking annotation of large training datasets. In this paper, we propose LocTex that takes advantage of the low-cost localized textual annotations (i.e., captions and synchronized mouse-over gestures) to reduce the annotation effort. We introduce a contrastive pre-training framework between images and captions and propose to supervise the cross-modal attention map with rendered mouse traces to provide coarse localization signals. Our learned visual features capture rich semantics (from free-form captions) and accurate localization (from mouse traces), which are very effective when transferred to various downstream vision tasks. Compared with ImageNet supervised pre-training, LocTex can reduce the size of the pre-training dataset by 10x or the target dataset by 2x while achieving comparable or even improved performance on COCO instance segmentation. When provided with the same amount of annotations, LocTex achieves around 4% higher accuracy than the previous state-of-the-art "vision+language" pre-training approach on the task of PASCAL VOC image classification.

计算机视觉任务，如目标检测和语义/实例分割，依赖于对大型训练数据集的精心标注。在本文中，我们提出了LocTex，它利用低成本的本地化文本注释（即标题和同步鼠标移动手势）来减少注释工作量。我们介绍了一种图像和字幕之间的对比预训练框架，并建议使用渲染的鼠标轨迹监控跨模态意图，以提供粗略的定位信号。我们学习到的视觉特征捕捉到丰富的语义（来自自由形式的标题）和精确的定位（来自鼠标轨迹），这在转移到各种下游视觉任务时非常有效。与ImageNet监督的预训练相比，LocTex可以将预训练数据集的大小减少10倍，或将目标数据集的大小减少2倍，同时在COCO实例分割方面实现相当甚至改进的性能。当提供相同数量的注释时，LocTex在PASCAL VOC图像分类任务上的准确率比之前最先进的“视觉+语言”预训练方法高出约4%。

Most supervised image segmentation methods require delicate and time-consuming pixel-level labeling of buildings or objects, especially for small objects. In this paper, we present a weakly supervised segmentation network for aerial/satellite images, separately considering small and large objects. First, we propose a simple point labeling method for small objects, while large objects are fully labeled. Then, we present a segmentation network trained with a small object mask to separate small and large objects in the loss function. During training, we employ a memory bank to cope with the limited number of point labels. Experiments results with three public datasets demonstrate the feasibility of our approach.

大多数有监督的图像分割方法需要对建筑物或物体（尤其是小物体）进行精细且耗时的像素级标记。在本文中，我们提出了一个弱监督的航空/卫星图像分割网络，分别考虑小目标和大目标。首先，我们提出了一种简单的小对象点标记方法，而大对象则完全标记。然后，我们提出了一个用小对象掩模训练的分割网络来分离损失函数中的大小对象。在训练期间，我们使用一个内存库来处理有限数量的点标签。在三个公共数据集上的实验结果证明了该方法的可行性。

outlier rejection and equivalently inlier set optimization is a key ingredient in numerous applications in computer vision such as filtering point-matches in camera pose estimation or plane and normal estimation in point clouds. Several approaches exist, yet at large scale we face a combinatorial explosion of possible solutions and state-of-the-art methods like RANSAC, Hough transform or Branch&Bound require a minimum inlier ratio or prior knowledge to remain practical. In fact, for problems such as camera posing in very large scenes these approaches become useless as they have exponential runtime growth if these conditions aren't met. To approach the problem we present a efficient and general algorithm for outlier rejection based on "intersecting" k-dimensional surfaces in  $R^d$ . We provide a recipe for casting a variety of geometric problems as finding a point in  $R^d$  which maximizes the number of nearby surfaces (and thus inliers). The resulting algorithm has linear worst-case complexity with a better runtime dependency in the approximation factor than competing algorithms while not requiring domain specific bounds. This is achieved by introducing a space decomposition scheme that bounds the number of computations by successively rounding and grouping samples. Our recipe (and open-source code) enables anybody to derive such fast approaches to new problems across a wide range of domains. We demonstrate the versatility of the approach on several camera posing problems with a high number of matches at low inlier ratio achieving state-of-the-art results at significantly lower processing times.

在计算机视觉的许多应用中，如在摄像机姿态估计中过滤点匹配或在点云中过滤平面和法线估计中，离群点抑制和等价内联集优化是一个关键因素。有几种方法存在，但在大范围内，我们面临着可能解决方案的组合爆炸，最先进的方法，如RANSAC、Hough变换或分枝定界，需要最小的内联比或先验知识才能保持实用性。事实上，对于诸如在非常大的场景中设置相机姿势之类的问题，这些方法变得毫无用处，因为如果不满足这些条件，它们的运行时会呈指数增长。为了解决这个问题，我们提出了一种基于 $R^d$ 中k维相交曲面的离群点剔除算法。我们提供了一个解决各种几何问题的方法，如在 $R^d$ 中找到一个点，该点可以最大化附近曲面的数量（从而最大化内联线）。结果算法具有线性最坏情况复杂度，在近似因子方面比竞争算法具有更好的运行时依赖性，同时不需要特定于域的边界。这是通过引入空间分解方案来实现的，该方案通过依次舍入和分组样本来限制计算数量。我们的配方（和开源代码）使任何人都能够在广泛的领域中找到解决新问题的快速方法。我们证明了该方法的多功能性，可以解决在低内联比下大量匹配的多个摄像机问题，从而在显著缩短处理时间的情况下获得最先进的结果。

We introduce DiscoBox, a novel framework that jointly learns instance segmentation and semantic correspondence using bounding box supervision. Specifically, we propose a self-ensembling framework where instance segmentation and semantic correspondence are jointly guided by a structured teacher in addition to the bounding box supervision. The teacher is a structured energy model incorporating a pairwise potential and a cross-image potential to model the pairwise pixel relationships both within and across the boxes. Minimizing the teacher energy simultaneously yields refined object masks and dense correspondences between intra-class objects, which are taken as pseudo-labels to supervise the task network and provide positive/negative correspondence pairs for dense contrastive learning. We show a symbiotic relationship where the two tasks mutually benefit from each other. Our best model achieves 37.9% AP on COCO instance segmentation, surpassing prior weakly supervised methods and is competitive to supervised methods. We also obtain state of the art weakly supervised results on PASCAL VOC12 and PF-PASCAL with real-time inference.

我们介绍了DiscoBox，这是一个新的框架，它使用边界框监控来联合学习实例分割和语义对应。具体来说，我们提出了一个自我理解框架，在这个框架中，除了边界框监督外，实例分割和语义对应由结构化教师共同指导。教师是一个结构化的能量模型，包含一个成对势和一个交叉图像势，以模拟盒子内和盒子之间的成对像素关系。最小化教师能量的同时，产生细化的对象掩码和类内对象之间的稠密对应，作为伪标签来监督任务网络，并为稠密对比学习提供正/负对应。我们展示了一种共生关系，其中两个任

务相互受益。我们的最佳模型在COCO实例分割上达到37.9%的AP，超过了以前的弱监督方法，并且与监督方法具有竞争性。我们还通过实时推理获得了PASCAL VOC12和PF-PASCAL的最新弱监督结果。

Hough voting, as has been demonstrated in VoteNet, is effective for 3D object detection, where voting is a key step. In this paper, we propose a novel VoteNet-based 3D detector with vote enhancement to improve the detection accuracy in cluttered indoor scenes. It addresses the limitations of current voting schemes, i.e., votes from neighboring objects and background have significant negative impacts. Specifically, before voting, we replace the classic MLP with the proposed Attentive MLP (AMLP) in the backbone network to get better feature description of seed points. During voting, we design a new vote attraction loss (VALoss) to enforce vote centers to locate closely and compactly to the corresponding object centers. After voting, we then devise a vote weighting module to integrate the foreground/background prediction into the vote aggregation process to enhance the capability of the original VoteNet to handle noise from background voting. The three proposed strategies all contribute to more effective voting and improved performance, resulting in a novel 3D object detector, termed VENet. Experiments show that our method outperforms state-of-the-art methods on benchmark datasets. Ablation studies demonstrate the effectiveness of the proposed components.

正如VoteNet中所展示的那样，Hough投票对于3D对象检测是有效的，其中投票是一个关键步骤。在本文中，我们提出了一种新的基于VoteNet的3D检测器，该检测器具有投票增强功能，以提高在杂乱的室内场景中的检测精度。它解决了当前投票方案的局限性，即来自相邻对象和背景的投票具有显著的负面影响，具体而言，在投票之前，我们在骨干网络中用建议的关注MLP (AMLP) 替换经典MLP，以获得更好的种子点特征描述。在投票过程中，我们设计了一种新的投票吸引损失 (VALoss) 算法，使投票中心与相应的目标中心紧密地定位。投票后，我们设计了一个投票权重模块，将前景/背景预测集成到投票聚合过程中，以增强原始VoteNet处理背景投票噪声的能力。提出的三种策略都有助于更有效的投票和改进性能，从而产生了一种新的三维对象检测器，称为VENet。实验表明，在基准数据集上，我们的方法优于最新的方法。烧蚀研究证明了所提出组件的有效性。

Aggregating features from different depths of a network is widely adopted to improve the network capability. Lots of modern architectures are equipped with skip connections, which actually makes the feature aggregation happen in all these networks. Since different features tell different semantic meanings, there are inconsistencies and incompatibilities to be solved. However, existing works naively blend deep features via element-wise summation or concatenation with a convolution behind. Better feature aggregation method beyond summation or concatenation is rarely explored. In this paper, given two layers of features to be aggregated together, we first detect and identify where and what needs to be updated in one layer, then replace the feature at the identified location with the information of the other layer. This process, which we call DEtect-rePLace (DEPLA), enables us to avoid inconsistent patterns while keeping useful information in the merged outputs. Experimental results demonstrate our method largely boosts multiple baselines e.g. ResNet, FishNet and FPN on three major vision tasks including ImageNet classification, MS COCO object detection and instance segmentation.

为了提高网络性能，人们广泛采用从网络不同深度聚集特征的方法。许多现代体系结构都配备了跳过连接，这实际上使得所有这些网络中都发生了功能聚合。由于不同的特征具有不同的语义，因此存在不一致和不兼容的问题。然而，现有的作品通过元素的求和或后面的卷积级联，天真地融合了深层特征。除了求和或连接之外，很少有人探索更好的特性聚合方法。在本文中，给定要聚合在一起的两层特征，我们首先检测并确定其中一层中需要更新的位置和内容，然后用另一层的信息替换已识别位置的特征。这个过程，我们称之为DEtect-rePLace (DEPLA)，使我们能够避免不一致的模式，同时在合并的输出中

保留有用的信息。实验结果表明，我们的方法在三个主要的视觉任务（包括ImageNet分类、MS COCO目标检测和实例分割）上大大增强了多基线，例如ResNet、FishNet和FPN。

LiDAR sensors can be used to obtain a wide range of measurement signals other than a simple 3D point cloud, and those signals can be leveraged to improve perception tasks like 3D object detection. A single laser pulse can be partially reflected by multiple objects along its path, resulting in multiple measurements called echoes. Multi-echo measurement can provide information about object contours and semi-transparent surfaces which can be used to better identify and locate objects. LiDAR can also measure surface reflectance (intensity of laser pulse return), as well as ambient light of the scene (sunlight reflected by objects). These signals are already available in commercial LiDAR devices but have not been used in most LiDAR-based detection models. We present a 3D object detection model which leverages the full spectrum of measurement signals provided by LiDAR. First, we propose a multi-signal fusion (MSF) module to combine (1) the reflectance and ambient features extracted with a 2D CNN, and (2) point cloud features extracted using a 3D graph neural network (GNN). Second, we propose a multi-echo aggregation (MEA) module to combine the information encoded in different set of echo points. Compared with traditional single echo point cloud methods, our proposed multi-signal LiDAR Detector (MSLiD) extracts richer context information from a wider range of sensing measurements and achieves more accurate 3D object detection. Experiments show that by incorporating the multi-modality of LiDAR, our method outperforms the state-of-the-art by up to relatively 9.1%.

激光雷达传感器可用于获取除简单的3D点云以外的各种测量信号，这些信号可用于改进3D物体检测等感知任务。单个激光脉冲可以被多个物体沿其路径部分反射，从而产生称为回波的多次测量。多回波测量可以提供有关物体轮廓和半透明表面的信息，用于更好地识别和定位物体。激光雷达还可以测量表面反射率（激光脉冲返回的强度）以及场景的环境光（物体反射的阳光）。这些信号已经在商用激光雷达设备中可用，但尚未在大多数基于激光雷达的探测模型中使用。我们提出了一个利用激光雷达提供的全谱测量信号的三维目标检测模型。首先，我们提出了一个多信号融合（MSF）模块，用于结合（1）使用2D CNN提取的反射和环境特征，以及（2）使用3D图形神经网络（GNN）提取的点云特征。第二，我们提出了一个多回波聚合（MEA）模块来组合编码在不同回波点集中的信息。与传统的单回波点云方法相比，我们提出的多信号激光雷达检测器（MSLiD）从更大范围的传感测量中提取了更丰富的背景信息，实现了更精确的三维目标检测。实验表明，通过结合多模态激光雷达，我们的方法比最先进的方法高出9.1%。

In this paper, we seek reasons for the two major failure cases in Semantic Segmentation (SS): 1) missing small objects or minor object parts, and 2) mislabeling minor parts of large objects as wrong classes. We have an interesting finding that Failure-1 is due to the underuse of detailed features and Failure-2 is due to the underuse of visual contexts. To help the model learn a better trade-off, we introduce several Self-Regulation (SR) losses for training SS neural networks. By "self", we mean that the losses are from the model per se without using any additional data or supervision. By applying the SR losses, the deep layer features are regulated by the shallow ones to preserve more details; meanwhile, shallow layer classification logits are regulated by the deep ones to capture more semantics. We conduct extensive experiments on both weakly and fully supervised SS tasks, and the results show that our approach consistently surpasses the baselines. We also validate that SR losses are easy to implement in various state-of-the-art SS models, e.g., SPGNet and OCRNet, incurring little computational overhead during training and none for testing.

在本文中，我们寻找语义切分 (SS) 中两个主要失败案例的原因：1) 缺少小对象或小对象部分，2) 将大对象的小部分错误标记为错误的类。我们有一个有趣的发现，Failure-1是由于细节功能的使用不足，Failure-2是由于视觉上下文的使用不足。为了帮助模型更好地进行权衡，我们引入了几种自调节 (SR) 损耗来训练SS神经网络。所谓“自我”，我们的意思是损失来自模型本身，而不使用任何额外的数据或监督。通过应用SR损耗，深层特征由浅层特征调节，以保留更多细节；同时，浅层分类逻辑由深层分类逻辑调节，以获取更多的语义。我们在弱监督和完全监督的SS任务上进行了大量实验，结果表明我们的方法始终优于基线。我们还验证了SR损失在各种最先进的SS模型（如SPGNet和OCRNet）中易于实现，在训练期间产生的计算开销很小，而在测试中则没有

Binary neural networks (BNNs) have received increasing attention due to their superior reductions of computation and memory. Most existing works focus on either lessening the quantization error by minimizing the gap between the full-precision weights and their binarization or designing a gradient approximation to mitigate the gradient mismatch, while leaving the "dead weights" untouched. This leads to slow convergence when training BNNs. In this paper, for the first time, we explore the influence of "dead weights" which refer to a group of weights that are barely updated during the training of BNNs, and then introduce rectified clamp unit (ReCU) to revive the "dead weights" for updating. We prove that reviving the "dead weights" by ReCU can result in a smaller quantization error. Besides, we also take into account the information entropy of the weights, and then mathematically analyze why the weight standardization can benefit BNNs. We demonstrate the inherent contradiction between minimizing the quantization error and maximizing the information entropy, and then propose an adaptive exponential scheduler to identify the range of the "dead weights". By considering the "dead weights", our method offers not only faster BNN training, but also state-of-the-art performance on CIFAR-10 and ImageNet, compared with recent methods. Code can be available at <https://github.com/z-hxu/ReCU>.

二元神经网络 (BNN) 由于其优越的计算和记忆能力而受到越来越多的关注。现有的大多数工作要么通过最小化全精度权值与其二值化之间的差距来减小量化误差，要么设计梯度近似来减小梯度失配，同时保持“死权值”不变。这导致在训练BNN时收敛缓慢。在本文中，我们首次探讨了“静重”的影响，静重指的是在BNNs训练期间几乎没有更新的一组重量，然后引入矫正钳单元 (ReCU) 来恢复“静重”进行更新。我们证明了通过ReCU恢复“静重”可以导致更小的量化误差。此外，我们还考虑了权重的信息熵，并从数学上分析了权重标准化对BNN的好处。我们证明了最小化量化误差和最大化信息熵之间的内在矛盾，然后提出了一种自适应指数调度器来识别“死权值”的范围。通过考虑“静重”，与最近的方法相比，我们的方法不仅提供了更快的BNN训练，而且在CIFAR-10和ImageNet上具有最先进的性能。代码可在<https://github.com/z-hXu/ReCU>。

This paper investigates the problem of reconstructing hyperspectral (HS) images from single RGB images captured by commercial cameras, without using paired HS and RGB images during training. To tackle this challenge, we propose a new lightweight and end-to-end learning-based framework. Specifically, on the basis of the intrinsic imaging degradation model of RGB images from HS images, we progressively spread the differences between input RGB images and re-projected RGB images from recovered HS images via effective unsupervised camera spectral response function estimation. To enable the learning without paired ground-truth HS images as supervision, we adopt the adversarial learning manner and boost it with a simple yet effective L1 gradient clipping scheme. Besides, we embed the semantic information of input RGB images to locally regularize the unsupervised learning, which is expected to promote pixels with identical semantics to have consistent spectral signatures. In addition to conducting quantitative experiments over two widely-used datasets for HS image reconstruction from synthetic RGB images, we also evaluate our method by applying recovered HS images from real RGB images to HS-based visual tracking. Extensive results show that our method significantly outperforms state-of-the-art unsupervised methods and even exceeds the latest supervised method under some settings. The source code is public available at <https://github.com/zbzhzhyy/Unsupervised-Spectral-Reconstruction>.

本文研究了在训练过程中不使用成对的高光谱 (HS) 和RGB图像的情况下，从商用相机拍摄的单个RGB图像重建高光谱 (HS) 图像的问题。为了应对这一挑战，我们提出了一个新的轻量级和基于端到端学习的框架。具体而言，基于HS图像中RGB图像的固有成像退化模型，我们通过有效的无监督相机光谱响应函数估计，逐步扩大输入RGB图像和恢复HS图像中重新投影RGB图像之间的差异。为了使学习不需要成对的地面对真实HS图像作为监督，我们采用了对抗式学习方式，并使用一种简单而有效的L1梯度剪裁方案对其进行增强。此外，我们还嵌入了输入RGB图像的语义信息，对无监督学习进行局部正则化，以促进语义相同的像素具有一致的光谱特征。除了对合成RGB图像重建HS图像的两个广泛使用的数据集进行定量实验外，我们还通过将真实RGB图像恢复的HS图像应用于基于HS的视觉跟踪来评估我们的方法。大量的实验结果表明，我们的方法明显优于最新的无监督方法，甚至在某些情况下超过了最新的有监督方法。源代码可在<https://github.com/zbzhzhyy/Unsupervised-Spectral-Reconstruction>.

As a challenging task of high-level video understanding, weakly supervised temporal action localization has been attracting increasing attention. With only video annotations, most existing methods seek to handle this task with a localization-by-classification framework, which generally adopts a selector to select snippets of high probabilities of actions or namely the foreground. Nevertheless, the existing foreground selection strategies have a major limitation of only considering the unilateral relation from foreground to actions, which cannot guarantee the foreground-action consistency. In this paper, we present a framework named FAC-Net based on the I3D backbone, on which three branches are appended, named class-wise foreground classification branch, class-agnostic attention branch and multiple instance learning branch. First, our class-wise foreground classification branch regularizes the relation between actions and foreground to maximize the foreground-background separation. Besides, the class-agnostic attention branch and multiple instance learning branch are adopted to regularize the foreground-action consistency and help to learn a meaningful foreground classifier. Within each branch, we introduce a hybrid attention mechanism, which calculates multiple attention scores for each snippet, to focus on both discriminative and less-discriminative snippets to capture the full action boundaries. Experimental results on THUMOS14 and ActivityNet1.3 demonstrate the superior performance over state-of-the-art approaches.

作为一项具有挑战性的高级视频理解任务，弱监督时间动作定位越来越受到人们的关注。由于只有视频注释，大多数现有方法试图通过分类框架进行本地化来处理此任务，该框架通常采用选择器来选择动作概率较高的片段或前景。然而，现有的前景选择策略主要局限于只考虑前景与动作之间的单向关系，不能保证前景动作的一致性。在本文中，我们提出了一个基于I3D主干的FAC-Net框架，在该框架上增加了三个分支，分别为类前景分类分支、类无关注意分支和多实例学习分支。首先，我们的类级前景分类分支将动作和前景之间的关系规则化，以最大限度地实现前景-背景分离。此外，采用类无关注意分支和多实例学习分支对前景动作一致性进行正则化，帮助学习有意义的前景分类器。在每个分支中，我们引入了一种混合注意机制，该机制为每个片段计算多个注意分数，以关注区分性和不那么区分性的片段，从而捕获完整的动作边界。THUMOS14和ActivityNet1的实验结果。3证明其性能优于最先进的方法。

Recent strategies achieved ensembling ""for free"" by fitting concurrently diverse subnetworks inside a single base network. The main idea during training is that each subnetwork learns to classify only one of the multiple inputs simultaneously provided. However, the question of how to best mix these multiple inputs has not been studied so far. In this paper, we introduce MixMo, a new generalized framework for learning multi-input multi-output deep subnetworks. Our key motivation is to replace the suboptimal summing operation hidden in previous approaches by a more appropriate mixing mechanism. For that purpose, we draw inspiration from successful mixed sample data augmentations. We show that binary mixing in features - particularly with rectangular patches from CutMix - enhances results by making subnetworks stronger and more diverse. We improve state of the art for image classification on CIFAR-100 and Tiny ImageNet datasets. Our easy to implement models notably outperform data augmented deep ensembles, without the inference and memory overheads. As we operate in features and simply better leverage the expressiveness of large networks, we open a new line of research complementary to previous works.

最近的策略通过在单个基础网络中同时安装不同的子网络来实现“免费”的集成。训练期间的主要思想是，每个子网络只学习对同时提供的多个输入中的一个进行分类。然而，到目前为止还没有研究如何最好地混合这些多个输入的问题。本文介绍了一种新的学习多输入多输出深子网络的通用框架MixMo。我们的主要动机是用一种更合适的混合机制取代以前方法中隐藏的次优求和操作。为此，我们从成功的混合样本数据扩充中获得灵感。我们展示了特征中的二进制混合——特别是来自CutMix的矩形面片——通过使子网络更强大和更多样化来增强结果。我们改进了CIFAR-100和微型ImageNet数据集的图像分类技术。我们易于实现的模型明显优于数据增强的深层集成，没有推理和内存开销。随着我们在功能上的操作和更好地利用大型网络的表现力，我们开启了一条新的研究路线，以补充以前的工作。

Recently, DETR pioneered the solution of vision tasks with transformers, it directly translates the image feature map into the object detection result. Though effective, translating the full feature map can be costly due to redundant computation on some area like the background. In this work, we encapsulate the idea of reducing spatial redundancy into a novel poll and pool (PnP) sampling module, with which we build an end-to-end PnP-DETR architecture that adaptively allocates its computation spatially to be more efficient. Concretely, the PnP module abstracts the image feature map into fine foreground object feature vectors and a small number of coarse background contextual feature vectors. The transformer models information interaction within the fine-coarse feature space and translates the features into the detection result. Moreover, the PnP-augmented model can instantly achieve various desired trade-offs between performance and computation with a single model by varying the sampled feature length, without requiring to train multiple models as existing methods. Thus it offers greater flexibility for deployment in diverse scenarios with varying computation constraint. We further validate the generalizability of the PnP module on panoptic segmentation and the recent transformer-based image recognition model ViT and show consistent efficiency gain. We believe our method makes a step for efficient visual analysis with transformers, wherein spatial redundancy is commonly observed. Code and models will be available.

最近，DETR率先提出了用变形金刚解决视觉任务的方法，它直接将图像特征映射转化为目标检测结果。虽然有效，但由于在某些区域（如背景）上存在冗余计算，转换全特征地图的成本可能会很高。在这项工作中，我们将减少空间冗余的思想封装到一个新的轮询池（poll-and-pool，PnP）采样模块中，利用该模块我们构建了一个端到端PnP-DETR体系结构，该体系结构可以自适应地在空间上分配计算以提高效率。具体而言，PnP模块将图像特征映射抽象为精细的前景对象特征向量和少量粗略的背景上下文特征向量。变压器在精细粗糙的特征空间中建模信息交互，并将特征转换为检测结果。此外，通过改变采样特征长度，PnP增强模型可以通过单个模型立即实现性能和计算之间的各种期望权衡，而无需像现有方法那样训练多个模型。因此，它为具有不同计算约束的不同场景中的部署提供了更大的灵活性。我们进一步验证了PnP模块在全景分割和最新基于变压器的图像识别模型ViT上的通用性，并显示了一致的效率增益。我们相信，我们的方法为变压器的高效视觉分析迈出了一步，其中空间冗余是常见的。代码和模型将可用。

This paper presents a neural network built upon Transformers, namely PlaneTR, to simultaneously detect and reconstruct planes from a single image. Different from previous methods, PlaneTR jointly leverages the context information and the geometric structures in a sequence-to-sequence way to holistically detect plane instances in one forward pass. Specifically, we represent the geometric structures as line segments and conduct the network with three main components: (i) context and line segments encoders, (ii) a structure-guided plane decoder, (iii) a pixel-wise plane embedding decoder. Given an image and its detected line segments, PlaneTR generates the context and line segment sequences via two specially designed encoders and then feeds them into a Transformers-based decoder to directly predict a sequence of plane instances by simultaneously considering the context and global structure cues. Finally, the pixel-wise embeddings are computed to assign each pixel to one predicted plane instance which is nearest to it in embedding space. Comprehensive experiments demonstrate that PlaneTR achieves state-of-the-art performance on the ScanNet and NYUV2 datasets.

本文提出了一种基于变换器的神经网络，即PlaneTR，用于从单个图像中同时检测和重建平面。与以前的方法不同，PlaneTR以顺序对顺序的方式联合利用上下文信息和几何结构，在一次向前传递中整体检测平面实例。具体来说，我们将几何结构表示为线段，并使用三个主要组件进行网络：(i) 上下文和线段编码器，(ii) 结构引导平面解码器，(iii) 像素平面嵌入解码器。给定一幅图像及其检测到的线段，PlaneTR通过两个专门设计的编码器生成上下文和线段序列，然后将它们输入基于转换器的解码器，通

过同时考虑上下文和全局结构线索直接预测平面实例序列。最后，计算像素嵌入，将每个像素分配到嵌入空间中最靠近它的一个预测平面实例。综合实验表明，PlaneTR在ScanNet和NYUV2数据集上实现了最先进的性能。

We present an image segmentation algorithm that is developed in an unsupervised deep learning framework. The delineation of object boundaries often fails due to the nuisance factors such as illumination changes and occlusions. Thus, we initially propose an unsupervised image decomposition algorithm to obtain an intrinsic representation that is robust with respect to undesirable bias fields based on a multiplicative image model. The obtained intrinsic image is subsequently provided to an unsupervised segmentation procedure that is developed based on a piecewise smooth model. The segmentation model is further designed to incorporate a geometric constraint imposed in the generative adversarial network framework where the discrepancy between the distribution of partitioning functions and the distribution of prior shapes is minimized. We demonstrate the effectiveness and robustness of the proposed algorithm in particular with bias fields and occlusions using simple yet illustrative synthetic examples and a benchmark dataset for image segmentation.

我们提出了一种在无监督的深度学习框架下开发的图像分割算法。由于照明变化和遮挡等干扰因素，物体边界的划定往往失败。因此，我们最初提出了一种无监督的图像分解算法，以获得一种内在的表示，该表示对基于乘法图像模型的不期望的偏置场具有鲁棒性。随后将获得的固有图像提供给基于分段平滑模型开发的无监督分割过程。分割模型进一步设计为在生成性对抗网络框架中加入几何约束，其中分割函数的分布和先验形状的分布之间的差异最小化。我们使用简单但说明性的合成示例和用于图像分割的基准数据集，证明了所提算法的有效性和鲁棒性，特别是在存在偏差场和遮挡的情况下。

In this work, we propose a Dynamic ResBlock Generative Adversarial Network (DRB-GAN) for artistic style transfer. The style code is modeled as the shared parameters for Dynamic ResBlocks connecting both the style encoding network and the style transfer network. In the style encoding network, a style class-aware attention mechanism is used to attend the style feature represent for generating the style codes. In the style transfer network, multiple Dynamic ResBlocks are designed to integrate the style code and the extracted CNN semantic feature and and then feed into the spatial window Layer-Instance Normalization (SW-LIN) decoder, which enables high-quality synthetic images with artistic style transfer. Moreover, the style collection conditional discriminator is designed to ensure our DRB-GAN model to equip with abilities for both arbitrary style transfer and collection style transfer during the training stage. No matter for arbitrary style transfer or collection style transfer, extensive experimental results strongly demonstrate that our proposed DRB-GAN beats state-of-the-art methods and exhibits its superior performance in terms of visual quality and efficiency.

在这项工作中，我们提出了一种用于艺术风格转换的动态重块生成对抗网络（DRB-GAN）。样式代码被建模为连接样式编码网络和样式传输网络的动态resblock的共享参数。在样式编码网络中，使用一种样式类感知注意机制来关注样式特征表示以生成样式代码。在风格传递网络中，设计了多个动态Resblock，将风格代码和提取的CNN语义特征进行集成，然后输入到空间窗口层实例规范化（SW-LIN）解码器中，从而实现了具有艺术风格传递的高质量合成图像。此外，设计了风格收集条件鉴别器，以确保我们的DRB-GAN模型在训练阶段具备任意风格转换和收集风格转换的能力。无论是任意风格的转换还是集合风格的转换，大量的实验结果都有力地证明了我们提出的DRB-GAN优于最先进的方法，并且在视觉质量和效率方面表现出优越的性能。

Spiking Neural Networks (SNNs) offer a promising alternative to traditional deep learning frameworks, since they provide higher computational efficiency due to event-driven information processing. SNNs distribute the analog values of pixel intensities into binary spikes over time. However, the most widely used input coding schemes, such as Poisson based rate-coding, do not leverage the additional temporal learning capability of SNNs effectively. Moreover, these SNNs suffer from high inference latency which is a major bottleneck to their deployment. To overcome this, we propose a time-based encoding scheme that utilizes the Discrete Cosine Transform (DCT) to reduce the number of timesteps required for inference. DCT decomposes an image into a weighted sum of sinusoidal basis images. At each time step, a single frequency base, taken in order and modulated by its corresponding DCT coefficient, is input to an accumulator that generates spikes upon crossing a threshold. We use the proposed scheme to learn DCT-SNN, a low-latency deep SNN with leaky-integrate-and-fire neurons, trained using surrogate gradient descent based backpropagation. We achieve top-1 accuracy of 89.94%, 68.3% and 52.43% on CIFAR-10, CIFAR-100 and TinyImageNet, respectively using VGG architectures. Notably, DCT-SNN performs inference with 2-14X reduced latency compared to other state-of-the-art SNNs, while achieving comparable accuracy to their standard deep learning counterparts. The dimension of the transform allows us to control the number of timesteps required for inference. Additionally, we can trade-off accuracy with latency in a principled manner by dropping the highest frequency components during inference. The code is publicly available\*.

尖峰神经网络 (SNN) 为传统的深度学习框架提供了一个很有前途的替代方案，因为它们由于事件驱动的信息处理提供了更高的计算效率。SNN将像素强度的模拟值随时间分布为二进制峰值。然而，最广泛使用的输入编码方案，例如基于泊松的速率编码，不能有效地利用SNN的额外时间学习能力。此外，这些SNN还存在较高的推理延迟，这是其部署的主要瓶颈。为了克服这一问题，我们提出了一种基于时间的编码方案，该方案利用离散余弦变换 (DCT) 来减少推理所需的时间步数。DCT将图像分解为正弦基图像的加权和。在每个时间步，一个单一的频率基，按顺序进行并由其相应的DCT系数调制，被输入到累加器，累加器在超过阈值时产生尖峰。我们使用所提出的方案学习DCT-SNN，这是一种低延迟的深度SNN，具有漏积分和激发神经元，使用基于代理梯度下降的反向传播进行训练。在使用VGG架构的CIFAR-10、CIFAR-100和TinyImageNet上，我们分别实现了89.94%、68.3%和52.43%的顶级精度。值得注意的是，与其他最先进的SNN相比，DCT-SNN执行推理的延迟降低了2-14倍，同时实现了与标准深度学习对应项相当的准确性。转换的维度允许我们控制推理所需的时间步数。此外，我们可以通过在推理过程中删除最高频率分量，原则性地权衡准确性和延迟。该代码可公开获取\*。

For all the ways convolutional neural nets have revolutionized computer vision in recent years, one important aspect has received surprisingly little attention: the effect of image size on the accuracy of tasks being trained for. Typically, to be efficient, the input images are resized to a relatively small spatial resolution (e.g. 224x224), and both training and inference are carried out at this resolution. The actual mechanism for this re-scaling has been an afterthought: Namely, off-the-shelf image resizers such as bilinear and bicubic are commonly used in most machine learning software frameworks. But do these resizers limit the on task performance of the trained networks? The answer is yes. Indeed, we show that the typical linear resizer can be replaced with learned resizers that can substantially improve performance. Importantly, while the classical resizers typically result in better perceptual quality of the downsampled images, our proposed learned resizers do not necessarily give better visual quality, but instead improve task performance. Our learned image resizer is jointly trained with a baseline vision model. This learned CNN-based resizer creates machine friendly visual manipulations that lead to a consistent improvement of the end task metric over the baseline model. Specifically, here we focus on the classification task with the ImageNet dataset, and experiment with four different models to learn resizers adapted to each model. Moreover, we show that the proposed resizer can also be useful for fine-tuning the classification baselines for other vision tasks. To this end, we experiment with three different baselines to develop image quality assessment (IQA) models on the AVA dataset.

尽管近年来卷积神经网络在计算机视觉领域掀起了一场革命，但有一个重要方面却很少受到关注：图像大小对训练任务准确性的影响。通常，为了提高效率，将输入图像调整为相对较小的空间分辨率（例如 224x224），并在此分辨率下执行训练和推断。这种重新缩放的实际机制是事后才想到的：即，在大多数机器学习软件框架中，通常使用现成的图像大小调整器，如双线性和双三次。但是，这些调整器是否限制了经过训练的网络的任务性能？答案是肯定的。事实上，我们表明，典型的线性大小调整器可以被学习的大小调整器所取代，从而显著提高性能。重要的是，虽然经典的大小调整器通常可以提高缩小图像的感知质量，但我们提出的学习大小调整器不一定能够提供更好的视觉质量，而是可以提高任务性能。我们学习的图像大小调整器与基线视觉模型联合训练。这个基于CNN的学识重定器创建了机器友好的视觉操作，与基线模型相比，最终任务度量得到了一致的改进。具体来说，这里我们将重点放在 ImageNet 数据集的分类任务上，并使用四种不同的模型进行实验，以了解适合每个模型的大小调整器。此外，我们还表明，建议的大小调整器也可以用于微调其他视觉任务的分类基线。为此，我们使用三种不同的基线进行实验，以在 AVA 数据集上开发图像质量评估 (IQA) 模型。

Cryo-Electron Tomography (cryo-ET) is a powerful tool for 3D cellular visualization. Due to instrumental limitations, cryo-ET images and their volumetric reconstruction suffer from extremely low signal-to-noise ratio. In this paper, we propose a novel end-to-end self-supervised learning model, the Sparsity Constrained Network (SC-Net), to restore volumetric image from single noisy data in cryo-ET. The proposed method only requires a single noisy data as training input and no ground-truth is needed in the whole training procedure. A new target function is proposed to preserve both local smoothness and detailed structure. Additionally, a novel procedure for the simulation of electron tomographic photographing is designed to help the evaluation of methods. Experiments are done on three simulated data and four real-world data. The results show that our method could produce a strong enhancement for a single very noisy cryo-ET volumetric data, which is much better than the state-of-the-art Noise2Void, and with a competitive performance comparing with Noise2Noise. Code is available at <https://github.com/icthrm/SC-Net>.

低温电子层析成像 (Cryo ET) 是三维细胞可视化的有力工具。由于仪器的限制, cryo ET图像及其体积重建的信噪比极低。在本文中, 我们提出了一种新的端到端自监督学习模型, 稀疏约束网络 (SC-Net), 用于从cryo-ET中的单个噪声数据中恢复体积图像。该方法只需要单个噪声数据作为训练输入, 整个训练过程不需要地面真值。提出了一种新的目标函数来保持局部光滑性和细节结构。此外, 还设计了一种新的电子断层摄影模拟程序, 以帮助评估方法。在三个模拟数据和四个真实数据上进行了实验。结果表明, 我们的方法可以对单个非常有噪声的cryo ET体积数据产生很强的增强, 这比最先进的Noise2Void好得多, 并且与Noise2Noise相比具有竞争力。代码可在<https://github.com/icthrm/SC-Net>。

We introduce a bottom-up model for simultaneously finding many boundary elements in an image, including contours, corners and junctions. The model explains boundary shape in each small patch using a 'generalized M-junction' comprising M angles and a freely-moving vertex. Images are analyzed using non-convex optimization to cooperatively find M+2 junction values at every location, with spatial consistency being enforced by a novel regularizer that reduces curvature while preserving corners and junctions. The resulting 'field of junctions' is simultaneously a contour detector, corner/junction detector, and boundary-aware smoothing of regional appearance. Notably, its unified analysis of contours, corners, junctions and uniform regions allows it to succeed at high noise levels, where other methods for segmentation and boundary detection fail.

我们引入了一个自底向上的模型来同时寻找图像中的许多边界元素, 包括轮廓、角点和连接点。该模型使用由M个角度和自由移动顶点组成的“广义M连接”来解释每个小面片中的边界形状。使用非凸优化对图像进行分析, 以协同查找每个位置的M+2连接值, 通过一种新的正则化器实现空间一致性, 该正则化器在保留角点和连接的同时减少曲率。由此产生的“连接场”同时是轮廓检测器、角点/连接检测器和区域外观的边界感知平滑。值得注意的是, 它对轮廓、角点、接合点和均匀区域的统一分析使它能够在高噪声水平下取得成功, 而其他分割和边界检测方法都失败了。

Generative models for 3D shapes represented by hierarchies of parts can generate realistic and diverse sets of outputs. However, existing models suffer from the key practical limitation of modelling shapes holistically and thus cannot perform conditional sampling, i.e. they are not able to generate variants on individual parts of generated shapes without modifying the rest of the shape. This is limiting for applications such as 3D CAD design that involve adjusting created shapes at multiple levels of detail. To address this, we introduce LSD-StructureNet, an augmentation to the StructureNet architecture that enables re-generation of parts situated at arbitrary positions in the hierarchies of its outputs. We achieve this by learning individual, probabilistic conditional decoders for each hierarchy depth. We evaluate LSD-StructureNet on the PartNet dataset, the largest dataset of 3D shapes represented by hierarchies of parts. Our results show that contrarily to existing methods, LSD-StructureNet can perform conditional sampling without impacting inference speed or the realism and diversity of its outputs.

由零件层次结构表示的三维形状的生成模型可以生成真实和多样的输出集。但是, 现有模型在整体建模形状方面存在关键的实际限制, 因此无法执行条件采样, 即, 如果不修改形状的其余部分, 它们无法在生成的形状的各个部分上生成变体。这限制了3D CAD设计等涉及在多个细节级别调整创建形状的应用。为了解决这个问题, 我们引入了LSD StructureNet, 它是StructureNet体系结构的一个扩展, 可以重新生成位于其输出层次结构中任意位置的部件。我们通过学习每个层次深度的单个概率条件解码器来实现这一点。我们在PartNet数据集上评估LSD StructureNet, 该数据集是由零件层次结构表示的最大3D形状数据集。我们的结果表明, 与现有方法相反, LSD StructureNet可以在不影响推理速度或其输出的真实性和多样性的情况下执行条件采样。

Most recent studies on detecting and localizing temporal anomalies have mainly employed deep neural networks to learn the normal patterns of temporal data in an unsupervised manner. Unlike them, the goal of our work is to fully utilize instance-level (or weak) anomaly labels, which only indicate whether any anomalous events occurred or not in each instance of temporal data. In this paper, we present WETAS, a novel framework that effectively identifies anomalous temporal segments (i.e., consecutive time points) in an input instance. WETAS learns discriminative features from the instance-level labels so that it infers the sequential order of normal and anomalous segments within each instance, which can be used as a rough segmentation mask. Based on the dynamic time warping (DTW) alignment between the input instance and its segmentation mask, WETAS obtains the result of temporal segmentation, and simultaneously, it further enhances itself by using the mask as additional supervision. Our experiments show that WETAS considerably outperforms other baselines in terms of the localization of temporal anomalies, and also it provides more informative results than point-level detection methods.

最近关于检测和定位时间异常的研究主要是利用深度神经网络以无监督的方式学习时间数据的正常模式。与它们不同的是，我们工作的目标是充分利用实例级（或弱）异常标签，它只指示在每个时态数据实例中是否发生了任何异常事件。在本文中，我们提出了一种新的框架WETAS，它可以有效地识别输入实例中的异常时间段（即连续时间点）。WETAS从实例级别的标签中学习鉴别特征，从而推断每个实例中正常和异常段的顺序，这可以用作粗分割掩码。基于输入实例与其分割掩码之间的动态时间扭曲（DTW）对齐，WETAS获得时间分割的结果，同时，通过使用掩码作为附加监控，它进一步增强了自身。我们的实验表明，在时间异常的定位方面，WETAS大大优于其他基线，并且与点级检测方法相比，它提供了更多信息。

While self-training has advanced semi-supervised semantic segmentation, it severely suffers from the long-tailed class distribution on real-world semantic segmentation datasets that make the pseudo-labeled data bias toward majority classes. In this paper, we present a simple and yet effective Distribution Alignment and Random Sampling (DARS) method to produce unbiased pseudo labels that match the true class distribution estimated from the labeled data. Besides, we also contribute a progressive data augmentation and labeling strategy to facilitate model training with pseudo-labeled data. Experiments on both Cityscapes and PASCAL VOC 2012 datasets demonstrate the effectiveness of our approach. Albeit simple, our method performs favorably in comparison with state-of-the-art approaches. Code will be available at <https://github.com/CVMI-Lab/DARS>.

虽然自训练促进了半监督语义分割，但它严重受到现实世界语义分割数据集上的长尾类分布的影响，这使得伪标记数据偏向大多数类。在本文中，我们提出了一种简单而有效的分布对齐和随机抽样（DARS）方法来生成无偏伪标签，该标签与从标签数据估计的真实类分布相匹配。此外，我们还提供了一种渐进的数据扩充和标记策略，以便于使用伪标记数据进行模型训练。在城市景观和PASCAL VOC 2012数据集上的实验证明了我们方法的有效性。虽然简单，但与最先进的方法相比，我们的方法表现良好。代码将在<https://github.com/CVMI-Lab/DARS>。

The past year has witnessed the rapid development of applying the Transformer module to vision problems. While some researchers have demonstrated that Transformer-based models enjoy a favorable ability of fitting data, there are still growing number of evidences showing that these models suffer over-fitting especially when the training data is limited. This paper offers an empirical study by performing step-by-step operations to gradually transit a Transformer-based model to a convolution-based model. The results we obtain during the transition process deliver useful messages for improving visual recognition. Based on these observations, we propose a new architecture named Visformer, which is abbreviated from the 'Vision-friendly Transformer'. With the same computational complexity, Visformer outperforms both the Transformer-based and convolution-based models in terms of ImageNet classification accuracy, and the advantage becomes more significant when the model complexity is lower or the training set is smaller. The code is available at <https://github.com/danczs/Visformer>.

在过去的一年里，变压器模块在视觉问题上的应用得到了快速发展。虽然一些研究人员已经证明基于变压器的模型具有良好的数据拟合能力，但仍有越来越多的证据表明，这些模型存在过度拟合的问题，尤其是在训练数据有限的情况下。本文提供了一个实证研究，通过逐步操作，逐步将基于变压器的模型转换为基于卷积的模型。我们在转换过程中获得的结果为改进视觉识别提供了有用的信息。基于这些观察，我们提出了一种新的架构，名为Visformer，它是“视觉友好型转换器”的缩写。在计算复杂度相同的情况下，Visformer在ImageNet分类精度方面优于基于变压器和基于卷积的模型，并且当模型复杂度较低或训练集较小时，优势更加显著。该守则可于<https://github.com/danczs/Visformer>.

Real-world machine learning systems need to analyze novel testing data that differs from the training data. In K-way classification, this is crisply formulated as open-set recognition, core to which is the ability to discriminate open-set data outside the K closed-set classes. Two conceptually elegant ideas for open-set discrimination are: 1) discriminatively learning an open-vs-closed binary discriminator by exploiting some outlier data as the open-set, and 2) unsupervised learning the closed-set data distribution with a GAN and using its discriminator as the open-set likelihood function. However, the former generalizes poorly to diverse open test data due to overfitting to the training outliers, which unlikely exhaustively span the open-world. The latter does not work well, presumably due to the unstable training of GANs. Motivated by the above, we propose OpenGAN, which addresses the limitation of each approach by combining them with several technical insights. First, we show that a carefully selected GAN-discriminator on some real outlier data already achieves the state-of-the-art. Second, we augment the available set of real open training examples with adversarially synthesized "fake" data. Third and most importantly, we build the discriminator over the features computed by the closed-world K-way networks. Extensive experiments show that OpenGAN significantly outperforms prior open-set methods.

现实世界的机器学习系统需要分析与训练数据不同的新测试数据。在K-way分类中，这被清晰地表述为开集识别，其核心是能够区分K闭集类之外的开集数据。关于开集判别，有两个概念上很好的想法：1) 通过利用一些离群数据作为开集，对开集与闭二元判别器进行判别学习；2) 使用GAN，对闭集数据分布进行无监督学习，并使用其判别器作为开集似然函数。然而，由于过度拟合训练异常值，前者不能很好地推广到不同的开放测试数据，这不可能完全跨越开放世界。后者效果不好，可能是由于GANs训练不稳定所致。基于上述动机，我们提出了OpenGAN，它通过将每种方法与一些技术见解结合起来，解决了每种方法的局限性。首先，我们证明了在一些真实的离群数据上精心选择的GAN鉴别器已经达到了最先进的水平。第二，我们用敌对合成的“假”数据来扩充可用的真实开放训练示例集。第三，也是最重要的一点，我们在封闭世界K-way网络计算的特征上构建鉴别器。大量实验表明，OpenGAN的性能明显优于以前的开放集方法。

Recent work has shown impressive results on data-driven defocus deblurring using the two-image views available on modern dual-pixel (DP) sensors. One significant challenge in this line of research is access to DP data. Despite many cameras having DP sensors, only a limited number provide access to the low-level DP sensor images. In addition, capturing training data for defocus deblurring involves a time-consuming and tedious setup requiring the camera's aperture to be adjusted. Some cameras with DP sensors (e.g., smartphones) do not have adjustable apertures, further limiting the ability to produce the necessary training data. We address the data capture bottleneck by proposing a procedure to generate realistic DP data synthetically. Our synthesis approach mimics the optical image formation found on DP sensors and can be applied to virtual scenes rendered with standard computer software. Leveraging these realistic synthetic DP images, we introduce a recurrent convolutional network (RCN) architecture that improves deblurring results and is suitable for use with single-frame and multi-frame data (e.g., video) captured by DP sensors. Finally, we show that our synthetic DP data is useful for training DNN models targeting video deblurring applications where access to DP data remains challenging.

最近的工作表明，使用现代双像素（DP）传感器上的两个图像视图，数据驱动的散焦去模糊效果令人印象深刻。这一研究领域的一个重大挑战是获取DP数据。尽管许多摄像机都有DP传感器，但只有有限数量的摄像机能够访问低级别的DP传感器图像。此外，捕获用于散焦去模糊的训练数据涉及耗时且繁琐的设置，需要调整相机的光圈。一些带有DP传感器的摄像头（如智能手机）没有可调光圈，进一步限制了生成必要训练数据的能力。我们通过提出一个综合生成真实DP数据的过程来解决数据捕获瓶颈。我们的合成方法模拟了DP传感器上的光学图像形成，可以应用于使用标准计算机软件渲染的虚拟场景。利用这些真实的合成DP图像，我们引入了一种循环卷积网络（RCN）体系结构，该体系结构改进了去模糊结果，并适用于DP传感器捕获的单帧和多帧数据（例如视频）。最后，我们展示了我们的合成DP数据对于针对视频去模糊应用的DNN模型的训练是有用的，在这些应用中，DP数据的访问仍然具有挑战性。

This paper studies the context aggregation problem in semantic image segmentation. The existing researches focus on improving the pixel representations by aggregating the contextual information within individual images. Though impressive, these methods neglect the significance of the representations of the pixels of the corresponding class beyond the input image. To address this, this paper proposes to mine the contextual information beyond individual images to further augment the pixel representations. We first set up a feature memory module, which is updated dynamically during training, to store the dataset-level representations of various categories. Then, we learn class probability distribution of each pixel representation under the supervision of the ground-truth segmentation. At last, the representation of each pixel is augmented by aggregating the dataset-level representations based on the corresponding class probability distribution. Furthermore, by utilizing the stored dataset-level representations, we also propose a representation consistent learning strategy to make the classification head better address intra-class compactness and inter-class dispersion. The proposed method could be effortlessly incorporated into existing segmentation frameworks (e.g., FCN, PSPNet, OCRNet and DeepLabV3) and brings consistent performance improvements. Mining contextual information beyond image allows us to report state-of-the-art performance on various benchmarks: ADE20K, LIP, Cityscapes and COCO-Stuff.

本文研究了语义图像分割中的上下文聚合问题。现有的研究集中在通过聚合单个图像中的上下文信息来改进像素表示。尽管这些方法令人印象深刻，但它们忽略了输入图像之外对应类别像素表示的重要性。为了解决这个问题，本文提出挖掘单个图像之外的上下文信息，以进一步增强像素表示。我们首先建立一个特征存储模块，在训练过程中动态更新，以存储各种类别的数据集级表示。然后，在地面真值分割的监督下，学习每个像素表示的类概率分布。最后，通过根据相应的类概率分布聚合数据集级表示来增强每个像素的表示。此外，通过利用存储的数据集级表示，我们还提出了一种表示一致性学习策略，使

分类头能够更好地解决类内紧凑性和类间分散性问题。所提出的方法可以轻松地融入现有的分割框架（如FCN、PSPNet、OCRNet和DeepLabV3），并带来一致的性能改进。通过挖掘图像之外的上下文信息，我们可以报告各种基准的最新性能：ADE20K、LIP、Cityscapes和COCO等。

In dynamic neural networks that adapt computations to different inputs, gating-based methods have demonstrated notable generality and applicability in trading-off the model complexity and accuracy. However, existing works only explore the redundancy from a single point of the network, limiting the performance. In this paper, we propose dual gating, a new dynamic computing method, to reduce the model complexity at run-time. For each convolutional block, dual gating identifies the informative features along two separate dimensions, spatial and channel. Specifically, the spatial gating module estimates which areas are essential, and the channel gating module predicts the salient channels that contribute more to the results. Then the computation of both unimportant regions and irrelevant channels can be skipped dynamically during inference. Extensive experiments on a variety of datasets demonstrate that our method can achieve higher accuracy under similar computing budgets compared with other dynamic execution methods. In particular, dynamic dual gating can provide 59.7% saving in computing of ResNet50 with 76.41% top-1 accuracy on ImageNet, which has advanced the state-of-the-art.

在使计算适应不同输入的动态神经网络中，基于选通的方法在权衡模型复杂性和准确性方面表现出显著的通用性和适用性。然而，现有的工作仅从网络的单个点探索冗余，限制了性能。在本文中，我们提出了双选通，一种新的动态计算方法，以降低运行时的模型复杂性。对于每个卷积块，双选通沿空间和通道两个独立维度识别信息特征。具体而言，空间选通模块估计哪些区域是必要的，通道选通模块预测对结果贡献更大的显著通道。然后在推理过程中动态跳过不重要区域和无关通道的计算。在各种数据集上进行的大量实验表明，与其他动态执行方法相比，该方法在计算量相近的情况下可以获得更高的精度。特别是，动态双选通可以为ResNet50的计算节省59.7%，在ImageNet上的最高精度为76.41%，这提高了最先进的技术水平。

Transformers have been recently adapted for large scale image classification, achieving high scores shaking up the long supremacy of convolutional neural networks. However the optimization of vision transformers has been little studied so far. In this work, we build and optimize deeper transformer networks for image classification. In particular, we investigate the interplay of architecture and optimization of such dedicated transformers. We make two architecture changes that significantly improve the accuracy of deep transformers. This leads us to produce models whose performance does not saturate early with more depth, for instance we obtain 86.5% top-1 accuracy on Imagenet when training with no external data, we thus attain the current state of the art with less floating-point operations and parameters. Our best model establishes the new state of the art on Imagenet with Reassessed labels and Imagenet-V2 / match frequency, in the setting with no additional training data. We share our code and models

变压器最近被用于大规模图像分类，取得了高分，动摇了卷积神经网络的长期优势。然而，迄今为止对视觉变换器的优化研究很少。在这项工作中，我们建立和优化更深层次的变压器网络用于图像分类。特别是，我们研究了这种专用变压器的架构和优化的相互作用。我们进行了两次架构更改，显著提高了deep transformers的精度。这导致我们生产的模型的性能不会随着深度的增加而提前饱和，例如，当在没有外部数据的情况下进行训练时，我们在Imagenet上获得86.5%的top-1精度，因此我们可以用更少的浮点运算和参数达到目前的水平。我们的最佳模型通过重新评估标签和Imagenet-V2/匹配频率，在没有额外训练数据的情况下，在Imagenet上建立了最新的技术状态。我们共享我们的代码和模型

weakly supervised salient object detection (wsod) targets to train a CNNs-based saliency network using only low-cost annotations. Existing wsod methods take various techniques to pursue single "high-quality" pseudo label from low-cost annotations and then develop their saliency networks. Though these methods have achieved good performance, the generated single label is inevitably affected by adopted refinement algorithms and shows prejudiced characteristics which further influence the saliency networks. In this work, we introduce a new multiple-pseudo label framework to integrate more comprehensive and accurate saliency cues from multiple labels, avoiding the aforementioned problem. Specifically, we propose a multi-filer directive network (MFNet) including a saliency network as well as multiple directive filters. The directive filter (DF) is designed to extract and filter more accurate saliency cues from the noisy pseudo labels. The multiple accurate cues from multiple DFs are then simultaneously propagated to the saliency network with a multi-guidance loss. Extensive experiments on five datasets over four metrics demonstrate that our method outperforms all the existing congeneric methods. Moreover, it is also worth noting that our framework is flexible enough to apply to existing methods and improve their performance.

弱监督显著性目标检测 (WSOD) 目标仅使用低成本注释来训练基于CNNs的显著性网络。现有的WSOD方法采用各种技术从低成本注释中寻找单一的“高质量”伪标签，然后开发其显著性网络。虽然这些方法都取得了良好的性能，但是生成的单个标签不可避免地受到所采用的细化算法的影响，并且显示出偏见的特征，这进一步影响了显著性网络。在这项工作中，我们引入了一个新的多伪标签框架来整合来自多个标签的更全面和准确的显著性线索，避免了上述问题。具体来说，我们提出了一种多文件器指令网络 (MFNet)，包括一个显著性网络和多个指令过滤器。指令滤波器 (DF) 设计用于从噪声伪标签中提取和过滤更精确的显著性线索。然后，来自多个DFs的多个精确线索同时传播到显著性网络，并伴有多个制导损失。在五个数据集上对四个指标进行的大量实验表明，我们的方法优于所有现有的同类方法。此外，还值得注意的是，我们的框架足够灵活，可以应用于现有方法并提高其性能。

Traffic accident anticipation aims to accurately and promptly predict the occurrence of a future accident from dashcam videos, which is vital for a safety-guaranteed self-driving system. To encourage an early and accurate decision, existing approaches typically focus on capturing the cues of spatial and temporal context before a future accident occurs. However, their decision-making lacks visual explanation and ignores the dynamic interaction with the environment. In this paper, we propose Deep Reinforced accident anticipation with visual Explanation, named DRIVE. The method simulates both the bottom-up and top-down visual attention mechanism in a dashcam observation environment so that the decision from the proposed stochastic multi-task agent can be visually explained by attentive regions. Moreover, the proposed dense anticipation reward and sparse fixation reward are effective in training the DRIVE model with our improved reinforcement learning algorithm. Experimental results show that the DRIVE model achieves state-of-the-art performance on multiple real-world traffic accident datasets. Code and pre-trained models are available at <https://www.rit.edu/actionlab/drive>.

交通事故预测旨在通过dashcam视频准确、及时地预测未来事故的发生，这对于安全保证的自动驾驶系统至关重要。为了鼓励早期和准确的决策，现有的方法通常侧重于在未来事故发生之前捕捉空间和时间背景的线索。然而，他们的决策缺乏视觉解释，忽视了与环境的动态互动。在这篇文章中，我们提出了一种深度强化的事故预测和视觉解释，名为DRIVE。该方法在dashcam观察环境中模拟了自下而上和自上而下的视觉注意机制，使得所提出的随机多任务智能体的决策可以用注意区域直观地解释。此外，本文提出的密集预期奖励和稀疏固定奖励在训练驱动模型时是有效的。实验结果表明，该驱动模型在多个真实交通事故数据集上达到了最先进的性能。代码和预先培训的模型可在<https://www.rit.edu/actionlab/drive>。

Understanding videos to localize moments with natural language often requires large expensive annotated video regions paired with language queries. To eliminate the annotation costs, we make a first attempt to train a natural language video localization model in zero-shot manner. Inspired by unsupervised image captioning setup, we merely require random text corpora, unlabeled video collections, and an off-the-shelf object detector to train a model. With the unrelated and unpaired data, we propose to generate pseudo-supervision of candidate temporal regions and corresponding query sentences, and develop a simple NLVL model to train with the pseudo-supervision. Our empirical validations show that the proposed pseudo-supervised method outperforms several baseline approaches and a number of methods using stronger supervision on Charades-STA and ActivityNet-Captions.

理解视频以使用自然语言定位瞬间通常需要大量昂贵的带注释的视频区域和语言查询。为了消除注释成本，我们首次尝试以零镜头方式训练自然语言视频定位模型。受无监督图像字幕设置的启发，我们只需要随机文本语料库、未标记的视频集合和现成的对象检测器来训练模型。对于不相关和不成对的数据，我们建议生成候选时态区域和相应查询语句的伪监督，并开发一个简单的NLVL模型来进行伪监督训练。我们的实验验证表明，所提出的伪监督方法优于几种基线方法和一些对Charades STA和ActivityNet字幕使用更强监督的方法。

An optimal transportation map finds the most economical way to transport one probability measure to the other. It has been applied in a broad range of applications in vision, deep learning and medical images. By Brenier theory, computing the optimal transport map is equivalent to solving a Monge-Ampere equation. Due to the highly non-linear nature, the computation of optimal transportation maps in large scale is very challenging. This work proposes a simple but powerful method, the FFT-OT algorithm, to tackle this difficulty based on three key ideas. First, solving Monge-Ampere equation is converted to a fixed point problem; Second, the obliqueness property of optimal transportation maps are reformulated as Neumann boundary conditions on rectangular domains; Third, FFT is applied in each iteration to solve a Poisson equation in order to improve the efficiency. Experiments on surfaces captured from 3D scanning and reconstructed from medical imaging are conducted, and compared with other existing methods. Our experimental results show that the proposed FFT-OT algorithm is simple, general and scalable with high efficiency and accuracy.

最优运输图可以找到将一个概率度量值运输到另一个概率度量值的最经济的方法。它在视觉、深度学习和医学图像等领域有着广泛的应用。根据Brenier理论，计算最优运输图相当于求解Monge-Ampere方程。由于其高度非线性的特点，大比例尺最优交通图的计算具有很大的挑战性。这项工作提出了一种简单但功能强大的方法，FFT-OT算法，基于三个关键思想来解决这一难题。首先，将求解Monge-Ampere方程转化为不动点问题；其次，将最优运输映射的倾斜性转化为矩形区域上的Neumann边界条件；第三，在每次迭代中应用FFT来求解泊松方程，以提高效率。对三维扫描获取的表面和医学成像重建的表面进行了实验，并与现有的其他方法进行了比较。实验结果表明，所提出的FFT-OT算法简单、通用、可扩展、效率高、精度高。

Correspondence pruning aims to correctly remove false matches (outliers) from an initial set of putative correspondences. The selection is challenging since putative matches are typically extremely unbalanced, largely dominated by outliers, and the random distribution of such outliers further complicates the learning process for learning-based methods. To address this issue, we propose to progressively prune the correspondences via a local-to-global consensus learning procedure. We introduce a "pruning" block that lets us identify reliable candidates among the initial matches according to consensus scores estimated using local-to-global dynamic graphs. We then achieve progressive pruning by stacking multiple pruning blocks sequentially. Our method outperforms state-of-the-arts on robust line fitting, camera pose estimation and retrieval-based image localization benchmarks by significant margins and shows promising generalization ability to different datasets and detector/descriptor combinations.

对应剪枝的目的是从一组假定的对应关系中正确地删除错误匹配（异常值）。这种选择是具有挑战性的，因为假定的匹配通常是极不平衡的，主要由离群值控制，并且这种离群值的随机分布进一步使基于学习的方法的学习过程复杂化。为了解决这个问题，我们建议通过一个从本地到全球的共识学习程序逐步删减通信。我们引入了一个“剪枝”块，使我们能够根据使用局部到全局动态图估计的一致性分数，在初始匹配中识别可靠的候选对象。然后，我们通过顺序堆叠多个修剪块来实现渐进修剪。我们的方法在鲁棒的直线拟合、摄像机姿态估计和基于检索的图像定位基准方面优于现有的方法，并且对不同的数据集和检测器/描述符组合显示出良好的泛化能力。

We present Multiscale Vision Transformers (MViT) for video and image recognition, by connecting the seminal idea of multiscale feature hierarchies with transformer models. Multiscale Transformers have several channel-resolution scale stages. Starting from the input resolution and a small channel dimension, the stages hierarchically expand the channel capacity while reducing the spatial resolution. This creates a multiscale pyramid of features with early layers operating at high spatial resolution to model simple low-level visual information, and deeper layers at spatially coarse, but complex, high-dimensional features. We evaluate this fundamental architectural prior for modeling the dense nature of visual signals for a variety of video recognition tasks where it outperforms concurrent vision transformers that rely on large scale external pre-training and are 5-10 more costly in computation and parameters. We further remove the temporal dimension and apply our model for image classification where it outperforms prior work on vision transformers. Code is available at:  
<https://github.com/facebookresearch/SlowFast>.

我们提出了用于视频和图像识别的多尺度视觉变换器 (MViT)，它将多尺度特征层次的基本思想与变换器模型相结合。多尺度变压器具有多个通道分辨率缩放级。从输入分辨率和较小的通道维度开始，阶段分层扩展通道容量，同时降低空间分辨率。这将创建一个多尺度特征金字塔，早期层以高空间分辨率运行，以模拟简单的低级视觉信息，而深层层以空间粗糙但复杂的高维特征运行。我们评估了这一基本的体系结构先验知识，用于为各种视频识别任务建模视觉信号的密集性质，其性能优于依赖大规模外部预训练的并发视觉转换器，且计算和参数成本高5-10%。我们进一步去除了时间维度，并将我们的模型应用于图像分类，其性能优于先前关于视觉变压器的工作。代码可从以下网址获取：<https://github.com/facebookresearch/SlowFast>。

While learning-based multi-view stereo (MVS) methods have recently shown successful performances in quality and efficiency, limited MVS data hampers generalization to unseen environments. A simple solution is to generate various large-scale MVS datasets, but generating dense ground truth for 3D structure requires a huge amount of time and resources. On the other hand, if the reliance on dense ground truth is relaxed, MVS systems will generalize more smoothly to new environments. To this end, we first introduce a novel semi-supervised multi-view stereo framework called a Sparse Ground truth-based MVS Network (SGT-MVSNet) that can reliably reconstruct the 3D structures even with a few ground truth 3D points. Our strategy is to divide the accurate and erroneous regions and individually conquer them based on our observation that a probability map can separate these regions. We propose a self-supervision loss called the 3D Point Consistency Loss to enhance the 3D reconstruction performance, which forces the 3D points back-projected from the corresponding pixels by the predicted depth values to meet at the same 3D coordinates. Finally, we propagate these improved depth predictions toward edges and occlusions by the Coarse-to-fine Reliable Depth Propagation module. We generate the spare ground truth of the DTU dataset for evaluation and extensive experiments verify that our SGT-MVSNet outperforms the state-of-the-art MVS methods on the sparse ground truth setting. Moreover, our method shows comparable reconstruction results to the supervised MVS methods though we only used tens and hundreds of ground truth 3D points.

虽然基于学习的多视图立体 (MVS) 方法最近在质量和效率方面取得了成功，但有限的MVS数据妨碍了对未知环境的推广。一个简单的解决方案是生成各种大规模MVS数据集，但为3D结构生成密集的地面真相需要大量的时间和资源。另一方面，如果放松对密集地面真相的依赖，MVS系统将更顺利地推广到新环境。为此，我们首先介绍了一种新的半监督多视角立体框架，称为基于稀疏地面真值的MVS网络 (SGT MVSNet)，它可以可靠地重建3D结构，即使只有少量地面真值3D点。我们的策略是根据概率图可以分隔这些区域的观察结果，将准确区域和错误区域分开，并单独征服它们。我们提出了一种称为3D点一致性损失的自我监督损失来提高3D重建性能，该损失迫使从相应像素反向投影的3D点通过预测的深度值在相同的3D坐标处相遇。最后，我们通过从粗到细的可靠深度传播模块向边缘和遮挡传播这些改进的深度预测。我们生成DTU数据集的备用地面真实值进行评估，大量实验证明了我们的SGT MVSNet在稀疏地面真实值设置上优于最先进的MVS方法。此外，尽管我们只使用了数十个和数百个地面真实三维点，但我们的方法显示了与监督MVS方法相当的重建结果。

This paper examines the problem of illumination spectra estimation in multispectral images. We cast the problem into a constrained matrix factorization problem and present a method for both single-global and multiple illumination estimation in which a deep unrolling network is constructed from the alternating direction method of multipliers (ADMM) optimization for solving the matrix factorization problem. To alleviate the lack of multispectral training data, we build a large multispectral reflectance image dataset for generating synthesized data and use them for training and evaluating our model. The results of simulations and real experiments demonstrate that the proposed method is able to outperform state-of-the-art spectral illumination estimation methods, and that it generalizes well to a wide variety of scenes and spectra.

本文研究了多光谱图像中的光谱估计问题。我们将该问题转化为一个约束矩阵分解问题，并提出了一种单全局和多全局光谱估计方法，其中，从求解矩阵分解问题的交替方向乘子法 (ADMM) 优化构造了一个深度展开网络。为了缓解多光谱训练数据的不足，我们构建了一个大型多光谱反射图像数据集，用于生成合成数据，并使用它们来训练和评估我们的模型。仿真和实际实验结果表明，该方法优于现有的光谱光谱估计方法，并能很好地推广到各种场景和光谱。

The recently-developed DETR approach applies the transformer encoder and decoder architecture to object detection and achieves promising performance. In this paper, we handle the critical issue, slow training convergence, and present a conditional cross-attention mechanism for fast DETR training. Our approach is motivated by that the cross-attention in DETR relies highly on the content embeddings for localizing the four extremities and predicting the box, which increases the need for high-quality content embeddings and thus the training difficulty. Our approach, named conditional DETR, learns a conditional spatial query from the decoder embedding for decoder multi-head cross-attention. The benefit is that through the conditional spatial query, each cross-attention head is able to attend to a band containing a distinct region, e.g., one object extremity or a region inside the object box. This narrows down the spatial range for localizing the distinct regions for object classification and box regression, thus relaxing the dependence on the content embeddings and easing the training. Empirical results show that conditional DETR converges 6.7x faster for the backbones R50 and R101 and 10x faster for stronger backbones DC5-R50 and DC5-R101. Code is available at <https://github.com/Atten4Vis/ConditionalDETR>.

最近开发的DETR方法将transformer编码器和解码器架构应用于目标检测，并取得了良好的性能。在本文中，我们处理的关键问题，缓慢的训练收敛，并提出了一种条件交叉注意机制的快速DETR训练。我们的方法的动机是，DETR中的交叉注意高度依赖于内容嵌入来定位四肢和预测盒子，这增加了对高质量内容嵌入的需求，从而增加了训练难度。我们的方法称为条件DETR，它从解码器嵌入中学习一个条件空间查询，用于解码器多头交叉注意。这样做的好处是，通过条件空间查询，每个交叉注意头部能够注意到包含不同区域的波段，例如，一个物体末端或物体盒内的区域。这缩小了用于对象分类和盒回归的不同区域定位的空间范围，从而放松了对内容嵌入的依赖并简化了训练。实证结果表明，对于主干R50和R101，条件DETR收敛速度快6.7倍，对于更强的主干DC5-R50和DC5-R101，条件DETR收敛速度快10倍。代码可在<https://github.com/Atten4Vis/ConditionalDETR>。

In prevalent knowledge distillation, logits in most image recognition models are computed by global average pooling, then used to learn to encode the high-level and task-relevant knowledge. In this work, we solve the limitation of this global logit transfer in this distillation context. We point out that it prevents the transfer of informative spatial information, which provides localized knowledge as well as rich relational information across contexts of an input scene. To exploit the rich spatial information, we propose a simple yet effective logit distillation approach. We add a local spatial pooling layer branch to the penultimate layer, thereby our method extends the standard logit distillation and enables learning of both finely-localized knowledge and holistic representation. Our proposed method shows favorable accuracy improvement against the state-of-the-art methods on several image classification datasets. We show that our distilled students trained on the image classification task can be successfully leveraged for object detection and semantic segmentation tasks; this result demonstrates our method's high transferability.

在流行的知识提取中，大多数图像识别模型中的logit都是通过全局平均池计算出来的，然后用来学习对高层和任务相关的知识进行编码。在这项工作中，我们解决了这种全局logit转移在这种蒸馏上下文中的局限性。我们指出，它阻止了信息性空间信息的传递，而信息性空间信息在输入场景的上下文中提供了本地化的知识以及丰富的关系信息。为了利用丰富的空间信息，我们提出了一种简单而有效的logit蒸馏方法。我们在倒数第二层添加了一个局部空间池层分支，因此我们的方法扩展了标准的logit蒸馏，并支持精细局部知识和整体表示的学习。我们提出的方法在多个图像分类数据集上与最新的方法相比，显示出良好的精度改进。我们证明，我们在图像分类任务上训练的学生可以成功地用于目标检测和语义分割任务；这一结果表明我们的方法具有很高的可移植性。

Multi-view projection methods have demonstrated their ability to reach state-of-the-art performance on 3D shape recognition. Those methods learn different ways to aggregate information from multiple views. However, the camera view-points for those views tend to be heuristically set and fixed for all shapes. To circumvent the lack of dynamism of current multi-view methods, we propose to learn those view-points. In particular, we introduce the Multi-View Transformation Network (MVTN) that regresses optimal view-points for 3D shape recognition, building upon advances in differentiable rendering. As a result, MVTN can be trained end-to-end along with any multi-view network for 3D shape classification. We integrate MVTN in a novel adaptive multi-view pipeline that can render either 3D meshes or point clouds. MVTN exhibits clear performance gains in the tasks of 3D shape classification and 3D shape retrieval without the need for extra training supervision. In these tasks, MVTN achieves state-of-the-art performance on ModelNet40, ShapeNet Core55, and the most recent and realistic ScanObjectNN dataset (up to 6 % improvement). Interestingly, we also show that MVTN can provide network robustness against rotation and occlusion in the 3D domain.

多视图投影方法已经证明了其在三维形状识别方面达到最先进水平的能力。这些方法学习从多个视图聚合信息的不同方法。但是，对于所有形状，这些视图的摄影机视点往往是启发式设置和固定的。为了避免当前多视角方法缺乏动态性，我们建议学习这些视角。特别是，我们介绍了多视图转换网络（MVTN），该网络基于可微绘制的进展，回归用于三维形状识别的最佳视点。因此，MVTN可以与任何多视图网络一起进行端到端的训练，用于三维形状分类。我们将MVTN集成到一个新的自适应多视图管道中，该管道可以渲染三维网格或点云。MVTN在3D形状分类和3D形状检索任务中表现出明显的性能提升，无需额外的培训监督。在这些任务中，MVTN在ModelNet40、ShapeNet Core55和最新和最真实的ScanObjectNN数据集上实现了最先进的性能（提高了6%）。有趣的是，我们还表明，MVTN可以在3D域中提供针对旋转和遮挡的网络鲁棒性。

We introduce GNeRF, a framework to marry Generative Adversarial Networks (GAN) with Neural Radiance Field (NeRF) reconstruction for the complex scenarios with unknown and even randomly initialized camera poses. Recent NeRF-based advances have gained popularity for remarkable realistic novel view synthesis. However, most of them heavily rely on accurate camera poses estimation, while few recent methods can only optimize the unknown camera poses in roughly forward-facing scenes with relatively short camera trajectories and require rough camera poses initialization. Differently, our GNeRF only utilizes randomly initialized poses for complex outside-in scenarios. We propose a novel two-phases end-to-end framework. The first phase takes the use of GANs into the new realm for optimizing coarse camera poses and radiance fields jointly, while the second phase refines them with additional photometric loss. We overcome local minima using a hybrid and iterative optimization scheme. Extensive experiments on a variety of synthetic and natural scenes demonstrate the effectiveness of GNeRF. More impressively, our approach outperforms the baselines favorably in those scenes with repeated patterns or even low textures that are regarded as extremely challenging before.

我们介绍了GNeRF，这是一个将生成对抗网络（GAN）与神经辐射场（NeRF）重建结合起来的框架，用于未知甚至随机初始化相机姿态的复杂场景。基于NeRF的最新进展因其出色的现实主义小说视图合成而广受欢迎。然而，大多数方法严重依赖于精确的相机姿态估计，而最近的一些方法只能在相机轨迹相对较短的大致前向场景中优化未知的相机姿态，并且需要粗略的相机姿态初始化。不同的是，我们的GNeRF只在复杂的内外场景中使用随机初始化的姿态。我们提出了一种新颖的两阶段端到端框架。第一阶段将GANs的使用带入了一个新的领域，用于联合优化粗略的相机姿态和辐射场，而第二阶段通过额外的光度损失对其进行细化。我们使用混合迭代优化方案来克服局部极小。在各种合成和自然场景上的大量实验证明了GNeRF的有效性。更令人印象深刻的是，我们的方法在那些具有重复图案甚至低纹理的场景中优于基线，这些场景以前被认为是非常具有挑战性的。

Localizing objects and estimating their extent in 3D is an important step towards high-level 3D scene understanding, which has many applications in Augmented Reality and Robotics. We present ODAM, a system for 3D Object Detection, Association, and Mapping using posed RGB videos. The proposed system relies on a deep-learning-based front-end to detect 3D objects from a given RGB frame and associate them to a global object-based map using a graph neural network (GNN). Based on these frame-to-model associations, our back-end optimizes object bounding volumes, represented as super-quadratics, under multi-view geometry constraints and the object scale prior. We validate the proposed system on ScanNet where we show a significant improvement over existing RGB-only methods.

在3D中定位对象并估计其范围是实现高水平3D场景理解的重要步骤，在增强现实和机器人技术中有广泛的应用。我们介绍了ODAM，一个使用姿势RGB视频进行3D对象检测、关联和映射的系统。该系统依赖于基于深度学习的前端，从给定的RGB帧中检测3D对象，并使用图形神经网络（GNN）将其与基于对象的全局地图相关联。基于这些帧到模型的关联，我们的后端在多视图几何约束和对象缩放之前优化对象边界体积（表示为超级二次曲面）。我们在ScanNet上验证了所提出的系统，与现有的仅使用RGB的方法相比，我们有了显著的改进。

Generative Adversarial Networks (GANs) produce impressive results on unconditional image generation when powered with large-scale image datasets. Yet generated images are still easy to spot especially on datasets with high variance (e.g. bedroom, church). In this paper, we propose various improvements to further push the boundaries in image generation. Specifically, we propose a novel dual contrastive loss and show that, with this loss, discriminator learns more generalized and distinguishable representations to incentivize generation. In addition, we revisit attention and extensively experiment with different attention blocks in the generator. We find attention to be still an important module for successful image generation even though it was not used in the recent state-of-the-art models. Lastly, we study different attention architectures in the discriminator, and propose a reference attention mechanism. By combining the strengths of these remedies, we improve the compelling state-of-the-art Frechet Inception Distance (FID) by at least 17.5% on several benchmark datasets. We obtain even more significant improvements on compositional synthetic scenes (up to 47.5% in FID).

生成性对抗网络（GAN）在使用大规模图像数据集时，在无条件生成图像方面产生了令人印象深刻的效果。然而，生成的图像仍然很容易被发现，特别是在差异较大的数据集上（例如卧室、教堂）。在本文中，我们提出了各种改进，以进一步推动图像生成中的边界。具体地说，我们提出了一种新的双重对比损失，并表明，利用这种损失，鉴别器可以学习更广义和更可区分的表示来激励生成。此外，我们还回顾了注意力，并在生成器中对不同的注意力块进行了广泛的实验。我们发现，关注仍然是成功生成图像的一个重要模块，即使它没有在最近最先进的模型中使用。最后，我们研究了鉴别器中不同的注意结构，并提出了一种参考注意机制。通过结合这些补救措施的优点，我们在几个基准数据集上将最先进的Frechet起始距离（FID）提高了至少17.5%。我们在合成场景方面取得了更显著的改进（FID中的改进率高达47.5%）。

Category-level 6D object pose and size estimation is to predict full pose configurations of rotation, translation, and size for object instances observed in single, arbitrary views of cluttered scenes. In this paper, we propose a new method of Dual Pose Network with refined learning of pose consistency for this task, shortened as DualPoseNet. DualPoseNet stacks two parallel pose decoders on top of a shared pose encoder, where the implicit decoder predicts object poses with a working mechanism different from that of the explicit one; they thus impose complementary supervision on the training of pose encoder. We construct the encoder based on spherical convolutions, and design a module of Spherical Fusion wherein for a better embedding of pose-sensitive features from the appearance and shape observations. Given no testing CAD models, it is the novel introduction of the implicit decoder that enables the refined pose prediction during testing, by enforcing the predicted pose consistency between the two decoders using a self-adaptive loss term. Thorough experiments on benchmarks of both category- and instance-level object pose datasets confirm efficacy of our designs. DualPoseNet outperforms existing methods with a large margin in the regime of high precision. Our code is released publicly at <https://github.com/Gorilla-Lab-SCUT/DualPoseNet>.

类别级别6D对象姿势和大小估计用于预测在杂乱场景的单个任意视图中观察到的对象实例的旋转、平移和大小的完整姿势配置。在本文中，我们提出了一种新的双姿态网络方法，该方法具有姿态一致性的精细学习，简称为双posenet。DualPoseNet将两个平行的姿势解码器堆叠在共享姿势编码器的顶部，其中隐式解码器使用不同于显式解码器的工作机制预测对象姿势；因此，他们对姿势编码器的培训实施补充监督。我们构建了基于球形卷积的编码器，并设计了一个球形融合模块，其中可以更好地嵌入来自外观和形状观测的姿势敏感特征。在没有测试CAD模型的情况下，通过使用自适应损失项在两个解码器之间强制实现预测的姿势一致性，隐式解码器的新颖引入使得在测试期间能够实现精确的姿势预测。在类别级和实例级对象姿态数据集的基准上进行的彻底实验证实了我们设计的有效性。DualPoseNet在高精度方面优于现有方法。我们的代码在<https://github.com/Gorilla-Lab-SCUT/DualPoseNet>.

Multi-modal reasoning systems rely on a pre-trained object detector to extract regions of interest from the image. However, this crucial module is typically used as a black box, trained independently of the downstream task and on a fixed vocabulary of objects and attributes. This makes it challenging for such systems to capture the long tail of visual concepts expressed in free form text. In this paper we propose MDETR, an end-to-end modulated detector that detects objects in an image conditioned on a raw text query, like a caption or a question. We use a transformer-based architecture to reason jointly over text and image by fusing the two modalities at an early stage of the model. We pre-train the network on 1.3M text-image pairs, mined from pre-existing multi-modal datasets having explicit alignment between phrases in text and objects in the image. We then fine-tune on several downstream tasks such as phrase grounding, referring expression comprehension and segmentation, achieving state-of-the-art results on popular benchmarks. We also investigate the utility of our model as an object detector on a given label set when fine-tuned in a few-shot setting. We show that our pre-training approach provides a way to handle the long tail of object categories which have very few labelled instances. Our approach can be easily extended for visual question answering, achieving competitive performance on GQA and CLEVR. The code and models are available at <https://github.com/ashkamath/mdetr>.

多模态推理系统依靠预先训练的目标检测器从图像中提取感兴趣的区域。然而，这一关键模块通常被用作一个黑匣子，独立于下游任务进行训练，并使用固定的对象和属性词汇表。这使得这类系统很难捕捉以自由形式文本表达的视觉概念的长尾。在本文中，我们提出了MDETR，一种端到端的调制检测器，用于检测图像中以原始文本查询为条件的对象，如标题或问题。我们使用基于转换器的架构，通过在模型的早期阶段融合这两种模式，对文本和图像进行联合推理。我们在130万个文本图像对上对网络进行预

训练，这些文本图像对是从预先存在的多模态数据集中挖掘出来的，这些数据集在文本中的短语和图像中的对象之间具有明确的对齐方式。然后，我们对几个下游任务进行微调，例如短语基础、引用表达式理解和分割，在流行基准上获得最先进的结果。我们还研究了当在几个镜头设置中进行微调时，我们的模型作为给定标签集上的对象检测器的实用性。我们表明，我们的预训练方法提供了一种处理对象类别的长尾的方法，这些对象类别具有很少的标记实例。我们的方法可以很容易地扩展为可视化问答，在GQA和CLEVR上实现有竞争力的性能。有关代码和模型，请访问<https://github.com/ashkamath/mdetr>.

In this paper, we propose the first minimal solutions for estimating the semi-generalized homography given a perspective and a generalized camera. The proposed solvers use five 2D-2D image point correspondences induced by a scene plane. One group of solvers assumes the perspective camera to be fully calibrated, while the other estimates the unknown focal length together with the absolute pose parameters. This setup is particularly important in structure-from-motion and visual localization pipelines, where a new camera is localized in each step with respect to a set of known cameras and 2D-3D correspondences might not be available. Thanks to a clever parametrization and the elimination ideal method, our solvers only need to solve a univariate polynomial of degree five or three, respectively a system of polynomial equations in two variables. All proposed solvers are stable and efficient as demonstrated by a number of synthetic and real-world experiments.

在本文中，我们提出了在给定透视和广义摄像机的情况下估计半广义单应的第一个极小解。所提出的解算器使用由场景平面诱导的五个2D-2D图像点对应。一组解算器假设透视摄影机已完全校准，而另一组解算器估计未知焦距以及绝对姿态参数。此设置在运动和视觉定位管道的结构中尤其重要，其中，在每个步骤中，新摄影机相对于一组已知摄影机进行定位，并且2D-3D对应可能不可用。由于巧妙的参数化和消去理想方法，我们的解算器只需要解五次或三次的一元多项式，分别解两个变量的多项式方程组。通过大量的合成和真实实验证明，所有提出的求解器都是稳定和高效的。

Instance segmentation models today are very accurate when trained on large annotated datasets, but collecting mask annotations at scale is prohibitively expensive. We address the partially supervised instance segmentation problem in which one can train on (significantly cheaper) bounding boxes for all categories but use masks only for a subset of categories. In this work, we focus on a popular family of models which apply differentiable cropping to a feature map and predict a mask based on the resulting crop. Under this family, we study Mask R-CNN and discover that instead of its default strategy of training the mask-head with a combination of proposals and groundtruth boxes, training the mask-head with only groundtruth boxes dramatically improves its performance on novel classes. This training strategy also allows us to take advantage of alternative mask-head architectures, which we exploit by replacing the typical mask-head of 2-4 layers with significantly deeper off-the-shelf architectures (e.g. ResNet, Hourglass models). While many of these architectures perform similarly when trained in fully supervised mode, our main finding is that they can generalize to novel classes in dramatically different ways. We call this ability of mask-heads to generalize to unseen classes the strong mask generalization effect and show that without any specialty modules or losses, we can achieve state-of-the-art results in the partially supervised COCO instance segmentation benchmark. Finally, we demonstrate that our effect is general, holding across underlying detection methodologies (including anchor-based, anchor-free or no detector at all) and across different backbone networks. Code and pre-trained models are available at <https://git.io/deepmac>.

今天的实例分割模型在大的带注释的数据集上训练时非常精确，但大规模收集掩码注释的成本高得令人望而却步。我们解决了部分监督的实例分割问题，在该问题中，可以对所有类别的边界框进行训练（非常便宜），但只对类别的子集使用遮罩。在这项工作中，我们关注一系列流行的模型，这些模型将可微裁剪应用于特征地图，并根据生成的裁剪预测遮罩。在这个家族中，我们研究了Mask R-CNN，发现它的默认策略不是用提案和背景真相盒组合训练面具头，而是只用背景真相盒训练面具头，这大大提高了它在小说类中的表现。该培训策略还允许我们利用替代掩模头架构，我们利用该架构将2-4层的典型掩模头替换为更深入的现成架构（如ResNet、沙漏模型）。虽然这些体系结构中的许多在完全监督模式下训练时表现类似，但我们的主要发现是，它们可以以截然不同的方式推广到新类。我们将面具头的这种能力称为强大的面具泛化效应，并表明在没有任何特殊模块或损失的情况下，我们可以在部分监督的COCO实例分割基准中获得最先进的结果。最后，我们证明了我们的效果是通用的，适用于底层检测方法（包括基于锚、无锚或根本没有检测器）和不同的主干网络。代码和预先培训的模型可在<https://git.io/deepmac>。

Contemporary domain generalization (DG) and multi-source unsupervised domain adaptation (UDA) methods mostly collect data from multiple domains together for joint optimization. However, this centralized training paradigm poses a threat to data privacy and is not applicable when data are non-shared across domains. In this work, we propose a new approach called collaborative optimization and Aggregation (COPA), which aims at optimizing a generalized target model for decentralized DG and UDA, where data from different domains are non-shared and private. Our base model consists of a domain-invariant feature extractor and an ensemble of domain-specific classifiers. In an iterative learning process, we optimize a local model for each domain, and then centrally aggregate local feature extractors and assemble domain-specific classifiers to construct a generalized global model, without sharing data from different domains. To improve generalization of feature extractors, we employ hybrid batch-instance normalization and collaboration of frozen classifiers. For better decentralized UDA, we further introduce a prediction agreement mechanism to overcome local disparities towards central model aggregation. Extensive experiments on five DG and UDA benchmark datasets show that COPA is capable of achieving comparable performance against the state-of-the-art DG and UDA methods without the need for centralized data collection in model training.

当代的领域综合 (DG) 和多源无监督领域自适应 (UDA) 方法主要是从多个领域收集数据进行联合优化。然而，这种集中化的培训模式对数据隐私构成威胁，并且在跨域数据不共享时不适用。在这项工作中，我们提出了一种称为协作优化和聚合 (COPA) 的新方法，该方法旨在优化分散DG和UDA的广义目标模型，其中来自不同领域的数据是非共享和私有的。我们的基本模型由一个领域不变特征提取器和一组领域特定分类器组成。在一个迭代学习过程中，我们为每个领域优化一个局部模型，然后集中聚局局部特征提取器并组装特定于领域的分类器来构造一个广义全局模型，而不共享来自不同领域的数据。为了提高特征抽取器的泛化能力，我们采用混合批量实例规范化和冻结分类器的协作。为了更好地分散 UDA，我们进一步引入了预测一致机制，以克服中央模型聚合的局部差异。在五个DG和UDA基准数据集上进行的大量实验表明，COPA能够实现与最先进的DG和UDA方法相当的性能，而无需在模型训练中集中收集数据。

Attention module does not always help deep models learn causal features that are robust in any confounding context, e.g., a foreground object feature is invariant to different backgrounds. This is because the confounders trick the attention to capture spurious correlations that benefit the prediction when the training and testing data are IID (identical & independent distribution); while harm the prediction when the data are OOD (out-of-distribution). The sole fundamental solution to learn causal attention is by causal intervention, which requires additional annotations of the confounders, e.g., a "dog" model is learned within "grass+dog" and "road+dog" respectively, so the "grass" and "road" contexts will no longer confound the "dog" recognition. However, such annotation is not only prohibitively expensive, but also inherently problematic, as the confounders are elusive in nature. In this paper, we propose a causal attention module (CaaM) that self-annotates the confounders in unsupervised fashion. In particular, multiple CaaMs can be stacked and integrated in conventional attention CNN and self-attention Vision Transformer. In OOD settings, deep models with CaaM outperform those without it significantly; even in IID settings, the attention localization is also improved by CaaM, showing a great potential in applications that require robust visual saliency. Codes are available at <https://github.com/Wangt-CN/CaaM>.

注意模块并不总是帮助深度模型学习在任何混杂背景下都具有鲁棒性的因果特征，例如，前景对象特征对不同背景是不变的。这是因为当训练和测试数据为IID（相同和独立分布）时，混杂因素欺骗了注意力，以捕获有利于预测的虚假相关性；而当数据分布不均时，会损害预测。学习因果注意的唯一根本解决方案是通过因果干预，这需要对混杂因素进行额外的注释，例如，分别在“草+狗”和“路+狗”中学习“狗”模型，因此“草”和“路”上下文将不再混淆“狗”识别。然而，这样的注释不仅昂贵得令人望而却步，而且本身就存在问题，因为混淆因素在本质上是难以捉摸的。在本文中，我们提出了一个因果注意模块（CaaM），它以无监督的方式对混杂因素进行自我注释。特别是，多个CAAM可以堆叠并集成在传统的注意力CNN和自我注意力视觉转换器中。在OOD设置中，有CaaM的深度模型的性能显著优于没有CaaM的深度模型；即使在IID设置中，CaaM也可以改进注意力定位，在需要鲁棒视觉显著性的应用中显示出巨大的潜力。代码可在<https://github.com/Wangt-CN/CaaM>。

We propose a new approach to detect atypicality in persuasive imagery. Unlike atypicality which has been studied in prior work, persuasive atypicality has a particular purpose to convey meaning, and relies on understanding the common-sense spatial relations of objects. We propose a self-supervised attention-based technique which captures contextual compatibility, and models spatial relations in a precise manner. We further experiment with capturing common sense through the semantics of co-occurring object classes. We verify our approach on a dataset of atypicality in visual advertisements, as well as a second dataset capturing atypicality that has no persuasive intent.

我们提出了一种新的方法来检测说服性意象中的非典型性。与先前研究的非典型性不同，说服性非典型性具有传达意义的特殊目的，并且依赖于理解物体的常识性空间关系。我们提出了一种基于自我监督注意的技术，该技术能够捕获上下文兼容性，并以精确的方式对空间关系进行建模。我们进一步尝试通过共现对象类的语义来获取常识。我们在视觉广告中的非典型性数据集上验证了我们的方法，以及第二个捕获非典型性的数据集，该数据集没有说服力。

Spatio-temporal video grounding (STVG) aims to localize a spatio-temporal tube of a target object in an untrimmed video based on a query sentence. In this work, we propose a one-stage visual-linguistic transformer based framework called STVGBert for the STVG task, which can simultaneously localize the target object in both spatial and temporal domains. Specifically, without resorting to pre-generated object proposals, our STVGBert directly takes a video and a query sentence as the input, and then produces the cross-modal features by using the newly introduced cross-modal feature learning module ST-ViLBert. Based on the cross-modal features, our method then generates bounding boxes and predicts the starting and ending frames to produce the predicted object tube. To the best of our knowledge, our STVGBert is the first one-stage method, which can handle the STVG task without relying on any pre-trained object detectors. Comprehensive experiments demonstrate our newly proposed framework outperforms the state-of-the-art multi-stage methods on two benchmark datasets Vid-STG and HC-STVG.

时空视频接地 (STVG) 旨在基于查询语句对未经剪辑的视频中的目标对象的时空管进行定位。在这项工作中，我们为STVG任务提出了一个基于视觉语言转换器的单阶段框架，称为STVGBert，它可以同时在空间和时间域中定位目标对象。具体地说，我们的STVGBert不借助预生成的对象建议，直接将视频和查询语句作为输入，然后使用新引入的跨模态特征学习模块ST ViLBert生成跨模态特征。基于交叉模态特征，我们的方法生成边界框并预测起始帧和结束帧，从而生成预测的目标管。据我们所知，我们的STVGBert是第一个单阶段方法，它可以处理STVG任务，而不依赖任何预先训练的对象检测器。综合实验表明，在Vid STG和HC-STVG两个基准数据集上，我们新提出的框架优于最先进的多阶段方法。

Deep Learning (DL)-based methods have achieved great success in solving the ill-posed JPEG compression artifacts removal problem. However, as most DL architectures are designed to directly learn pixel-level mapping relationships, they largely ignore semantic-level information and lack sufficient interpretability. To address the above issues, in this work, we propose an interpretable deep network to learn both pixel-level regressive prior and semantic-level discriminative prior. Specifically, we design a variational model to formulate the image de-blocking problem and propose two prior terms for the image content and gradient, respectively. The content-relevant prior is formulated as a DL-based image-to-image regressor to perform as a de-blocker from the pixel-level. The gradient-relevant prior serves as a DL-based classifier to distinguish whether the image is compressed from the semantic-level. To effectively solve the variational model, we design an alternating minimization algorithm and unfold it into a deep network architecture. In this way, not only the interpretability of the deep network is increased, but also the dual priors can be well estimated from training samples. By integrating the two priors into a single framework, the image de-blocking problem can be well-constrained, leading to a better performance. Experiments on benchmarks and real-world use cases demonstrate the superiority of our method to the existing state-of-the-art approaches.

基于深度学习 (DL) 的方法在解决病态JPEG压缩伪影去除问题方面取得了巨大成功。然而，由于大多数DL体系结构被设计为直接学习像素级映射关系，它们在很大程度上忽略了语义级信息，并且缺乏足够的可解释性。为了解决上述问题，在这项工作中，我们提出了一个可解释的深度网络来学习像素级的回归先验和语义级的判别先验。具体来说，我们设计了一个变分模型来描述图像去块问题，并分别提出了图像内容和梯度的两个先验项。与先前相关的内容被公式化为基于DL的图像到图像回归器，以从像素级执行去块。梯度相关先验作为一个基于DL的分类器，从语义层区分图像是否被压缩。为了有效地求解变分模型，我们设计了一种交替最小化算法，并将其展开为一个深层网络结构。这样，不仅提高了深度网络的可解释性，而且可以从训练样本中很好地估计双先验。通过将这两个先验知识集成到一个框架中，可以很好地约束图像去块问题，从而获得更好的性能。在基准测试和实际用例上的实验证明了我们的方法比现有的最先进的方法优越。

Conventional video models rely on a single stream to capture the complex spatial-temporal features. Recent work on two-stream video models, such as SlowFast network and AssembleNet, prescribe separate streams to learn complementary features, and achieve stronger performance. However, manually designing both streams as well as the in-between fusion blocks is a daunting task, requiring to explore a tremendously large design space. Such manual exploration is time-consuming and often ends up with sub-optimal architectures when computational resources are limited and the exploration is insufficient. In this work, we present a pragmatic neural architecture search approach, which is able to search for two-stream video models in giant spaces efficiently. We design a multivariate search space, including 6 search variables to capture a wide variety of choices in designing two-stream models. Furthermore, we propose a progressive search procedure, by searching for the architecture of individual streams, fusion blocks and attention blocks one after the other. We demonstrate two-stream models with significantly better performance can be automatically discovered in our design space. Our searched two-stream models, namely Auto-TSNet, consistently outperform other models on standard benchmarks. On Kinetics, compared with the SlowFast model, our Auto-TSNet-L model reduces FLOPS by nearly 11 times while achieving the same accuracy 78.9%. On Something-Something-v2, Auto-TSNet-M improves the accuracy by at least 2% over other methods which use less than 50 GFLOPS per video.

传统的视频模型依赖于单个流来捕获复杂的时空特征。最近关于两个流视频模型（如SlowFast network 和AssembleNet）的工作规定了单独的流来学习互补功能，并实现了更强的性能。然而，手动设计两个流以及中间融合块是一项艰巨的任务，需要探索巨大的设计空间。这样的人工探索非常耗时，当计算资源有限且探索不足时，往往会以次优架构告终。在这项工作中，我们提出了一种实用的神经结构搜索方法，它能够在巨大的空间中高效地搜索两个流视频模型。我们设计了一个多元搜索空间，包括6个搜索变量，以捕获设计两个流模型时的各种选择。此外，我们提出了一种渐进式搜索过程，通过逐个搜索各个流、融合块和注意块的结构。我们演示了在我们的设计空间中可以自动发现两个性能显著更好的流模型。我们搜索的两个流模型，即Auto-TSNet，在标准基准上始终优于其他模型。在动力学方面，与慢速模型相比，我们的Auto-TSNet-L模型减少了近11倍的失败，同时实现了78.9%的相同精度。在Something-Something-V2上，Auto-TSNet-M比其他每个视频使用不到50gflops的方法至少提高了2%的准确性。

We propose a novel loss weighting algorithm, called loss scale balancing (LSB), for multi-task learning (MTL) of pixelwise vision tasks. An MTL model is trained to estimate multiple pixelwise predictions using an overall loss, which is a linear combination of individual task losses. The proposed algorithm dynamically adjusts the linear weights to learn all tasks effectively. Instead of controlling the trend of each loss value directly, we balance the loss scale --- the product of the loss value and its weight --- periodically. In addition, by evaluating the difficulty of each task based on the previous loss record, the proposed algorithm focuses more on difficult tasks during training. Experimental results show that the proposed algorithm outperforms conventional weighting algorithms for MTL of various pixelwise tasks. Codes are available at <https://github.com/jaehanlee-mcl/LSB-MTL>.

我们提出了一种新的损失加权算法，称为损失尺度平衡 (LSB)，用于像素视觉任务的多任务学习 (MTL)。MTL模型经过训练，可以使用总体损失（单个任务损失的线性组合）来估计多个像素预测。该算法动态调整线性权重，有效地学习所有任务。我们不是直接控制每个损失值的趋势，而是周期性地平衡损失量表——损失值与其权重的乘积。此外，该算法通过根据之前的损失记录评估每个任务的难度，在训练过程中更加关注困难任务。实验结果表明，对于不同像素任务的MTL，该算法优于传统的加权算法。代码可在<https://github.com/jaehanlee-mcl/LSB-MTL>。

Benefiting from the excellent performance of Siamese-based trackers, huge progress on 2D visual tracking has been achieved. However, 3D visual tracking is still under-explored. Inspired by the idea of Hough voting in 3D object detection, in this paper, we propose a Multi-level Voting Siamese Network (MLVSNet) for 3D visual tracking from outdoor point cloud sequences. To deal with sparsity in outdoor 3D point clouds, we propose to perform Hough voting on multi-level features to get more vote centers and retain more useful information, instead of voting only on the final level feature as in previous methods. We also design an efficient and lightweight Target-Guided Attention (TGA) module to transfer the target information and highlight the target points in the search area. Moreover, we propose a Vote-cluster Feature Enhancement (VFE) module to exploit the relationships between different vote clusters. Extensive experiments on the 3D tracking benchmark of KITTI dataset demonstrate that our MLVSNet outperforms state-of-the-art methods with significant margins. Code will be available at <https://github.com/CodewZT/MLVSNet>.

得益于基于暹罗的跟踪器的出色性能，2D视觉跟踪取得了巨大的进步。然而，三维视觉跟踪仍处于探索阶段。受三维目标检测中Hough投票的启发，本文提出了一种用于室外点云序列三维视觉跟踪的多级投票连体网络（MLVSNet）。为了解决室外三维点云的稀疏性问题，我们建议对多层次特征进行Hough投票，以获得更多的投票中心并保留更多有用的信息，而不是像以前的方法那样只对最终层次特征进行投票。我们还设计了一个高效、轻量级的目标引导注意（TGA）模块，用于传递目标信息和突出搜索区域中的目标点。此外，我们还提出了一个投票簇特征增强（VFE）模块来利用不同投票簇之间的关系。在KITTI数据集的3D跟踪基准上进行的大量实验表明，我们的MLVSNet优于最先进的方法，具有显著的优势。代码将在<https://github.com/CodewZT/MLVSNet>。

One-stage long-tailed recognition methods improve the overall performance in a "seesaw" manner, i.e., either sacrifice the head's accuracy for better tail classification or elevate the head's accuracy even higher but ignore the tail. Existing algorithms bypass such trade-off by a multi-stage training process: pre-training on imbalanced set and fine-tuning on balanced set. Though achieving promising performance, not only are they sensitive to the generalizability of the pre-trained model, but also not easily integrated into other computer vision tasks like detection and segmentation, where pre-training of classifier solely is not applicable. In this paper, we propose a one-stage long-tailed recognition scheme, ally complementary experts (ACE), where the expert is the most knowledgeable specialist in a sub-set that dominates its training, and is complementary to other experts in the less-seen categories without disturbed by what it has never seen. We design a distribution-adaptive optimizer to adjust the learning pace of each expert to avoid over-fitting. Without special bells and whistles, the vanilla ACE outperforms the current one-stage SOTA method by 3-10% on CIFAR10-LT, CIFAR100-LT, ImageNet-LT and iNaturalist datasets. It is also shown to be the first one to break the "seesaw" trade-off by improving the accuracy of the majority and minority categories simultaneously in only one stage.

单阶段长尾识别方法以“跷跷板”的方式提高整体性能，即要么牺牲头部的准确度以获得更好的尾部分类，要么将头部的准确度提高得更高，但忽略尾部。现有的算法通过一个多阶段的训练过程来绕过这种权衡：对不平衡集进行预训练和对平衡集进行微调。虽然取得了很好的性能，但它们不仅对预先训练的模型的可推广性敏感，而且不容易集成到其他计算机视觉任务中，如检测和分割，而单独对分类器进行预训练是不适用的。在本文中，我们提出了一个单阶段长尾识别方案，即联合互补专家（ACE），其中专家是主导其训练的子集中知识最丰富的专家，并且在不太常见的类别中与其他专家互补，而不受从未见过的干扰。我们设计了一个分布自适应优化器来调整每个专家的学习速度，以避免过度拟合。在没有特殊提示的情况下，vanilla ACE在CIFAR10-LT、CIFAR100-LT、ImageNet LT和iNaturalist数据集上的

性能比当前的单阶段SOTA方法高出3~10%。它也被证明是第一个打破“跷跷板”权衡的方法，即在一个阶段内同时提高多数和少数类别的准确性。

The hyperspectral image (HSI) denoising has been widely utilized to improve HSI qualities. Recently, learning-based HSI denoising methods have shown their effectiveness, but most of them are based on synthetic dataset and lack the generalization capability on real testing HSI. Moreover, there is still no public paired real HSI denoising dataset to learn HSI denoising network and quantitatively evaluate HSI methods. In this paper, we mainly focus on how to produce realistic dataset for learning and evaluating HSI denoising network. On the one hand, we collect a paired real HSI denoising dataset, which consists of shortexposure noisy HSIs and the corresponding long-exposure clean HSIs. On the other hand, we propose an accurate HSI noise model which matches the distribution of real data well and can be employed to synthesize realistic dataset. On the basis of the noise model, we present an approach to calibrate the noise parameters of the given hyperspectral camera. The extensive experimental results show that a network learned with only synthetic data generated by our noise model performs as well as it is learned with paired real data.

高光谱图像 (HSI) 去噪已被广泛用于改善HSI质量。近年来，基于学习的HSI去噪方法已显示出其有效性，但大多数方法都是基于合成数据集，缺乏对真实HSI测试的泛化能力。此外，还没有公开的成对真实HSI去噪数据集来学习HSI去噪网络和定量评估HSI方法。在本文中，我们主要研究如何生成真实的数据集来学习和评估HSI去噪网络。一方面，我们收集一对真实HSI去噪数据集，该数据集由短曝光噪声HSI和相应的长曝光清洁HSI组成。另一方面，我们提出了一个精确的HSI噪声模型，该模型能够很好地匹配真实数据的分布，并可用于合成真实数据集。在噪声模型的基础上，提出了一种标定给定高光谱相机噪声参数的方法。大量的实验结果表明，仅使用我们的噪声模型生成的合成数据学习的网络性能与使用成对真实数据学习的网络性能相同。

The lack of large-scale real raw image denoising dataset gives the rise to challenges on synthesizing realistic raw image noise for training denoising models. However, the real raw image noise is contributed by many noise sources and varies greatly among different sensors. Existing methods are unable to model all noise sources accurately, and building a noise model for each sensor is also laborious. In this paper, we introduce a new perspective to synthesize noise by directly sampling from the sensor's real noise. It inherently generates accurate raw image noise for different camera sensors. Two efficient and generic techniques: pattern-aligned patch sampling and high-bit reconstruction help accurate synthesis of spatial-correlated noise and high-bit noise respectively. We conduct systematic experiments on SIDD and ELD datasets. The results show that (1) our method outperforms existing methods and demonstrates wide generalization on different sensors and lighting conditions. (2) Recent conclusions derived from DNN-based noise modeling methods are actually based on inaccurate noise parameters. The DNN-based methods still cannot outperform physics-based statistical methods.

由于缺乏大规模的真实原始图像去噪数据集，合成真实原始图像噪声以训练去噪模型成为一个挑战。然而，真实的原始图像噪声是由许多噪声源造成的，并且在不同的传感器之间差异很大。现有的方法无法对所有噪声源进行精确建模，为每个传感器建立噪声模型也很困难。本文介绍了一种从传感器真实噪声中直接采样合成噪声的新方法。它固有地为不同的摄像机传感器生成精确的原始图像噪声。两种高效且通用的技术：模式对齐面片采样和高位重建分别有助于精确合成空间相关噪声和高位噪声。我们在SIDD和ELD数据集上进行了系统的实验。结果表明：（1）我们的方法优于现有的方法，在不同的传感器和光照条件下具有广泛的泛化性。（2）最近从基于DNN的噪声建模方法得出的结论实际上是基于不准确的噪声参数。基于DNN的方法仍然不能优于基于物理的统计方法。

General face recognition has seen remarkable progress in recent years. However, large age gap still remains a big challenge due to significant alterations in facial appearance and bone structure. Disentanglement plays a key role in partitioning face representations into identity-dependent and age-dependent components for age-invariant face recognition (AIFR). In this paper we propose a multi-task learning framework based on mutual information minimization (MT-MIM), which casts the disentangled representation learning as an objective of information constraints. The method trains a disentanglement network to minimize mutual information between the identity component and age component of the face image from the same person, and reduce the effect of age variations during the identification process. For quantitative measure of the degree of disentanglement, we verify that mutual information can represent as metric. The resulting identity-dependent representations are used for age-invariant face recognition. We evaluate MT-MIM on popular public-domain face aging datasets (FG-NET, MORPH Album 2, CACD and AgeDB) and obtained significant improvements over previous state-of-the-art methods. Specifically, our method exceeds the baseline models by over 0.4% on MORPH Album 2, and over 0.7% on CACD subsets, which are impressive improvements at the high accuracy levels of above 99% and an average of 94%.

近年来，通用人脸识别技术取得了显著的进展。然而，由于面部外观和骨骼结构的显著改变，较大的年龄差距仍然是一个巨大的挑战。对于年龄不变的人脸识别（AIFR），解纠缠在将人脸表示划分为身份相关和年龄相关两个分量方面起着关键作用。本文提出了一种基于互信息最小化的多任务学习框架（MT-MIM），该框架将非纠缠表示学习作为信息约束的目标。该方法训练一个解纠缠网络，以最小化来自同一人的人脸图像的身份分量和年龄分量之间的互信息，并减少识别过程中年龄变化的影响。对于解纠缠度的定量度量，我们验证了互信息可以表示为度量。由此产生的身份相关表示用于年龄不变的人脸识别。我们在流行的公共领域人脸老化数据集（FG-NET、MORPH Album 2、CACD和AgeDB）上对MT-MIM进行了评估，并与以前最先进的方法相比取得了显著的改进。具体而言，我们的方法在MORPH Album 2上比基线模型高出0.4%以上，在CACD子集上比基线模型高出0.7%以上，在99%以上的高精度水平和平均94%的高精度水平上都有令人印象深刻的改进。

Ternary Neural Networks (TNNS) have received much attention due to being potentially orders of magnitude faster in inference, as well as more power efficient, than full-precision counterparts. However, 2 bits are required to encode the ternary representation with only 3 quantization levels leveraged. As a result, conventional TNNS have similar memory consumption and speed compared with the standard 2-bit models, but have worse representational capability. Moreover, there is still a significant gap in accuracy between TNNS and full-precision networks, hampering their deployment to real applications. To tackle these two challenges, in this work, we first show that, under some mild constraints, computational complexity of the ternary inner product can be reduced by 2x. Second, to mitigate the performance gap, we elaborately design an implementation-dependent ternary quantization algorithm. The proposed framework is termed Fast and Accurate Ternary Neural Networks (FATNN). Experiments on image classification demonstrate that our FATNN surpasses the state-of-the-arts by a significant margin in accuracy. More importantly, speedup evaluation compared with various precisions is analyzed on several platforms, which serves as a strong benchmark for further research.

三元神经网络（TNN）由于比全精度神经网络具有更快的推理速度和更高的能量效率而受到广泛关注。然而，仅使用3个量化级别对三值表示进行编码需要2位。因此，与标准2位模型相比，传统TNN具有相似的内存消耗和速度，但表示能力较差。此外，TNN和全精度网络之间在准确性方面仍然存在巨大差距，这阻碍了它们在实际应用中的部署。为了解决这两个挑战，在这项工作中，我们首先表明，在一些温和的约束下，三元内积的计算复杂度可以降低2倍。其次，为了缩小性能差距，我们精心设计了一种依赖于实现的三值量化算法。该框架被称为快速准确的三元神经网络（FATNN）。对图像分类的实验表

明，我们的FATNN在准确度上大大超过了现有技术。更重要的是，在多个平台上分析了不同精度下的加速比评估，为进一步研究提供了有力的基准。

Dark environment becomes a challenge for computer vision algorithms owing to insufficient photons and undesirable noises. Most of the existing studies tackle this by either targeting human vision for better visual perception or improving the machine vision for specific high-level tasks. In addition, these methods rely on data augmentation and directly train their models based on real-world or over-simplified synthetic datasets without exploring the intrinsic pattern behind illumination translation. Here, we propose a novel multitask auto encoding transformation (MAET) model that combines human vision and machine vision tasks to enhance object detection in a dark environment. With a self-supervision learning, the MAET learns an intrinsic visual structure by encoding and decoding the realistic illumination-degrading transformation considering the physical noise model and image signal processing (ISP). Based on this representation, we achieve object detection task by decoding the bounding box coordinates and classes. To avoid the over-entanglement of two tasks, our MAET disentangles the object and degrading features by imposing an orthogonal tangent regularity. This forms a parametric manifold along which multitask predictions can be geometrically formulated by maximizing the orthogonality between the tangents along the outputs of respective tasks. Our framework can be implemented based on the mainstream object detection architecture and directly trained end-to-end using the normal target detection datasets, such as COCO and VOC. We have achieved the state-of-the-art performance using synthetic and real-world datasets.

由于光子不足和噪声的影响，暗环境成为计算机视觉算法的一个挑战。现有的大多数研究都是通过针对人类视觉以获得更好的视觉感知或改善特定高级任务的机器视觉来解决这一问题。此外，这些方法依赖于数据论证，直接基于真实世界或过于简化的合成数据集训练其模型，而不探索照明转换背后的内在模式。在这里，我们提出了一种新的多任务自动编码转换（MAET）模型，该模型结合了人类视觉和机器视觉任务，以增强黑暗环境中的目标检测。通过自监督学习，MAET在考虑物理噪声模型和图像信号处理（ISP）的情况下，通过对真实光照退化变换进行编码和解码来学习内在的视觉结构。基于这种表示，我们通过对边界框坐标和类进行解码来实现目标检测任务。为了避免两个任务的过度纠缠，我们的MAET通过施加正交正切规则来分离对象和退化特征。这形成了一个参数流形，通过最大化各个任务输出的切线之间的正交性，可以在几何上表示多任务预测。我们的框架可以基于主流的目标检测体系结构来实现，并使用普通的目标检测数据集（如COCO和VOC）直接进行端到端训练。我们使用合成数据集和真实数据集实现了最先进的性能。

The goal of unsupervised co-localization is to locate the object in a scene under the assumptions that 1) the dataset consists of only one superclass, e.g., birds, and 2) there are no human-annotated labels in the dataset. The most recent method achieves impressive co-localization performance by employing self-supervised representation learning approaches such as predicting rotation. In this paper, we introduce a new contrastive objective directly on the attention maps to enhance co-localization performance. Our contrastive loss function exploits rich information of location, which induces the model to activate the extent of the object effectively. In addition, we propose a pixel-wise attention pooling that selectively aggregates the feature map regarding their magnitudes across channels. Our methods are simple and shown effective by extensive qualitative and quantitative evaluation, achieving state-of-the-art co-localization performances by large margins on four datasets: CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Stanford Dogs. Our code will be publicly available online for the research community.

无监督协同定位的目标是在以下假设下定位场景中的对象：1) 数据集仅包含一个超类，例如鸟类；2) 数据集中没有人类注释的标签。最新的方法通过使用自监督表示学习方法，如预测旋转，实现了令人印象深刻的共定位性能。在本文中，我们直接在注意图上引入一个新的对比目标，以提高共定位性能。我们的对比损失函数利用了丰富的位置信息，从而诱导模型有效地激活对象的范围。此外，我们还提出了一种像素级的注意力池，可以有选择地聚集通道上的特征图。通过广泛的定性和定量评估，我们的方法简单有效，在四个数据集（CUB-200-2011、斯坦福汽车、FGVC飞机和斯坦福狗）上大幅度实现了最先进的共定位性能。我们的代码将在研究社区的网上公开。

In real-life applications, machine learning models often face scenarios where there is a change in data distribution between training and test domains. When the aim is to make predictions on distributions different from those seen at training, we incur in a domain generalization problem. Methods to address this issue learn a model using data from multiple source domains, and then apply this model to the unseen target domain. Our hypothesis is that when training with multiple domains, conflicting gradients within each mini-batch contain information specific to the individual domains which is irrelevant to the others, including the test domain. If left untouched, such disagreement may degrade generalization performance. In this work, we characterize the conflicting gradients emerging in domain shift scenarios and devise novel gradient agreement strategies based on gradient surgery to alleviate their effect. We validate our approach in image classification tasks with three multi-domain datasets, showing the value of the proposed agreement strategy in enhancing the generalization capability of deep learning models in domain shift scenarios.

在实际应用中，机器学习模型通常面临训练域和测试域之间的数据分布发生变化的场景。当目标是预测不同于训练时所看到的分布时，我们产生了一个领域泛化问题。解决此问题的方法使用来自多个源域的数据学习模型，然后将此模型应用于看不见的目标域。我们的假设是，当使用多个域进行训练时，每个小批次中的冲突梯度包含与其他域（包括测试域）无关的单个域的特定信息。如果保持不变，这种分歧可能会降低泛化性能。在这项工作中，我们描述了领域转移场景中出现的冲突梯度，并设计了基于梯度手术的新梯度协议策略来缓解其影响。我们用三个多域数据集验证了我们的方法在图像分类任务中的有效性，表明了所提出的一致性策略在提高域转移场景下深度学习模型的泛化能力方面的价值。

The popularization of intelligent devices including smartphones and surveillance cameras results in more serious privacy issues. De-identification is regarded as an effective tool for visual privacy protection with the process of concealing or replacing identity information. Most of the existing de-identification methods suffer from some limitations since they mainly focus on the protection process and are usually non-reversible. In this paper, we propose a personalized and invertible de-identification method based on the deep generative model, where the main idea is introducing a user-specific password and an adjustable parameter to control the direction and degree of identity variation. Extensive experiments demonstrate the effectiveness and generalization of our proposed framework for both face de-identification and recovery.

智能手机和监控摄像头等智能设备的普及导致了更严重的隐私问题。反身身份认证被认为是一种有效的视觉隐私保护工具，其过程是隐藏或替换身份信息。现有的大多数去识别方法都存在一些局限性，因为它们主要关注保护过程，并且通常是不可逆的。在本文中，我们提出了一种基于深层生成模型的个性化可逆去识别方法，其主要思想是引入用户专用密码和可调参数来控制身份变化的方向和程度。大量实验证明了我们提出的人脸去识别和恢复框架的有效性和通用性。

Have you ever looked at a painting and wondered what is the story behind it? This work presents a framework to bring art closer to people by generating comprehensive descriptions of fine-art paintings. Generating informative descriptions for artworks, however, is extremely challenging, as it requires to 1) describe multiple aspects of the image such as its style, content, or composition, and 2) provide background and contextual knowledge about the artist, their influences, or the historical period. To address these challenges, we introduce a multi-topic and knowledgeable art description framework, which modules the generated sentences according to three artistic topics and, additionally, enhances each description with external knowledge. The framework is validated through an exhaustive analysis, both quantitative and qualitative, as well as a comparative human evaluation, demonstrating outstanding results in terms of both topic diversity and information veracity.

你有没有看过一幅画，想知道它背后的故事是什么？这部作品提供了一个框架，通过生成对美术画的全面描述，使艺术更贴近人们。然而，为艺术作品生成信息性描述是非常具有挑战性的，因为它需要1) 描述图像的多个方面，例如其风格、内容或构图，以及2) 提供关于艺术家、其影响或历史时期的背景和背景知识。为了应对这些挑战，我们引入了一个多主题和知识丰富的艺术描述框架，该框架根据三个艺术主题对生成的句子进行模块化，并使用外部知识增强每个描述。该框架通过详尽的定量和定性分析以及人类比较评估得到验证，展示了在主题多样性和信息准确性方面的杰出成果。

During the last years, convolutional neural networks (CNNs) have triumphed over video quality assessment (VQA) tasks. However, CNN-based approaches heavily rely on annotated data which are typically not available in VQA, leading to the difficulty of model generalization. Recent advances in domain adaptation technique makes it possible to adapt models trained on source data to unlabeled target data. However, due to the distortion diversity and content variation of the collected videos, the intrinsic subjectivity of VQA tasks hampers the adaptation performance. In this work, we propose a curriculum-style unsupervised domain adaptation to handle the cross-domain no-reference VQA problem. The proposed approach could be divided into two stages. In the first stage, we conduct an adaptation between source and target domains to predict the rating distribution for target samples, which can better reveal the subjective nature of VQA. From this adaptation, we split the data in target domain into confident and uncertain subdomains using the proposed uncertainty-based ranking function, through measuring their prediction confidences. In the second stage, by regarding samples in confident subdomain as the easy tasks in the curriculum, a fine-level adaptation is conducted between two subdomains to fine-tune the prediction model. Extensive experimental results on benchmark datasets highlight the superiority of the proposed method over the competing methods in both accuracy and speed. The source code is released at <https://github.com/cpf0079/UCDA>.

在过去的几年中，卷积神经网络（CNN）已经战胜了视频质量评估（VQA）任务。然而，基于CNN的方法严重依赖于VQA中通常不可用的注释数据，这导致了模型泛化的困难。领域自适应技术的最新进展使得将源数据上训练的模型自适应到未标记的目标数据成为可能。然而，由于所收集视频的失真多样性和内容变化，VQA任务固有的主观性阻碍了自适应性能。在这项工作中，我们提出了一种课程风格的无监督领域适应来处理跨领域无参考VQA问题。拟议的办法可分为两个阶段。在第一阶段，我们在源域和目标域之间进行自适应，以预测目标样本的评级分布，这可以更好地揭示VQA的主观性质。通过这种自适应，我们使用所提出的基于不确定性的排序函数，通过测量目标域中的数据的预测置信度，将目标域中的数据划分为置信子域和不确定子域。在第二阶段，通过将自信子域中的样本作为课程中的简单任务，在两个子域之间进行精细调整，以微调预测模型。在基准数据集上的大量实验结果表明，该方法在准确性和速度上均优于其他方法。源代码发布于<https://github.com/cpf0079/UCDA>。

We present GTT-Net, a supervised learning framework for the reconstruction of sparse dynamic 3D geometry. We build on a graph-theoretic formulation of the generalized trajectory triangulation problem, where non-concurrent multi-view imaging geometry is known but global image sequencing is not provided. GTT-Net learns pairwise affinities modeling the spatio-temporal relationships among our input observations and leverages them to determine 3D geometry estimates. Experiments reconstructing 3D motion-capture sequences show GTT-Net outperforms the state of the art in terms of accuracy and robustness. Within the context of articulated motion reconstruction, our proposed architecture is 1) able to learn and enforce semantic 3D motion priors for shared training and test domains, while being 2) able to generalize its performance across different training and test domains. Moreover, GTT-Net provides a computationally streamlined framework for trajectory triangulation with applications to multi-instance reconstruction and event segmentation.

我们提出了GTT网络，一个用于重建稀疏动态三维几何体的监督学习框架。我们建立在广义轨迹三角部分问题的图论公式上，其中非并发多视图成像几何已知，但不提供全局图像排序。GTT网络学习成对的亲和力建模我们的输入观测值之间的时空关系，并利用它们来确定3D几何估计。重建三维运动捕获序列的实验表明，GTT网络在准确性和鲁棒性方面优于现有技术。在关节运动重建的背景下，我们提出的架构1)能够学习和实施共享训练和测试域的语义3D运动先验，同时2)能够在不同的训练和测试域中推广其性能。此外，GTT网络为轨迹三角部分提供了一个计算简化的框架，并应用于多实例重建和事件分割。

The fluctuation of the water surface causes refractive distortions that severely downgrade the image of an underwater scene. Here, we present the distortion-guided network (DG-Net) for restoring distortion-free underwater images. The key idea is to use a distortion map to guide network training. The distortion map models the pixel displacement caused by water refraction. We first use a physically constrained convolutional network to estimate the distortion map from the refracted image. We then use a generative adversarial network guided by the distortion map to restore the sharp distortion-free image. Since the distortion map indicates correspondences between the distorted image and the distortion-free one, it guides the network to make better predictions. We evaluate our network on several real and synthetic underwater image datasets and show that it outperforms the state-of-the-art algorithms, especially in presence of large distortions. We also show results of complex scenarios, including outdoor swimming pool images captured by the drone and indoor aquarium images taken by cellphone camera.

水面的起伏会导致折射失真，严重降低水下场景的图像质量。在这里，我们提出了用于恢复无失真水下图像的失真引导网络 (DG-Net)。其关键思想是使用失真图来指导网络训练。畸变贴图模拟由水折射引起的像素位移。我们首先使用一个物理约束卷积网络来估计折射图像的失真图。然后，我们使用一个由畸变图引导的生成性对抗网络来恢复清晰的无畸变图像。由于畸变图指示畸变图像和无畸变图像之间的对应关系，因此它引导网络做出更好的预测。我们在几个真实的和合成的水下图像数据集上评估了我们的网络，并表明它的性能超过了最先进的算法，特别是在存在大失真的情况下。我们还展示了复杂场景的结果，包括无人驾驶飞机拍摄的室外游泳池图像和手机摄像头拍摄的室内水族馆图像。

Neural Radiance Fields (NeRF) have recently gained a surge of interest within the computer vision community for its power to synthesize photorealistic novel views of real-world scenes. One limitation of NeRF, however, is its requirement of known camera poses to learn the scene representations. In this paper, we propose Bundle-Adjusting Neural Radiance Fields (BARF) for training NeRF from imperfect camera poses -- the joint problem of learning neural 3D representations and registering camera frames. We establish a theoretical connection to classical planar image registration and show that coarse-to-fine registration is also applicable to NeRF. Furthermore, we demonstrate mathematically that positional encoding has a direct impact on the basin of attraction for registration with a synthesis-based objective. Experiments on synthetic and real-world data show that BARF can effectively optimize the neural scene representations and resolve large camera pose misalignment at the same time. This enables applications of view synthesis and localization of video sequences from unknown camera poses, opening up new avenues for visual localization systems (e.g. SLAM) towards sequential registration with NeRF.

神经辐射场 (NeRF) 由于其合成真实世界场景的照片级真实感新视图的能力，最近在计算机视觉界引起了极大的兴趣。然而，NeRF的一个限制是它需要已知的相机姿势来学习场景表示。在这篇论文中，我们提出了束调整神经辐射场 (BARF) 来训练来自不完美相机姿势的NeRF——学习神经3D表示和注册相机帧的联合问题。我们建立了一个与经典平面图像配准的理论联系，并表明从粗到精的配准也适用于NeRF。此外，我们从数学上证明了位置编码对基于合成的目标的注册吸引盆有直接影响。对合成数据和真实数据的实验表明，BARF可以有效地优化神经场景表示，同时解决较大的相机姿态失调问题。这使得来自未知摄像机姿势的视频序列的视图合成和定位应用成为可能，为视觉定位系统（如SLAM）向NeRF顺序注册开辟了新的途径。

The key point of language-guided person search is to construct the cross-modal association between visual and textual input. Existing methods focus on designing multimodal attention mechanisms and novel cross-modal loss functions to learn such association implicitly. We propose a representation learning method for language-guided person search based on color reasoning (LapsCore). It can explicitly build a fine-grained cross-modal association bidirectionally. Specifically, a pair of dual sub-tasks, image colorization and text completion, is designed. In the former task, rich text information is learned to colorize gray images, and the latter one requests the model to understand the image and complete color word vacancies in the captions. The two sub-tasks enable models to learn correct alignments between text phrases and image regions, so that rich multimodal representations can be learned. Extensive experiments on multiple datasets demonstrate the effectiveness and superiority of the proposed method.

语言引导的人称搜索的关键是在视觉输入和文本输入之间建立跨模态的关联。现有的方法侧重于设计多模态注意机制和新的跨模态损失函数来隐式地学习这种关联。我们提出了一种基于颜色推理 (LapsCore) 的语言引导的人物搜索表征学习方法。它可以显式地双向构建细粒度跨模态关联。具体来说，设计了一对双重子任务：图像着色和文本完成。在前一个任务中，通过学习富文本信息对灰度图像进行着色，而后一个任务则要求模型理解图像并完成标题中的颜色词空缺。这两个子任务使模型能够学习文本短语和图像区域之间的正确对齐，从而可以学习丰富的多模态表示。在多个数据集上的大量实验证明了该方法的有效性和优越性。

Existing deep methods produce highly accurate 3D reconstructions in stereo and multiview stereo settings, i.e., when cameras are both internally and externally calibrated. Nevertheless, the challenge of simultaneous recovery of camera poses and 3D scene structure in multiview settings with deep networks is still outstanding. Inspired by projective factorization for Structure from Motion (SfM) and by deep matrix completion techniques, we propose a neural network architecture that, given a set of point tracks in multiple images of a static scene, recovers both the camera parameters and a (sparse) scene structure by minimizing an unsupervised reprojection loss. Our network architecture is designed to respect the structure of the problem: the sought output is equivariant to permutations of both cameras and scene points. Notably, our method does not require initialization of camera parameters or 3D point locations. We test our architecture in two setups: (1) single scene reconstruction and (2) learning from multiple scenes. Our experiments, conducted on a variety of datasets in both internally calibrated and uncalibrated settings, indicate that our method accurately recovers pose and structure, on par with classical state of the art methods. Additionally, we show that a pre-trained network can be used to reconstruct novel scenes using inexpensive fine-tuning with no loss of accuracy.

现有的deep方法在立体和多视图立体设置中产生高精度的三维重建，即当摄像机进行内部和外部校准时。然而，在具有深度网络的多视图环境中，同时恢复相机姿势和三维场景结构的挑战仍然很突出。受运动结构投影分解（SfM）和深度矩阵完成技术的启发，我们提出了一种神经网络结构，该结构在给定静态场景多幅图像中的一组点轨迹的情况下，通过最小化无监督重投影损失来恢复相机参数和（稀疏）场景结构。我们的网络架构是为了尊重问题的结构而设计的：所寻求的输出与摄像机和场景点的排列相同。值得注意的是，我们的方法不需要初始化相机参数或3D点位置。我们在两个设置中测试我们的架构：（1）单场景重建和（2）从多个场景学习。我们的实验，进行了各种数据集在内部校准和未校准设置，表明我们的方法准确地恢复姿势和结构，与经典的最先进的方法相媲美。此外，我们还表明，预先训练好的网络可以在不损失精度的情况下使用廉价的微调来重建新场景。

Unsupervised representation learning has achieved outstanding performances using centralized data available on the Internet. However, the increasing awareness of privacy protection limits sharing of decentralized unlabeled image data that grows explosively in multiple parties (e.g. mobile phones and cameras). As such, a natural problem is how to leverage these data to learn visual representations for downstream tasks while preserving data privacy. To address this problem, we propose a novel federated unsupervised learning framework, FedU. In this framework, each party trains models from unlabeled data independently using contrastive learning with an online network and a target network. Then, a central server aggregates trained models and updates clients' models with the aggregated global model. It preserves data privacy as each party only has access to its raw data. Decentralized data among multiple parties is normally non-independent and identically distributed (non-IID), which leads to performance degradation. To tackle this challenge, we propose two simple but effective methods: (1) we design the communication protocol to upload only the encoders of online networks for server aggregation and update them with the aggregated encoder. (2) we introduce a new module to dynamically decide how to update the predictors based on the degree of divergence caused by non-IID. The predictor is the other component of the online network. Extensive experiments and ablations demonstrate the effectiveness and significance of FedU. It outperforms training with only one party by over 5% and other methods by over 14% in linear and semi-supervised evaluation on non-IID data.

利用互联网上的集中数据，无监督表征学习取得了优异的成绩。然而，隐私保护意识的提高限制了分散的未标记图像数据的共享，这些数据在多方（如手机和相机）中爆炸性增长。因此，一个自然的问题是如何利用这些数据来学习下游任务的可视化表示，同时保护数据隐私。为了解决这个问题，我们提出了一个新的联邦无监督学习框架FedU。在这个框架中，各方使用在线网络和目标网络的对比学习，独立地从未标记的数据中训练模型。然后，中央服务器聚合经过训练的模型，并使用聚合的全局模型更新客户机的模型。它保护了数据隐私，因为各方只能访问其原始数据。多方之间的分散数据通常是非独立的、相同分布的（非IID），这会导致性能下降。为了应对这一挑战，我们提出了两种简单但有效的方法：

(1) 设计通信协议，只上传在线网络的编码器进行服务器聚合，并使用聚合编码器进行更新。(2) 我们引入了一个新的模块，根据非IID引起的发散程度动态决定如何更新预测器。预测器是在线网络的另一个组成部分。大量的实验和烧蚀证明了FedU的有效性和重要性。在非IID数据的线性和半监督评估中，它的表现优于仅使用一方的培训，超过5%，其他方法超过14%。

We consider the problem of online and real-time registration of partial point clouds obtained from an unseen real-world rigid object without knowing its 3D model. The point cloud is partial as it is obtained by a depth sensor capturing only the visible part of the object from a certain viewpoint. It introduces two main challenges: 1) two partial point clouds do not fully overlap and 2) keypoints tend to be less reliable when the visible part of the object does not have salient local structures. To address these issues, we propose DeepPRO, a keypoint-free and an end-to-end trainable deep neural network. Its core idea is inspired by how humans align two point clouds: we can imagine how two point clouds will look like after the registration based on their shape. To realize the idea, DeepPRO has inputs of two partial point clouds and directly predicts the point-wise location of the aligned point cloud. By preserving the ordering of points during the prediction, we enjoy dense correspondences between input and predicted point clouds when inferring rigid transform parameters. We conduct extensive experiments on the real-world Linemod and synthetic ModelNet40 datasets. In addition, we collect and evaluate on the PRO1k dataset, a large-scale version of Linemod meant to test generalization to real-world scans. Results show that DeepPRO achieves the best accuracy against thirteen strong baseline methods, e.g., 2.2mm ADD on the Linemod dataset, while running 50 fps on mobile devices.

我们考虑的问题，在线和实时注册的部分点云从一个看不见的现实世界的刚性对象，而不知道其三维模型。点云是局部的，因为它是由深度传感器从某个视点仅捕获对象的可见部分获得的。它带来了两个主要挑战：1) 两个局部点云没有完全重叠；2) 当对象的可见部分没有显著的局部结构时，关键点往往不太可靠。为了解决这些问题，我们提出了DeepPRO，一种无关键点、端到端可训练的deep神经网络。它的核心思想是受人类如何对齐两点云的启发：我们可以想象两点云在根据形状进行注册后的样子。为了实现这个想法，DeepPRO有两个部分点云的输入，并直接预测对齐点云的逐点位置。通过在预测过程中保持点的顺序，我们可以在推断刚性变换参数时享受到输入和预测点云之间的紧密对应。我们在真实世界的Linemod和合成ModelNet40数据集上进行了广泛的实验。此外，我们收集并评估PRO1k数据集，这是一个大型版本的Linemod，旨在测试对真实世界扫描的概括。结果表明，与13种强基线方法相比，DeepPRO达到了最佳精度，例如，在Linemod数据集上添加2.2mm，同时在移动设备上运行50fps。

Deep networks allow to obtain outstanding results in semantic segmentation, however they need to be trained in a single shot with a large amount of data. Continual learning settings where new classes are learned in incremental steps and previous training data is no longer available are challenging due to the catastrophic forgetting phenomenon. Existing approaches typically fail when several incremental steps are performed or in presence of a distribution shift of the background class. We tackle these issues by recreating no longer available data for the old classes and outlining a content inpainting scheme on the background class. We propose two sources for replay data. The first resorts to a generative adversarial network to sample from the class space of past learning steps. The second relies on web-crawled data to retrieve images containing examples of old classes from online databases. In both scenarios no samples of past steps are stored, thus avoiding privacy concerns. Replay data are then blended with new samples during the incremental steps. Our approach, RECALL, outperforms state-of-the-art methods.

深度网络允许在语义分割中获得出色的结果，但是它们需要使用大量数据在单个镜头中进行训练。由于灾难性遗忘现象的存在，在持续学习环境中，新课程以渐进的方式学习，而以前的培训数据不再可用，这是一种挑战。当执行几个增量步骤或存在后台类的分布偏移时，现有方法通常会失败。我们通过为旧类重新创建不再可用的数据并在后台类上概述内容修复方案来解决这些问题。我们提出了两个重播数据源。第一种方法借助于生成性对抗网络，从过去学习步骤的课堂空间中取样。第二种方法依靠网络爬网数据从在线数据库中检索包含旧类示例的图像。在这两种情况下，不会存储过去步骤的样本，因此避免了隐私问题。然后，在增量步骤中将重播数据与新样本混合。回想一下，我们的方法优于最先进的方法。

While most neural video codecs address P-frame coding (predicting each frame from past ones), in this paper we address B-frame compression (predicting frames using both past and future reference frames). Our B-frame solution is based on the existing P-frame methods. As a result, B-frame coding capability can easily be added to an existing neural codec. The basic idea of our B-frame coding method is to interpolate the two reference frames to generate a single reference frame and then use it together with an existing P-frame codec to encode the input B-frame. Our studies show that the interpolated frame is a much better reference for the P-frame codec compared to using the previous frame as is usually done. Our results show that using the proposed method with an existing P-frame codec can lead to 28.5% saving in bit-rate on the UVG dataset compared to the P-frame codec while generating the same video quality.

虽然大多数神经视频编解码器处理P帧编码（从过去的帧预测每个帧），但在本文中，我们处理B帧压缩（使用过去和未来参考帧预测帧）。我们的B-frame解决方案基于现有的P-frame方法。因此，B帧编码能力可以很容易地添加到现有的神经编解码器中。我们的B帧编码方法的基本思想是对两个参考帧进行插值以生成单个参考帧，然后将其与现有的P帧编解码器一起使用以对输入的B帧进行编码。我们的研究表明，与通常使用前一帧相比，插值帧是P帧编解码器更好的参考。我们的结果表明，与P帧编解码器相比，在生成相同视频质量的同时，将所提出的方法与现有的P帧编解码器结合使用，可以在UVG数据集上节省28.5%的比特率。

Relational reasoning is at the heart of video question answering. However, existing approaches suffer from several common limitations: (1) they only focus on either object-level or frame-level relational reasoning, and fail to integrate the both; and (2) they neglect to leverage semantic knowledge for relational reasoning. In this work, we propose a Hierarchical VisuAl-Semantic Relational Reasoning (HAIR) framework to address these limitations. Specifically, we present a novel graph memory mechanism to perform relational reasoning, and further develop two types of graph memory: a) visual graph memory that leverages visual information of video for relational reasoning; b) semantic graph memory that is specifically designed to explicitly leverage semantic knowledge contained in the classes and attributes of video objects, and perform relational reasoning in the semantic space. Taking advantage of both graph memory mechanisms, we build a hierarchical framework to enable visual-semantic relational reasoning from object level to frame level. Experiments on four challenging benchmark datasets show that the proposed framework leads to state-of-the-art performance, with fewer parameters and faster inference speed. Besides, our approach also shows superior performance on other video+language task.

关系推理是视频问答的核心。然而，现有的方法有几个共同的局限性：（1）它们只关注对象级或框架级的关系推理，不能将两者结合起来；（2）他们忽略了利用语义知识进行关系推理。在这项工作中，我们提出了一个分层视觉语义关系推理（HAIR）框架来解决这些限制。具体来说，我们提出了一种新的图形记忆机制来执行关系推理，并进一步开发了两种类型的图形记忆：a）利用视频的视觉信息进行关系推理的视觉图形记忆；b）语义图内存，专门设计用于显式利用视频对象的类和属性中包含的语义知识，并在语义空间中执行关系推理。利用这两种图形存储机制，我们构建了一个分层框架，以支持从对象层到框架层的可视化语义关系推理。在四个具有挑战性的基准数据集上的实验表明，该框架具有最先进的性能，参数更少，推理速度更快。此外，我们的方法在其他视频+语言任务上也表现出优异的性能。

We present Voxel Transformer (VoTr), a novel and effective voxel-based Transformer backbone for 3D object detection from point clouds. Conventional 3D convolutional backbones in voxel-based 3D detectors cannot efficiently capture large context information, which is crucial for object recognition and localization, owing to the limited receptive fields. In this paper, we resolve the problem by introducing a Transformer-based architecture that enables long-range relationships between voxels by self-attention. Given the fact that non-empty voxels are naturally sparse but numerous, directly applying standard Transformer on voxels is non-trivial. To this end, we propose the sparse voxel module and the submanifold voxel module, which can operate on the empty and non-empty voxel positions effectively. To further enlarge the attention range while maintaining comparable computational overhead to the convolutional counterparts, we propose two attention mechanisms for multi-head attention in those two modules: Local Attention and Dilated Attention, and we further propose Fast Voxel Query to accelerate the querying process in multi-head attention. VoTr contains a series of sparse and submanifold voxel modules and can be applied in most voxel-based detectors. Our proposed VoTr shows consistent improvement over the convolutional baselines while maintaining computational efficiency on the KITTI dataset and the Waymo Open dataset.

我们提出了体素变换器（VoTr），一种新颖有效的基于体素的变换器主干，用于从点云中检测三维对象。在基于体素的三维检测器中，传统的三维卷积主干不能有效地捕获大的背景信息，这对于目标识别和定位是至关重要的，因为接收场有限。在本文中，我们通过引入基于转换器的体系结构来解决这个问题，该体系结构通过自我关注实现体素之间的远程关系。考虑到非空体素自然稀疏但数量众多的事实，直接在体素上应用标准变换器是非常重要的。为此，我们提出了稀疏体素模块和子流形体素模块，可以有效地对空体素和非空体素位置进行操作。为了进一步扩大注意范围，同时保持与卷积注意相当的计算开销，我们在这两个模块中提出了两种多头部注意的注意机制：局部注意和扩展注意，并进一步提出了快速体素查询来加速多头部注意的查询过程。VoTr包含一系列稀疏和子流形体素模块，可应用于大多数

基于体素的检测器。我们提出的VoTr在保持KITTI数据集和Waymo开放数据集的计算效率的同时，显示了对卷积基线的一致改进。

Warping-based video stabilizers smooth camera trajectory by constraining each pixel's displacement and warp stabilized frames from unstable ones accordingly. However, since the view outside the boundary is not available during warping, the resulting holes around the boundary of the stabilized frame must be discarded (i.e., cropping) to maintain visual consistency, and thus does leads to a tradeoff between stability and cropping ratio. In this paper, we make a first attempt to address this issue by proposing a new Out-of-boundary View Synthesis (OVS) method. By the nature of spatial coherence between adjacent frames and within each frame, OVS extrapolates the out-of-boundary view by aligning adjacent frames to each reference one. Technically, it first calculates the optical flow and propagates it to the outer boundary region according to the affinity, and then warps pixels accordingly. OVS can be integrated into existing warping-based stabilizers as a plug-and-play pre-processing module to significantly improve the cropping ratio of the stabilized results. In addition, stability is improved because the jitter amplification effect caused by cropping and resizing is reduced. Experimental results on the NUS benchmark show that OVS can improve the performance of five representative state-of-the-art methods in terms of objective metrics and subjective visual quality.

基于扭曲的视频稳定器通过约束每个像素的位移来平滑相机轨迹，并相应地从不稳定帧扭曲稳定帧。但是，由于在翘曲期间边界外的视图不可用，因此必须丢弃稳定框架边界周围产生的孔（即裁剪），以保持视觉一致性，从而在稳定性和裁剪率之间进行权衡。在本文中，我们首次尝试通过提出一种新的边界外视图综合（OVS）方法来解决这个问题。根据相邻帧之间和每个帧内的空间相关性，OVS通过将相邻帧与每个参考帧对齐来推断出边界外视图。从技术上讲，它首先计算光流并根据亲和性将其传播到外边界区域，然后相应地扭曲像素。OVS可以作为即插即用预处理模块集成到现有的基于翘曲的稳定剂中，以显著提高稳定结果的裁剪率。此外，由于减少了裁剪和调整大小引起的抖动放大效应，因此稳定性得到了提高。在NUS基准上的实验结果表明，OVS可以提高五种具有代表性的最新方法在客观度量和主观视觉质量方面的性能。

Multimodal self-supervised learning is getting more and more attention as it allows not only to train large networks without human supervision but also to search and retrieve data across various modalities. In this context, this paper proposes a framework that, starting from a pre-trained backbone, learns a common multimodal embedding space that, in addition to sharing representations across different modalities, enforces a grouping of semantically similar instances. To this end, we extend the concept of instance-level contrastive learning with a multimodal clustering step in the training pipeline to capture semantic similarities across modalities. The resulting embedding space enables retrieval of samples across all modalities, even from unseen datasets and different domains. To evaluate our approach, we train our model on the HowTo100M dataset and evaluate its zero-shot retrieval capabilities in two challenging domains, namely text-to-video retrieval, and temporal action localization, showing state-of-the-art results on four different datasets.

多模态自监督学习不仅可以在无需人工监督的情况下训练大型网络，而且可以在各种模式下搜索和检索数据，因此受到越来越多的关注。在此背景下，本文提出了一个框架，该框架从预先训练的主干开始，学习一个公共的多模态嵌入空间，该空间除了在不同模式之间共享表示之外，还强制执行一组语义相似的实例。为此，我们扩展了实例级对比学习的概念，在训练管道中采用多模态聚类步骤，以捕获不同模态之间的语义相似性。由此产生的嵌入空间可以跨所有模式检索样本，甚至可以从看不见的数据集和不同的域检索样本。为了评估我们的方法，我们在HowTo100M数据集上训练我们的模型，并评估其在两

个具有挑战性的领域中的零镜头检索能力，即文本到视频检索和时间动作定位，在四个不同的数据集上显示最先进的结果。

This paper proposes to handle the practical problem of learning a universal model for crowd counting across scenes and datasets. We dissect that the crux of this problem is the catastrophic sensitivity of crowd counters to scale shift, which is very common in the real world and caused by factors such as different scene layouts and image resolutions. Therefore it is difficult to train a universal model that can be applied to various scenes. To address this problem, we propose scale alignment as a prime module for establishing a novel crowd counting framework. We derive a closed-form solution to get the optimal image rescaling factors for alignment by minimizing the distances between their scale distributions. A novel neural network together with a loss function based on an efficient sliced Wasserstein distance is also proposed for scale distribution estimation. Benefiting from the proposed method, we have learned a universal model that generally works well on several datasets where can even outperform state-of-the-art models that are particularly fine-tuned for each dataset significantly. Experiments also demonstrate the much better generalizability of our model to unseen scenes.

本文提出处理学习跨场景和数据集人群计数通用模型的实际问题。我们剖析了这个问题的症结是人群计数器对尺度变化的灾难性敏感性，这在现实世界中非常常见，并且是由不同的场景布局和图像分辨率等因素引起的。因此，很难训练出适用于各种场景的通用模型。为了解决这个问题，我们提出了规模对齐作为建立一个新的人群计数框架的主要模块。我们推导了一个封闭形式的解决方案，通过最小化它们的尺度分布之间的距离来获得最佳的图像对齐重缩放因子。本文还提出了一种新的神经网络和基于有效切片Wasserstein距离的损失函数，用于尺度分布估计。得益于所提出的方法，我们学习了一个通用模型，该模型通常在多个数据集上运行良好，甚至可以优于对每个数据集进行了特别微调的最新模型。实验还表明，我们的模型对于看不见的场景具有更好的通用性。

Learning RAW-to-sRGB mapping has drawn increasing attention in recent years, wherein an input raw image is trained to imitate the target sRGB image captured by another camera. However, the severe color inconsistency makes it very challenging to generate well-aligned training pairs of input raw and target sRGB images. While learning with inaccurately aligned supervision is prone to causing pixel shift and producing blurry results. In this paper, we circumvent such issue by presenting a joint learning model for image alignment and RAW-to-sRGB mapping. To diminish the effect of color inconsistency in image alignment, we introduce to use a global color mapping (GCM) module to generate an initial sRGB image given the input raw image, which can keep the spatial location of the pixels unchanged, and the target sRGB image is utilized to guide GCM for converting the color towards it. Then a pre-trained optical flow estimation network (e.g., PWC-Net) is deployed to warp the target sRGB image to align with the GCM output. To alleviate the effect of inaccurately aligned supervision, the warped target sRGB image is leveraged to learn RAW-to-sRGB mapping. When training is done, the GCM module and optical flow network can be detached, thereby bringing no extra computation cost for inference. Experiments show that our method performs favorably against state-of-the-arts on ZRR and SR-RAW datasets. With our joint learning model, a light-weight backbone can achieve better quantitative and qualitative performance on ZRR dataset. Codes are available at <https://github.com/cszhilu1998/RAW-to-sRGB>.

近年来，学习RAW到sRGB映射引起了越来越多的关注，其中输入的RAW图像被训练成模仿由另一个相机捕获的目标sRGB图像。然而，严重的颜色不一致性使得生成输入原始图像和目标sRGB图像的对齐训练对非常具有挑战性。而使用不准确对齐的监控进行学习容易导致像素偏移并产生模糊结果。在本文中，我们提出了一种用于图像对齐和RAW到sRGB映射的联合学习模型，从而避免了这一问题。为了减少图像对齐中颜色不一致的影响，我们引入了一个全局颜色映射（GCM）模块，在给定输入原始图像的

情况下生成初始sRGB图像，该模块可以保持像素的空间位置不变，并利用目标sRGB图像引导GCM向其转换颜色。然后部署预先训练的光流估计网络（例如，PWC网络），以扭曲目标sRGB图像，使其与GCM输出对齐。为了减轻不精确对齐监控的影响，利用扭曲的目标sRGB图像来学习RAW到sRGB的映射。当完成训练时，GCM模块和光流网络可以分离，从而不会为推断带来额外的计算成本。实验表明，我们的方法在ZRR和SR-Raw数据集上表现良好。通过我们的联合学习模型，轻量级主干可以在ZRR数据集上实现更好的定量和定性性能。代码可在<https://github.com/cszhilu1998/RAW-to-sRGB>。

Deep convolutional neural networks (CNNs) are achieving great successes for image super-resolution (SR), where global context is crucial for accurate restoration. However, the basic convolutional layer in CNNs is designed to extract local patterns, lacking the ability to model global context. Many efforts have been devoted to augmenting SR networks with the global context information, especially by global feature interaction methods. These works incorporate the global context into local feature representation. However, recent advances in neuroscience show that it is necessary for the neurons to dynamically modulate their functions according to context, which is neglected in most CNN based SR methods. Motivated by those observations and analyses, we propose context reasoning attention network (CRAN) to adaptively modulate the convolution kernel according to the global context. Specifically, we extract global context descriptors, which are further enhanced with semantic reasoning. Channel and spatial interactions are then proposed to generate context reasoning attention mask, which is applied to modify the convolution kernel adaptively. Such a modulated convolution layer is utilized as basic component to build the network blocks and itself. Extensive experiments on benchmark datasets with multiple degradation models show that our CRAN achieves superior SR results and favourable efficiency trade-off.

深度卷积神经网络 (CNN) 在图像超分辨率 (SR) 方面取得了巨大成功，其中全局环境对于精确恢复至关重要。然而，CNN中的基本卷积层设计用于提取局部模式，缺乏对全局上下文建模的能力。人们致力于用全局上下文信息扩充SR网络，特别是通过全局特征交互方法。这些工作将全局上下文合并到局部特征表示中。然而，神经科学的最新进展表明，神经元有必要根据上下文动态调节其功能，这在大多数基于CNN的SR方法中被忽略。基于这些观察和分析，我们提出了上下文推理注意网络 (CRAN)，根据全局上下文自适应调整卷积核。具体来说，我们提取了全局上下文描述符，并通过语义推理进一步增强了这些描述符。然后，提出了通道和空间交互来生成上下文推理注意掩码，用于自适应地修改卷积核。这种调制卷积层被用作构建网络块及其自身的基本组件。在具有多种退化模型的基准数据集上进行的大量实验表明，我们的CRAN实现了优异的SR结果和有利的效率权衡。

We present 3DeepCT, a deep neural network for computed tomography, which performs 3D reconstruction of scattering volumes from multi-view images. The architecture is dictated by the stationary nature of atmospheric cloud fields. The task of volumetric scattering tomography aims at recovering a volume from its 2D projections. This problem has been approached by diverse inverse methods based on signal processing and physics models. However, such techniques are typically iterative, exhibiting a high computational load and a long convergence time. We show that 3DeepCT outperforms physics-based inverse scattering methods, in accuracy, as well as offering orders of magnitude improvement in computational run-time. We further introduce a hybrid model that combines 3DeepCT and physics-based analysis. The resultant hybrid technique enjoys fast inference time and improved recovery performance.

我们提出3DeepCT，一种用于计算机断层扫描的深度神经网络，它可以从多视图图像中对散射体进行三维重建。这种结构是由大气云场的静止性质决定的。体积散射层析成像的任务是从二维投影中恢复体积。基于信号处理和物理模型的各种反演方法已经解决了这个问题。然而，此类技术通常是迭代的，表现出高计算量和长收敛时间。我们表明3DeepCT在精度上优于基于物理的逆散射方法，并且在计算运行

时间上提供了数量级的改进。我们进一步介绍了一种结合3DeepCT和基于物理的分析的混合模型。由此产生的混合技术具有快速的推理时间和改进的恢复性能。

An event camera detects the scene radiance changes and sends a sequence of asynchronous event streams with high dynamic range, high temporal resolution, and low latency. However, the spatial resolution of event cameras is limited as a trade-off for these outstanding properties. To reconstruct high-resolution intensity images from event data, we propose EvIntSR-Net that converts event data to multiple latent intensity frames to achieve super-resolution on intensity images in this paper. EvIntSR-Net bridges the domain gap between event streams and intensity frames and learns to merge a sequence of latent intensity frames in a recurrent updating manner. Experimental results show that EvIntSR-Net can reconstruct SR intensity images with higher dynamic range and fewer blurry artifacts by fusing events with intensity frames for both simulated and real-world data. Furthermore, the proposed EvIntSR-Net is able to generate high-frame-rate videos with super-resolved frames.

事件摄影机检测场景辐射变化，并发送一系列具有高动态范围、高时间分辨率和低延迟的异步事件流。然而，事件摄影机的空间分辨率是有限的，作为这些突出特性的权衡。为了从事件数据中重建高分辨率的强度图像，本文提出了将事件数据转换为多个潜在强度帧的EvIntSR网络，以实现强度图像的超分辨率。EvIntSR网络在事件流和强度帧之间架起了桥梁，并学习以循环更新方式合并潜在强度帧序列。实验结果表明，EvIntSR网络通过融合模拟和真实数据中的事件和强度帧，可以重建具有更高动态范围和更少模糊伪影的SR强度图像。此外，提出的EvIntSR网络能够生成具有超分辨率帧的高帧速率视频。

Effectively structuring deep knowledge plays a pivotal role in transfer from teacher to student, especially in semantic vision tasks. In this paper, we present a simple knowledge structure to exploit and encode information inside the detection system to facilitate detector knowledge distillation. Specifically, aiming at solving the feature imbalance problem while further excavating the missing relation inside semantic instances, we design a graph whose nodes correspond to instance proposal-level features and edges represent the relation between nodes. To further refine this graph, we design an adaptive background loss weight to reduce node noise and background samples mining to prune trivial edges. We transfer the entire graph as encoded knowledge representation from teacher to student, capturing local and global information simultaneously. We achieve new state-of-the-art results on the challenging COCO object detection task with diverse student-teacher pairs on both one- and two-stage detectors. We also experiment with instance segmentation to demonstrate robustness of our method. It is notable that distilled Faster R-CNN with ResNet18-FPN and ResNet50-FPN yields 38.68 and 41.82 Box AP respectively on the COCO benchmark, Faster R-CNN with ResNet101-FPN significantly achieves 43.38 AP, which outperforms ResNet152-FPN teacher about 0.7 AP. Code: <https://github.com/dvlab-research/Dsig>.

有效构建深层知识在教师向学生的迁移中起着关键作用，尤其是在语义视觉任务中。在本文中，我们提出了一个简单的知识结构来开发和编码检测系统中的信息，以促进检测器知识提取。具体来说，为了解决特征不平衡问题，同时进一步挖掘语义实例中的缺失关系，我们设计了一个图，其节点对应于实例建议级特征，边代表节点之间的关系。为了进一步细化该图，我们设计了一个自适应背景损失权重来减少节点噪声，并通过背景样本挖掘来修剪平凡的边缘。我们将整个图形作为编码的知识表示从教师传递给学生，同时捕获局部和全局信息。我们在具有挑战性的COCO目标检测任务上取得了新的最新成果，在一级和两级探测器上都有不同的师生对。我们还通过实例分割实验证了该方法的鲁棒性。值得注意的是，在COCO基准上，使用ResNet18 FPN和ResNet50 FPN提取的更快的R-CNN分别产生38.68和41.82箱AP，使用ResNet101 FPN的更快的R-CNN显著达到43.38 AP，优于ResNet152 FPN约0.7 AP。代码：<https://github.com/dvlab-research/Dsig>。

We find that images contain intrinsic structure that enables the reversal of many adversarial attacks. Attack vectors cause not only image classifiers to fail, but also collaterally disrupt incidental structure in the image. We demonstrate that modifying the attacked image to restore the natural structure will reverse many types of attacks, providing a defense. Experiments demonstrate significantly improved robustness for several state-of-the-art models across the CIFAR-10, CIFAR-100, SVHN, and ImageNet datasets. Our results show that our defense is still effective even if the attacker is aware of the defense mechanism. Since our defense is deployed during inference instead of training, it is compatible with pre-trained networks as well as most other defenses. Our results suggest deep networks are vulnerable to adversarial examples partly because their representations do not enforce the natural structure of images.

我们发现图像包含能够逆转许多敌对攻击的内在结构。攻击向量不仅会导致图像分类器失败，还会间接破坏图像中的附带结构。我们证明，修改受攻击图像以恢复自然结构将逆转多种类型的攻击，从而提供防御。实验表明，跨CIFAR-10、CIFAR-100、SVHN和ImageNet数据集的多个最先进模型的鲁棒性显著提高。我们的结果表明，即使攻击者知道防御机制，我们的防御仍然有效。由于我们的防御是在推理而非训练期间部署的，因此它与预先训练的网络以及大多数其他防御兼容。我们的研究结果表明，深层网络很容易受到敌对例子的攻击，部分原因是它们的表现没有强化图像的自然结构。

Deep learning has made tremendous success in computer vision, natural language processing and even visual-semantic learning, which requires a huge amount of labeled training data. Nevertheless, the goal of human-level intelligence is to enable a model to quickly obtain an in-depth understanding given a small number of samples, especially with heterogeneity in the multi-modal scenarios such as visual question answering and image captioning. In this paper, we study the few-shot visual-semantic learning and present the Hierarchical Graph ATTention network (HGAT). This two-stage network models the intra- and inter-modal relationships with limited image-text samples. The main contributions of HGAT can be summarized as follows: 1) it sheds light on tackling few-shot multi-modal learning problems, which focuses primarily, but not exclusively on visual and semantic modalities, through better exploitation of the intra-relationship of each modality and an attention-based co-learning framework between modalities using a hierarchical graph-based architecture; 2) it achieves superior performance on both visual question answering and image captioning in the few-shot setting; 3) it can be easily extended to the semi-supervised setting where image-text samples are partially unlabeled. We show via extensive experiments that HGAT delivers state-of-the-art performance on three widely-used benchmarks of two visual-semantic learning tasks.

深度学习在计算机视觉、自然语言处理甚至视觉语义学习方面取得了巨大的成功，而视觉语义学习需要大量的标记训练数据。然而，人类水平智能的目标是使模型能够在给定少量样本的情况下快速获得深入理解，特别是在多模态场景（如视觉问答和图像字幕）中的异质性。本文研究了少镜头视觉语义学习，提出了层次图注意网络（HGAT）。这两个阶段的网络模型内和模态间的关系与有限的图像文本样本。HGAT的主要贡献可概括如下：1）它揭示了如何解决少数镜头多模式学习问题，主要关注但不完全关注视觉和语义模式，通过使用基于层次图的架构，更好地利用每个模态的内部关系和模态之间基于注意的共同学习框架；2）在少数镜头设置下，它在视觉问答和图像字幕方面都取得了优异的性能；3）它可以很容易地扩展到图像文本样本部分未标记的半监督设置。我们通过大量的实验表明，HGAT在两个视觉语义学习任务的三个广泛使用的基准上提供了最先进的性能。

In this paper, we tackle data-driven 3D point cloud registration. Given point correspondences, the standard Kabsch algorithm provides an optimal rotation estimate. This allows to train registration models in an end-to-end manner by differentiating the SVD operation. However, given the initial rotation estimate supplied by Kabsch, we show we can improve point correspondence learning during model training by extending the original optimization problem. In particular, we linearize the governing constraints of the rotation matrix and solve the resulting linear system of equations. We then iteratively produce new solutions by updating the initial estimate. Our experiments show that, by plugging our differentiable layer to existing learning-based registration methods, we improve the correspondence matching quality. This yields up to a 7% decrease in rotation error for correspondence-based data-driven registration methods.

在本文中，我们研究数据驱动的三维点云配准。给定点对应，标准Kabsch算法提供最佳旋转估计。这允许通过区分SVD操作，以端到端的方式培训注册模型。然而，考虑到Kabsch提供的初始旋转估计，我们可以通过扩展原始优化问题来改进模型训练期间的点对应学习。特别是，我们将旋转矩阵的控制约束线性化，并求解得到的线性方程组。然后，我们通过更新初始估计迭代生成新的解决方案。我们的实验表明，通过将我们的可微层插入现有的基于学习的配准方法，我们提高了对应匹配的质量。这使得基于对应关系的数据驱动配准方法的旋转误差降低了7%。

Image-level weakly supervised semantic segmentation is a challenging task. As classification networks tend to capture notable object features and are insensitive to overactivation, class activation map (CAM) is too sparse and rough to guide segmentation network training. Inspired by the fact that erasing distinguishing features force networks to collect new ones from non-discriminative object regions, we use relationships between CAMs to propose a novel weakly supervised method. In this work, we apply these features, learned from erased images, as segmentation supervision, driving network to study robust representation. In specifically, object regions obtained by CAM techniques are erased on images firstly. To provide other regions with segmentation supervision, Erased CAM Supervision Net (ECSNet) generates pixel-level labels by predicting segmentation results of those processed images. We also design the rule of suppressing noise to select reliable labels. Our experiments on PASCAL VOC 2012 dataset show that without data annotations except for ground truth image-level labels, our ECS-Net achieves 67.6% mIoU on test set and 66.6% mIoU on val set, outperforming previous state-of-the-art methods.

图像级弱监督语义分割是一项具有挑战性的任务。由于分类网络往往捕捉显著的对象特征，并且对过度激活不敏感，因此类激活图（CAM）过于稀疏和粗糙，无法指导分割网络训练。受消除区分特征迫使网络从非区分对象区域收集新特征这一事实的启发，我们利用CAM之间的关系提出了一种新的弱监督方法。在这项工作中，我们应用这些特征，从擦除的图像中学习，作为分割监控，驱动网络来研究鲁棒表示。具体来说，首先在图像上擦除CAM技术获得的目标区域。为了向其他区域提供分割监控，擦除CAM监控网（ECSNet）通过预测处理后图像的分割结果来生成像素级标签。我们还设计了抑制噪声的规则来选择可靠的标签。我们在PASCAL VOC 2012数据集上的实验表明，除了地面真实图像级别标签外，没有数据注释，我们的ECS网络在测试集上实现了67.6%的mIoU，在val集上实现了66.6%的mIoU，优于以前最先进的方法。

The crucial problem in vehicle re-identification is to find the same vehicle identity when reviewing this object from cross-view cameras, which sets a higher demand for learning viewpoint-invariant representations. In this paper, we propose to solve this problem from two aspects: constructing robust feature representations and proposing camera-sensitive evaluations. We first propose a novel Heterogeneous Relational Complement Network (HRCN) by incorporating region-specific features and cross-level features as complements for the original high-level output. Considering the distributional differences and semantic misalignment, we propose graph-based relation modules to embed these heterogeneous features into one unified high-dimensional space. On the other hand, considering the deficiencies of cross-camera evaluations in existing measures (i.e., CMC and AP), we then propose a Cross-camera Generalization Measure (CGM) to improve the evaluations by introducing position-sensitivity and cross-camera generalization penalties. We further construct a new benchmark of existing models with our proposed CGM and experimental results reveal that our proposed HRCN model achieves new state-of-the-art in VeRi-776, VehicleID, and VeRI-Wild.

车辆再识别中的关键问题是在从交叉摄像机查看该对象时找到相同的车辆标识，这对学习视点不变表示提出了更高的要求。在本文中，我们建议从两个方面来解决这个问题：构造健壮的特征表示和提出相机敏感评估。我们首先提出了一种新的异构关系补码网络（HRCN），该网络将特定于区域的特征和跨级别的特征作为原始高级输出的补码。考虑到分布差异和语义错位，我们提出了基于图的关系模块，将这些异构特征嵌入到一个统一的高维空间中。另一方面，考虑到现有措施（即CMC和AP）中跨摄像头评估的不足，我们提出了一种跨摄像头综合措施（CGM），通过引入位置敏感性和跨摄像头综合惩罚来改进评估。我们用我们提出的CGM进一步构建了现有模型的新基准，实验结果表明，我们提出的HRCN模型在VeRi-776、VehicleID和VeRi-Wild中达到了最新水平。

Recently RGB-D sensors have become very popular in the area of simultaneous Localisation and Mapping (SLAM). The RGB-D SLAM approach relies heavily on the accuracy of the input depth map. However, refraction and reflection of transparent objects will result in false depth input of RGB-D cameras, which makes the traditional RGB-D SLAM algorithm unable to work correctly in the presence of transparent objects. In this paper, we propose a novel SLAM approach called transfusion that allows transparent object existence and recovery in the video input. Our method is composed of two parts. Transparent Objects Cut Iterative Closest Points (TC-ICP) is first used to recover camera pose, detecting and removing transparent objects from input to reduce the trajectory errors. Then Transparent Objects Reconstruction (TO-Reconstruction) is used to reconstruct the transparent objects and opaque objects separately. The opaque objects are reconstructed with the traditional method, and the transparent objects are reconstructed with the visual hull-based method. To evaluate our algorithm, we construct a new RGB-D SLAM database containing 25 video sequences. Each sequence has at least one transparent object. Experiments show that our approach can work adequately in scenes contain transparent objects while the existing approach can not handle them. Our approach significantly improves the accuracy of the camera trajectory and the quality of environment reconstruction.

最近，RGB-D传感器在同步定位和绘图（SLAM）领域变得非常流行。RGB-D SLAM方法在很大程度上依赖于输入深度贴图的精度。然而，由于透明物体的折射和反射会导致RGB-D摄像机的深度输入错误，使得传统的RGB-D SLAM算法无法在透明物体存在的情况下正常工作。在本文中，我们提出了一种新的SLAM方法，称为输血，它允许视频输入中透明对象的存在和恢复。我们的方法由两部分组成。透明物体切割迭代最近点（TC-ICP）首先用于恢复相机姿态，检测并从输入中移除透明物体以减少轨迹误差。然后利用透明对象重构（TO Reconstruction）分别对透明对象和不透明对象进行重构。用传统的方法重建不透明对象，用基于视觉外壳的方法重建透明对象。为了评估我们的算法，我们构建了一个新的包含25个视频序列的RGB-D SLAM数据库。每个序列至少有一个透明对象。实验表明，我们的方法可以在包含

透明对象的场景中充分工作，而现有的方法无法处理透明对象。我们的方法显著提高了摄像机轨迹的准确性和环境重建的质量。

Deep neural networks (DNNs) are vulnerable to adversarial noise. Pre-processing based defenses could largely remove adversarial noise by processing inputs. However, they are typically affected by the error amplification effect, especially in the front of continuously evolving attacks. To solve this problem, in this paper, we propose to remove adversarial noise by implementing a self-supervised adversarial training mechanism in a class activation feature space. To be specific, we first maximize the disruptions to class activation features of natural examples to craft adversarial examples. Then, we train a denoising model to minimize the distances between the adversarial examples and the natural examples in the class activation feature space. Empirical evaluations demonstrate that our method could significantly enhance adversarial robustness in comparison to previous state-of-the-art approaches, especially against unseen adversarial attacks and adaptive attacks.

深层神经网络（DNN）易受对抗性噪声的影响。基于预处理的防御可以通过处理输入在很大程度上消除对抗性噪声。然而，它们通常会受到错误放大效应的影响，特别是在不断演变的攻击面前。为了解决这个问题，在本文中，我们提出通过在类激活特征空间中实现一种自我监督的对抗性训练机制来消除对抗性噪声。具体来说，我们首先最大限度地破坏自然示例的类激活特性，以构建对抗性示例。然后，我们训练一个去噪模型，以最小化类激活特征空间中敌对示例和自然示例之间的距离。经验评估表明，与以前的最新方法相比，我们的方法可以显著增强对抗鲁棒性，尤其是对不可见的对抗攻击和自适应攻击。

Existing blind image super-resolution (SR) methods mostly assume blur kernels are spatially invariant across the whole image. However, such an assumption is rarely applicable for real images whose blur kernels are usually spatially variant due to factors such as object motion and out-of-focus. Hence, existing blind SR methods would inevitably give rise to poor performance in real applications. To address this issue, this paper proposes a mutual affine network (MANet) for spatially variant kernel estimation. Specifically, MANet has two distinctive features. First, it has a moderate receptive field so as to keep the locality of degradation. Second, it involves a new mutual affine convolution (MAConv) layer that enhances feature expressiveness without increasing receptive field, model size and computation burden. This is made possible through exploiting channel interdependence, which applies each channel split with an affine transformation module whose input are the rest channel splits. Extensive experiments on synthetic and real images show that the proposed MANet not only performs favorably for both spatially variant and invariant kernel estimation, but also leads to state-of-the-art blind SR performance when combined with non-blind SR methods.

现有的盲图像超分辨率 (SR) 方法大多假设模糊核在整个图像中具有空间不变性。然而，这样的假设很少适用于真实图像，其模糊核通常由于对象运动和失焦等因素而在空间上变化。因此，现有的盲SR方法在实际应用中不可避免地会导致性能下降。为了解决这个问题，本文提出了一种用于空间变异核估计的互仿射网络 (MANet)。具体来说，MANet有两个显著的特点。首先，它有一个适度的感受野，以保持退化的位置。其次，它涉及一个新的互仿射卷积 (MAConv) 层，该层在不增加感受野、模型大小和计算负担的情况下增强了特征表达能力。这是通过利用通道相互依赖性实现的，它使用仿射变换模块应用每个通道分割，仿射变换模块的输入是剩余通道分割。在合成图像和真实图像上的大量实验表明，所提出的MANet不仅具有良好的空间变异和不变核估计性能，而且在与非盲SR方法相结合时还具有最先进的盲SR性能。

Developing deep neural networks to generate 3D scenes is a fundamental problem in neural synthesis with immediate applications in architectural CAD, computer graphics, as well as in generating virtual robot training environments. This task is challenging because 3D scenes exhibit diverse patterns, ranging from continuous ones, such as object sizes and the relative poses between pairs of shapes, to discrete patterns, such as occurrence and co-occurrence of objects with symmetrical relationships. This paper introduces a novel neural scene synthesis approach that can capture diverse feature patterns of 3D scenes. Our method combines the strength of both neural network-based and conventional scene synthesis approaches. We use the parametric prior distributions learned from training data, which provide uncertainties of object attributes and relative attributes, to regularize the outputs of feed-forward neural models. Moreover, instead of merely predicting a scene layout, our approach predicts an over-complete set of attributes. This methodology allows us to utilize the underlying consistency constraints among the predicted attributes to prune infeasible predictions. Experimental results show that our approach outperforms existing methods considerably. The generated 3D scenes interpolate the training data faithfully while preserving both continuous and discrete feature patterns.

开发深度神经网络以生成三维场景是神经合成中的一个基本问题，它直接应用于建筑CAD、计算机图形学以及生成虚拟机器人训练环境。这项任务具有挑战性，因为3D场景显示各种模式，从连续模式（如对象大小和形状对之间的相对姿势）到离散模式（如具有对称关系的对象的出现和共存）。本文介绍了一种新的神经场景合成方法，该方法可以捕获三维场景的各种特征模式。我们的方法结合了基于神经网络和传统场景合成方法的优点。我们使用从训练数据中学习的参数先验分布（提供对象属性和相对属性的不确定性）来正则化前馈神经模型的输出。此外，我们的方法不只是预测场景布局，而是预测一组过于完整的属性。这种方法允许我们利用预测属性之间的潜在一致性约束来修剪不可行的预测。实验结果表明，该方法的性能明显优于现有方法。生成的三维场景忠实地插值训练数据，同时保留连续和离散的特征模式。

Detection of adversarial examples with high accuracy is critical for the security of deployed deep neural network-based models. we present the first graph-based adversarial detection method that constructs a Latent Neighborhood Graph (LNG) around an input example to determine if the input example is adversarial. Given an input example, selected reference adversarial and benign examples are used to capture the local manifold in the vicinity of the input example. The LNG node connectivity parameters are optimized jointly with the parameters of a graph attention network in an end-to-end manner to determine the optimal graph topology for adversarial example detection. The graph attention network is used to determine if the LNG is derived from an adversarial or benign input example. Experimental evaluations on CIFAR-10, STL-10, and ImageNet datasets, using six adversarial attack methods, demonstrate that the proposed method outperforms state-of-the-art adversarial detection methods in white-box and gray-box settings. The proposed method is able to successfully detect adversarial examples crafted with small perturbations using unseen attacks.

高精度的对抗性示例检测对于部署的基于深度神经网络的模型的安全性至关重要。我们提出了第一种基于图的对抗性检测方法，该方法围绕输入示例构造一个潜在邻域图（LNG），以确定输入示例是否具有对抗性。给定一个输入示例，使用选定的参考敌对示例和良性示例捕获输入示例附近的局部流形。LNG节点连接性参数与图形注意网络的参数以端到端的方式联合优化，以确定对抗性示例检测的最佳图形拓扑。图形注意网络用于确定LNG是否来自敌对或良性输入示例。使用六种对抗性攻击方法对CIFAR-10、STL-10和ImageNet数据集进行的实验评估表明，所提出的方法在白盒和灰盒设置下优于最先进的对抗性检测方法。所提出的方法能够成功地检测对抗性的例子制作小扰动使用看不见的攻击。

We present an algorithm for generating novel views at arbitrary viewpoints and any input time step given a monocular video of a dynamic scene. Our work builds upon recent advances in neural implicit representation and uses continuous and differentiable functions for modeling the time-varying structure and the appearance of the scene. We jointly train a time-invariant static NeRF and a time-varying dynamic NeRF, and learn how to blend the results in an unsupervised manner. However, learning this implicit function from a single video is highly ill-posed (with infinitely many solutions that match the input video). To resolve the ambiguity, we introduce regularization losses to encourage a more physically plausible solution. We show extensive quantitative and qualitative results of dynamic view synthesis from casually captured videos.

我们提出了一种算法，用于在给定动态场景的单目视频的任意视点和任意输入时间步长下生成新视图。我们的工作建立在神经隐式表示的最新进展之上，并使用连续和可微函数来建模时变结构和场景外观。我们联合训练一个时不变的静态NeRF和一个时变的动态NeRF，并学习如何以无监督的方式混合结果。然而，从单个视频学习此隐式函数是非常不稳定的（有无限多个与输入视频匹配的解）。为了解决这种模糊性，我们引入正则化损失，以鼓励一种物理上更合理的解决方案。我们展示了从随意捕获的视频中进行动态视图合成的大量定量和定性结果。

Matching local features across images is a fundamental problem in computer vision. Targeting towards high accuracy and efficiency, we propose Seeded Graph Matching Network, a graph neural network with sparse structure to reduce redundant connectivity and learn compact representation. The network consists of 1) Seeding Module, which initializes the matching by generating a small set of reliable matches as seeds. 2) Seeded Graph Neural Network, which utilizes seed matches to pass messages within/across images and predicts assignment costs. Three novel operations are proposed as basic elements for message passing: 1) Attentional Pooling, which aggregates keypoint features within the image to seed matches. 2) Seed Filtering, which enhances seed features and exchanges messages across images. 3) Attentional Unpooling, which propagates seed features back to original keypoints. Experiments show that our method reduces computational and memory complexity significantly compared with typical attention-based networks while competitive or higher performance is achieved.

跨图像匹配局部特征是计算机视觉中的一个基本问题。针对高精度和高效率的目标，我们提出了种子图匹配网络，一种稀疏结构的图神经网络，以减少冗余连接和学习紧凑表示。该网络由1) 种子模块组成，该模块通过生成一小组可靠的匹配作为种子来初始化匹配。2) 种子图神经网络，利用种子匹配在图像内/图像间传递消息，并预测分配成本。提出了三种新的操作作为消息传递的基本元素：1) 注意池，它聚集图像中的关键点特征以进行种子匹配。2) 种子过滤，增强种子功能并在图像间交换消息。3) 注意解除冷却，将种子特征传播回原始关键点。实验表明，与典型的基于注意的网络相比，该方法显著降低了计算复杂度和存储复杂度，同时获得了具有竞争力或更高的性能。

In egocentric videos, the face of a wearer capturing the video is never captured. This gives a false sense of security that the wearer's privacy is preserved while sharing such videos. However, egocentric cameras are typically harnessed to wearer's head, and hence, also capture wearer's gait. Recent works have shown that wearer gait signatures can be extracted from egocentric videos, which can be used to determine if two egocentric videos have the same wearer. In a more damaging scenario, one can even recognize a wearer using hand gestures from egocentric videos, or identify a wearer in third person videos such as from a surveillance camera. We believe, this could be a death knell in sharing of egocentric videos, and fatal for egocentric vision research. In this work, we suggest a novel technique to anonymize egocentric videos, which create carefully crafted, but small, and imperceptible optical flow perturbations in an egocentric video's frames. Importantly, these perturbations do not affect object detection or action/activity recognition from egocentric videos but are strong enough to dis-balance the gait recovery process. In our experiments on benchmark \epic dataset, the proposed perturbation degrades the wearer recognition performance of [??], from 66.3% to 13.4%, while preserving the activity recognition performance of [??] from 89.6% to 87.4%. To test our anonymization with more wearer recognition techniques, we also developed a stronger, and more generalizable wearer recognition method based on camera egomotion cues. The approach achieves state-of-the-art (SOTA) performance of 59.67% on \epicns, compared to 55.06% by [??]. However, the accuracy of our recognition technique also drops to 12% using the proposed anonymizing perturbations.

在以自我为中心的视频中，捕获视频的佩戴者的脸永远不会被捕获。这给人一种虚假的安全感，即在共享此类视频时，佩戴者的隐私得到了保护。然而，以自我为中心的摄像机通常被固定在佩戴者的头部，因此也可以捕捉佩戴者的步态。最近的研究表明，可以从以自我为中心的视频中提取佩戴者的步态特征，这可以用来确定两个以自我为中心的视频是否有相同的佩戴者。在一个更具破坏性的场景中，人们甚至可以通过以自我为中心的视频中的手势识别佩戴者，或者通过监控摄像头等第三人称视频识别佩戴者。我们相信，这可能是分享以自我为中心的视频的丧钟，对以自我为中心的视觉研究来说是致命的。在这项工作中，我们提出了一种新的匿名化自我中心视频的技术，该技术在自我中心视频的帧中创建精心制作但很小且不易察觉的光流扰动。重要的是，这些扰动不会影响以自我为中心的视频中的目标检测或动作/活动识别，但足以破坏步态恢复过程的平衡。在我们在benchmark\epic数据集上的实验中，提出的扰动降低了[? ?]的佩戴者识别性能，从66.3%到13.4%，同时保持[? ?]的活动识别性能从89.6%到87.4%。为了用更多的佩戴者识别技术来测试我们的匿名性，我们还开发了一种更强大、更通用的基于摄像头运动提示的佩戴者识别方法。该方法在epicns上实现了59.67%的最先进（SOTA）性能，而在[? ?]上达到了55.06%。然而，我们的识别技术的准确性也下降到12%，使用建议的匿名扰动。

State-of-the-art object detection approaches typically rely on pre-trained classification models to achieve better performance and faster convergence. We hypothesize that classification pre-training strives to achieve translation invariance, and consequently ignores the localization aspect of the problem. We propose a new large-scale pre-training strategy for detection, where noisy class labels are available for all images, but not bounding-boxes. In this setting, we augment standard classification pre-training with a new detection-specific pretext task. Motivated by the noise-contrastive learning based self-supervised approaches, we design a task that forces bounding boxes with high-overlap to have similar representations in different views of an image, compared to non-overlapping boxes. We redesign Faster R-CNN modules to perform this task efficiently. Our experimental results show significant improvements over existing weakly-supervised and self-supervised pre-training approaches in both detection accuracy as well as fine-tuning speed.

最先进的目标检测方法通常依靠预先训练的分类模型来实现更好的性能和更快的收敛。我们假设分类预训练努力实现翻译不变性，因此忽略了问题的本地化方面。我们提出了一种新的大规模预训练检测策略，其中噪声类标签适用于所有图像，但不适用于边界框。在此设置中，我们使用新的检测特定借口任务来增强标准分类预训练。基于基于噪声对比学习的自监督方法，我们设计了一个任务，与非重叠框相比，具有高重叠的边界框在图像的不同视图中具有相似的表示。我们重新设计了更快的R-CNN模块，以有效地执行此任务。我们的实验结果表明，与现有的弱监督和自监督预训练方法相比，在检测精度和微调速度方面都有显著的改进。

Real-time video inference on edge devices like mobile phones and drones is challenging due to the high computation cost of Deep Neural Networks. We present Adaptive Model Streaming (AMS), a new approach to improving the performance of efficient lightweight models for video inference on edge devices. AMS uses a remote server to continually train and adapt a small model running on the edge device, boosting its performance on the live video using online knowledge distillation from a large, state-of-the-art model. We discuss the challenges of over-the-network model adaptation for video inference and present several techniques to reduce communication the cost of this approach: avoiding excessive overfitting, updating a small fraction of important model parameters, and adaptive sampling of training frames at edge devices. On the task of video semantic segmentation, our experimental results show 0.4--17.8 percent mean Intersection-over-Union improvement compared to a pre-trained model across several video datasets. Our prototype can perform video segmentation at 30 frames-per-second with 40 milliseconds camera-to-label latency on a Samsung Galaxy S10+ mobile phone, using less than 300 Kbps uplink and downlink bandwidth on the device.

由于深度神经网络的高计算成本，在手机和无人机等边缘设备上进行实时视频推断具有挑战性。我们提出了自适应模型流（AMS），这是一种改进边缘设备上视频推断的高效轻量级模型性能的新方法。AMS 使用远程服务器持续培训和调整运行在边缘设备上的小型模型，通过在线知识提炼从大型、最先进的模型中提升其在实时视频上的性能。我们讨论了视频推理的网络模型自适应的挑战，并提出了几种降低通信成本的技术：避免过度拟合，更新一小部分重要模型参数，以及在边缘设备上对训练帧进行自适应采样。在视频语义分割方面，我们的实验结果表明，与预先训练的模型相比，在多个视频数据集中，联合改进的平均交集为0.4%-17.8%。我们的原型可以在三星Galaxy S10+手机上以每秒30帧的速度执行视频分割，使用40毫秒的摄像头标记延迟，在设备上使用小于300 Kbps的上行和下行带宽。

In this paper, we introduce a novel audio-visual multi-modal bridging framework that can utilize both audio and visual information, even with uni-modal inputs. We exploit a memory network that stores source (i.e., visual) and target (i.e., audio) modal representations, where source modal representation is what we are given, and target modal representations are what we want to obtain from the memory network. We then construct an associative bridge between source and target memories that considers the interrelationship between the two memories. By learning the interrelationship through the associative bridge, the proposed bridging framework is able to obtain the target modal representations inside the memory network, even with the source modal input only, and it provides rich information for its downstream tasks. We apply the proposed framework to two tasks: lip reading and speech reconstruction from silent video. Through the proposed associative bridge and modality-specific memories, each task knowledge is enriched with the recalled audio context, achieving state-of-the-art performance. We also verify that the associative bridge properly relates the source and target memories.

在本文中，我们介绍了一种新的视听多模态桥接框架，它可以利用音频和视频信息，即使是单模态输入。我们开发了一个存储源（即视觉）和目标（即音频）模态表示的内存网络，其中源模态表示是我们得到的，目标模态表示是我们希望从内存网络中获得的。然后，我们在源存储器和目标存储器之间构建一个关联桥梁，考虑两个存储器之间的相互关系。通过联想桥学习相互关系，所提出的桥接框架能够获得内存网络中的目标模态表示，即使只有源模态输入，它也能为其下游任务提供丰富的信息。我们将该框架应用于两项任务：唇读和无声视频语音重建。通过所提出的联想桥和特定于模态的记忆，每个任务知识都通过回忆的音频上下文来丰富，从而实现最先进的性能。我们还验证了关联桥正确地关联了源和目标存储器。

Rotation augmentations generally improve a model's invariance/equivariance to rotation – except in object detection. In object detection the shape is not known, therefore rotation creates a label ambiguity. We show that the de-facto method for bounding box label rotation, the Largest Box Method, creates very large labels, leading to poor performance and in many cases worse performance than using no rotation at all. We propose a new method of rotation augmentation that can be implemented in a few lines of code. First, we create a differentiable approximation of label accuracy and show that axis-aligning the bounding box around an ellipse is optimal. We then introduce Rotation Uncertainty (RU) Loss, allowing the model to adapt to the uncertainty of the labels. On five different datasets (including COCO, PascalVOC, and Transparent Object Bin Picking), this approach improves the rotational invariance of both one-stage and two-stage architectures when measured with AP, AP50, and AP75.

旋转增强通常会提高模型对旋转的不变性/等效性，但在对象检测中除外。在目标检测中，形状未知，因此旋转会造成标签模糊。我们展示了边界框标签旋转的实际方法，即最大框方法，会创建非常大的标签，导致性能较差，并且在许多情况下，性能比完全不使用旋转更差。我们提出了一种新的旋转增强方法，只需几行代码即可实现。首先，我们创建一个标签精度的可微近似，并表明围绕椭圆对齐边界框的轴是最优的。然后，我们引入旋转不确定性 (RU) 损失，使模型适应标签的不确定性。在五个不同的数据集（包括COCO、PascalVOC和透明对象箱拾取）上，当使用AP、AP50和AP75进行测量时，该方法提高了一级和两级体系结构的旋转不变性。

weakly supervised semantic segmentation (WSSS) using image-level classification labels usually utilizes the Class Activation Maps (CAMs) to localize objects of interest in images. While pointing out that CAMs only highlight the most discriminative regions of the classes of interest, adversarial erasing (AE) methods have been proposed to further explore the less discriminative regions. In this paper, we review the potential of the pre-trained classifier which is trained on the raw images. We experimentally verify that the ordinary classifier already has the capability to activate the less discriminative regions if the most discriminative regions are erased to some extent. Based on that, we propose a class-specific AE-based framework that fully exploits the potential of an ordinary classifier. Our framework (1) adopts the ordinary classifier to notify the regions to be erased and (2) generates a class-specific mask for erasing by randomly sampling a single specific class to be erased (target class) among the existing classes on the image for obtaining more precise CAMs. Specifically, with the guidance of the ordinary classifier, the proposed CAMs Generation Network (CGNet) is enforced to generate a CAM of the target class while constraining the CAM not to intrude the object regions of the other classes. Along with the pseudo-labels refined from our CAMs, we achieve the state-of-the-art WSSS performance on both PASCAL VOC 2012 and MS-COCO dataset only with image-level supervision. The code is available at <https://github.com/KAIST-vilab/OC-CSE>.

使用图像级分类标签的弱监督语义分割 (WSSS) 通常利用类激活映射 (CAM) 来定位图像中感兴趣的对像。虽然指出CAM只突出感兴趣类别中最具辨别力的区域，但提出了对抗性擦除 (AE) 方法来进一步探索辨别力较低的区域。在本文中，我们回顾了在原始图像上训练的预训练分类器的潜力。我们通过实验证，如果在一定程度上删除了最具辨别力的区域，普通分类器已经能够激活较少辨别力的区域。在此基础上，我们提出了一个基于类特定AE的框架，该框架充分利用了普通分类器的潜力。我们的框架

(1) 采用普通分类器来通知要擦除的区域，(2) 通过在图像上现有类中随机抽样单个要擦除的特定类 (目标类) 来生成用于擦除的特定类掩码，以获得更精确的CAM。具体来说，在普通分类器的指导下，建议的CAM生成网络 (CGNet) 被强制生成目标类的CAM，同时约束CAM不侵入其他类的对象区域。除了从我们的CAMs中改进的伪标签外，我们在PASCAL VOC 2012和MS-COCO数据集上都实现了最先进的WSSS性能，仅在图像级监控的情况下。该守则可于<https://github.com/KAIST-vilab/OC-CSE>.

Domain adaptation is critical for success when confronting with the lack of annotations in a new domain. As the huge time consumption of labeling process on 3D point cloud, domain adaptation for 3D semantic segmentation is of great expectation. With the rise of multi-modal datasets, large amount of 2D images are accessible besides 3D point clouds. In light of this, we propose to further leverage 2D data for 3D domain adaptation by intra and inter domain cross modal learning. As for intra-domain cross modal learning, most existing works sample the dense 2D pixel-wise features into the same size with sparse 3D point-wise features, resulting in the abandon of numerous useful 2D features. To address this problem, we propose Dynamic sparse-to-dense Cross Modal Learning (DsCML) to increase the sufficiency of multi-modality information interaction for domain adaptation. For inter-domain cross modal learning, we further advance Cross Modal Adversarial Learning (CMAL) on 2D and 3D data which contains different semantic content aiming to promote high-level modal complementarity. We evaluate our model under various multi-modality domain adaptation settings including day-to-night, country-to-country and dataset-to-dataset, brings large improvements over both uni-modal and multi-modal domain adaptation methods on all settings.

当面对新领域中注释的缺乏时，领域适应是成功的关键。由于在三维点云上进行标注的过程耗时巨大，因此对三维语义分割的领域自适应具有很大的期望。随着多模态数据集的兴起，除了三维点云之外，还可以访问大量的二维图像。有鉴于此，我们建议通过域内和域间跨模式学习进一步利用2D数据进行3D域自适应。对于域内跨模态学习，现有的研究大多将密集的二维像素特征采样到与稀疏的三维点特征相同的大小，导致大量有用的二维特征被丢弃。为了解决这个问题，我们提出了动态稀疏到密集的跨模态学习 (DsCML)，以增加多模态信息交互对域自适应的充分性。对于跨域跨模态学习，我们进一步提出了基于二维和三维数据的跨模态对抗学习 (CMAL)，该数据包含不同的语义内容，旨在促进高级模态互补性。我们在不同的多模态域适配设置下（包括白天到晚上、国家到国家和数据集到数据集）评估了我们的模型，在所有设置下都比单模态和多模态域适配方法有很大的改进。

Image Signal Processor (ISP) is a crucial component in digital cameras that transforms sensor signals into images for us to perceive and understand. Existing ISP designs always adopt a fixed architecture, e.g., several sequential modules connected in a rigid order. Such a fixed ISP architecture may be suboptimal for real-world applications, where camera sensors, scenes and tasks are diverse. In this study, we propose a novel Reconfigurable ISP (ReconfigISP) whose architecture and parameters can be automatically tailored to specific data and tasks. In particular, we implement several ISP modules, and enable backpropagation for each module by training a differentiable proxy, hence allowing us to leverage the popular differentiable neural architecture search and effectively search for the optimal ISP architecture. A proxy tuning mechanism is adopted to maintain the accuracy of proxy networks in all cases. Extensive experiments conducted on image restoration and object detection, with different sensors, light conditions and efficiency constraints, validate the effectiveness of ReconfigISP. Only hundreds of parameters need tuning for every task.

图像信号处理器 (ISP) 是数码相机中的一个关键部件，它将传感器信号转换成图像供我们感知和理解。现有的ISP设计总是采用固定的体系结构，例如，以刚性顺序连接的几个顺序模块。这种固定的ISP体系结构对于现实世界的应用来说可能是次优的，因为在现实世界中，摄像机传感器、场景和任务是多样的。在这项研究中，我们提出了一种新的可重构ISP (ReconfigISP)，其结构和参数可以根据特定的数据和任务自动调整。特别是，我们实现了几个ISP模块，并通过训练可微代理为每个模块启用反向传播，从而允许我们利用流行的可微神经结构搜索，有效地搜索最佳ISP结构。在所有情况下，采用代理调优机制来保持代理网络的准确性。在不同传感器、光照条件和效率约束下，对图像恢复和目标检测进行了大量实验，验证了ReconfigISP的有效性。每个任务只需要调整数百个参数。

Video instance segmentation (VIS) aims to segment and associate all instances of predefined classes for each frame in videos. Prior methods usually obtain segmentation for a frame or clip first, and merge the incomplete results by tracking or matching. These methods may cause error accumulation in the merging step. Contrarily, we propose a new paradigm -- Propose-Reduce, to generate complete sequences for input videos by a single step. We further build a sequence propagation head on the existing image-level instance segmentation network for long-term propagation. To ensure robustness and high recall of our proposed framework, multiple sequences are proposed where redundant sequences of the same instance are reduced. We achieve state-of-the-art performance on two representative benchmark datasets -- we obtain 47.6% in terms of AP on YouTube-VIS validation set and 70.4% for J&F on DAVIS-UVOS validation set.

视频实例分割 (VIS) 旨在分割和关联视频中每个帧的预定义类的所有实例。以往的方法通常是先对一帧或一个片段进行分割，然后通过跟踪或匹配将不完整的结果进行合并。这些方法可能会导致合并步骤中的错误累积。相反，我们提出了一种新的范例——propose Reduce，通过一个步骤为输入视频生成完整的序列。我们进一步在现有的图像级实例分割网络上构建了一个序列传播头，用于长期传播。为了保证我们提出的框架的健壮性和高召回率，我们提出了多个序列，其中减少了同一实例的冗余序列。我们在两个具有代表性的基准数据集上实现了最先进的性能——我们在YouTube VIS验证集上获得了47.6%的AP，在DAVIS-UVOS验证集上获得了70.4%的J&F。

Snapshot compressive imaging (SCI) aims to record three-dimensional signals via a two-dimensional camera. For the sake of building a fast and accurate SCI recovery algorithm, we incorporate the interpretability of model-based methods and the speed of learning-based ones and present a novel dense deep unfolding network (DUN) with 3D-CNN prior for SCI, where each phase is unrolled from an iteration of Half-Quadratic Splitting (HQS). To better exploit the spatial-temporal correlation among frames and address the problem of information loss between adjacent phases in existing DUNs, we propose to adopt the 3D-CNN prior in our proximal mapping module and develop a novel dense feature map (DFM) strategy, respectively. Besides, in order to promote network robustness, we further propose a dense feature map adaption (DFMA) module to allow inter-phase information to fuse adaptively. All the parameters are learned in an end-to-end fashion. Extensive experiments on simulation data and real data verify the superiority of our method. The source code is available at \href{https://github.com/jianzhangcs/SCI3D}{https://github.com/jianzhangcs/SCI3D} .

快照压缩成像 (SCI) 旨在通过二维摄像机记录三维信号。为了构建快速准确的SCI恢复算法，我们结合基于模型的方法的可解释性和基于学习的方法的速度，提出了一种新的具有3D-CNN先验的SCI密集深度展开网络 (DUN)，其中每个阶段都从半二次分裂 (HQS) 的迭代中展开。为了更好地利用帧间的时空相关性并解决现有DUN中相邻相位之间的信息丢失问题，我们建议在我们的近端映射模块中采用3D-CNN先验，并分别开发一种新的密集特征映射 (DFM) 策略。此外，为了提高网络的鲁棒性，我们进一步提出了一种密集特征映射自适应 (DFMA) 模块，允许相位间信息自适应融合。所有参数都以端到端的方式学习。仿真数据和实际数据的大量实验证明了该方法的优越性。源代码位于\href{https://github.com/jianzhangcs/SCI3D}{https://github.com/jianzhangcs/SCI3D} .

Surface reconstruction from point clouds is a fundamental problem in the computer vision and graphics community. Recent state-of-the-arts solve this problem by individually optimizing each local implicit field during inference. Without considering the geometric relationships between local fields, they typically require accurate normals to avoid the sign conflict problem in overlapped regions of local fields, which severely limits their applicability to raw scans where surface normals could be unavailable. Although SAL breaks this limitation via sign-agnostic learning, further works still need to explore how to extend this technique for local shape modeling. To this end, we propose to learn implicit surface reconstruction by sign-agnostic optimization of convolutional occupancy networks, to simultaneously achieve advanced scalability to large-scale scenes, generality to novel shapes, and applicability to raw scans in a unified framework. Concretely, we achieve this goal by a simple yet effective design, which further optimizes the pre-trained occupancy prediction networks with an unsigned cross-entropy loss during inference. The learning of occupancy fields is conditioned on convolutional features from an hourglass network architecture. Extensive experimental comparisons with previous state-of-the-arts on both object-level and scene-level datasets demonstrate the superior accuracy of our approach for surface reconstruction from un-orientated point clouds. The code is available at <https://github.com/tangjiapeng/SA-ConvONet>.

点云曲面重建是计算机视觉和图形学领域的一个基本问题。最新的技术通过在推理过程中单独优化每个局部隐式场来解决这个问题。在不考虑局部场之间的几何关系的情况下，它们通常需要精确的法线以避免局部场重叠区域中的符号冲突问题，这严重限制了它们对可能无法使用曲面法线的原始扫描的适用性。虽然SAL通过符号不可知学习突破了这一限制，但进一步的工作仍需要探索如何将该技术扩展到局部形状建模。为此，我们建议通过卷积占用网络的符号不可知优化来学习隐式曲面重建，以在统一的框架中同时实现对大规模场景的高级可扩展性、对新颖形状的通用性以及对原始扫描的适用性。具体来说，我们通过一个简单而有效的设计实现了这一目标，该设计进一步优化了预训练的占用预测网络，在推理过程中具有无符号交叉熵损失。占用域的学习以沙漏网络结构的卷积特征为条件。在对象级数据集和场景级数据集上与以前的最新技术进行了广泛的实验比较，结果表明我们的方法对于从未定向点云重建曲面具有更高的精度。该守则可于<https://github.com/tangjiapeng/SA-ConvONet>.

We introduce the active audio-visual source separation problem, where an agent must move intelligently in order to better isolate the sounds coming from an object of interest in its environment. The agent hears multiple audio sources simultaneously (e.g., a person speaking down the hall in a noisy household) and it must use its eyes and ears to automatically separate out the sounds originating from a target object within a limited time budget. Towards this goal, we introduce a reinforcement learning approach that trains movement policies controlling the agent's camera and microphone placement over time, guided by the improvement in predicted audio separation quality. We demonstrate our approach in scenarios motivated by both augmented reality (system is already co-located with the target object) and mobile robotics (agent begins arbitrarily far from the target object). Using state-of-the-art realistic audio-visual simulations in 3D environments, we demonstrate our model's ability to find minimal movement sequences with maximal payoff for audio source separation. Project: <http://vision.cs.utexas.edu/projects/move2hear>

我们介绍了主动音频-视频源分离问题，其中代理必须智能移动，以便更好地隔离来自其环境中感兴趣对象的声音。代理可以同时听到多个音频源（例如，在嘈杂的家庭中，一个人在大厅里讲话），并且必须使用眼睛和耳朵在有限的时间预算内自动分离来自目标对象的声音。为了实现这一目标，我们引入了一种强化学习方法，该方法通过预测音频分离质量的改善来训练运动策略，随着时间的推移控制代理的摄像头和麦克风的放置。我们在增强现实（系统已经与目标对象位于同一位置）和移动机器人（代理从任意远离目标对象的位置开始）的场景中演示了我们的方法。通过在3D环境中使用最先进的真实感视听模

拟，我们证明了我们的模型能够找到最小的运动序列和最大的音源分离回报。项目：<http://vision.cs.utexas.edu/projects/move2hear>

In this paper, we present a novel recurrent multi-view stereo network based on long short-term memory (LSTM) with adaptive aggregation, namely AA-RMVSNet. We firstly introduce an intra-view aggregation module to adaptively extract image features by using context-aware convolution and multi-scale aggregation, which efficiently improves the performance on challenging regions, such as thin objects and large low-textured surfaces. To overcome the difficulty of varying occlusion in complex scenes, we propose an inter-view cost volume aggregation module for adaptive pixel-wise view aggregation, which is able to preserve better-matched pairs among all views. The two proposed adaptive aggregation modules are lightweight, effective and complementary regarding improving the accuracy and completeness of 3D reconstruction. Instead of conventional 3D CNNs, we utilize a hybrid network with recurrent structure for cost volume regularization, which allows high-resolution reconstruction and finer hypothetical plane sweep. The proposed network is trained end-to-end and achieves excellent performance on various datasets. It ranks 1st among all submissions on Tanks and Temples benchmark and achieves competitive results on DTU dataset, which exhibits strong generalizability and robustness. Implementation of our method is available at <https://github.com/QT-Zhu/AA-RMVSNet>.

在本文中，我们提出了一种新的基于长短时记忆 (LSTM) 和自适应聚合的递归多视点立体网络，即AA RMVSNet。我们首先引入了一个视图内聚集模块，通过上下文感知卷积和多尺度聚集自适应地提取图像特征，有效地提高了在具有挑战性的区域（如薄对象和大型低纹理表面）上的性能。为了克服复杂场景中遮挡变化的困难，我们提出了一种用于自适应像素级视图聚合的视图间代价体积聚合模块，该模块能够在所有视图之间保留更好的匹配对。提出的两个自适应聚合模块在提高三维重建的准确性和完整性方面是轻量级、有效和互补的。与传统的3D CNN不同，我们使用具有循环结构的混合网络进行成本-体积正则化，这允许高分辨率重建和更精细的假设平面扫描。该网络经过端到端的训练，在各种数据集上都取得了优异的性能。它在坦克和庙宇基准的所有提交中排名第一，在DTU数据集上取得了具有竞争力的结果，具有很强的通用性和健壮性。我们的方法的实现可在<https://github.com/QT-Zhu/AA-RMVSNet>。

Unsupervised disentanglement learning is a crucial issue for understanding and exploiting deep generative models. Recently, SeFa tries to find latent disentangled directions by performing SVD on the first projection of a pre-trained GAN. However, it is only applied to the first layer and works in a post-processing way. Hessian Penalty minimizes the off-diagonal entries of the output's Hessian matrix to facilitate disentanglement, and can be applied to multi-layers. However, it constrains each entry of output independently, making it not sufficient in disentangling the latent directions (e.g., shape, size, rotation, etc.) of spatially correlated variations. In this paper, we propose a simple Orthogonal Jacobian Regularization (OroJaR) to encourage deep generative model to learn disentangled representations. It simply encourages the variation of output caused by perturbations on different latent dimensions to be orthogonal, and the Jacobian with respect to the input is calculated to represent this variation. We show that our OroJaR also encourages the output's Hessian matrix to be diagonal in an indirect manner. In contrast to the Hessian Penalty, our OroJaR constrains the output in a holistic way, making it very effective in disentangling latent dimensions corresponding to spatially correlated variations. Quantitative and qualitative experimental results show that our method is effective in disentangled and controllable image generation, and performs favorably against the state-of-the-art methods. Our code is available at <https://github.com/csyxwei/OroJaR>.

无监督解纠缠学习是理解和利用深层生成模型的关键问题。最近，SeFa试图通过对预先训练的GAN的第一个投影执行SVD来寻找潜在的分离方向。但是，它仅应用于第一层，并以后处理方式工作。Hessian惩罚最小化输出Hessian矩阵的非对角项，以便于解纠缠，并可应用于多层。然而，它独立地约束输出的每个条目，使得它不足以分离空间相关变化的潜在方向（例如，形状、大小、旋转等）。在本文中，我们提出了一种简单的正交雅可比正则化（OroJaR），以鼓励深层生成模型学习解纠缠表示。它简单地鼓励由不同潜在维度上的扰动引起的输出变化是正交的，并且计算关于输入的雅可比矩阵来表示这种变化。我们证明了我们的OroJaR还鼓励输出的Hessian矩阵以间接方式是对角的。与Hessian惩罚相反，我们的OroJaR以一种整体的方式约束输出，使得它在分离与空间相关变化对应的潜在维度方面非常有效。定量和定性的实验结果表明，我们的方法在分离和可控的图像生成方面是有效的，并且与现有的方法相比表现良好。我们的代码可在<https://github.com/csyxwei/OroJaR>。

Image harmonization aims to improve the quality of image compositing by matching the "appearance" (e.g., color tone, brightness and contrast) between foreground and background images. However, collecting large-scale annotated datasets for this task requires complex professional retouching. Instead, we propose a novel Self-Supervised Harmonization framework (SSH) that can be trained using just "free" natural images without being edited. We reformulate the image harmonization problem from a representation fusion perspective, which separately processes the foreground and background examples, to address the background occlusion issue. This framework design allows for a dual data augmentation method, where diverse [foreground, background, pseudo GT] triplets can be generated by cropping an image with perturbations using 3D color lookup tables (LUTs). In addition, we build a real-world harmonization dataset as carefully created by expert users, for evaluation and benchmarking purposes. Our results show that the proposed self-supervised method outperforms previous state-of-the-art methods in terms of reference metrics, visual quality, and subject user study. Code and dataset will be publicly available.

图像协调旨在通过匹配“外观”（例如色调、亮度和对比度）来提高图像合成的质量在前景和背景图像之间。然而，为此任务收集大规模带注释的数据集需要复杂的专业修饰。相反我们提出了一种新的自我监督协调框架（SSH），可以使用“免费”进行培训“未经编辑的自然图像。我们从表示融合的角度重新阐述了图像协调问题，分别处理前景和背景示例，以解决背景遮挡问题。该框架设计允许在不同的情况下使用双重数据增强方法[前景、背景、伪GT]三元组可以通过使用3D颜色查找表（LUT）对图像进行扰动裁剪来生成。此外，我们还构建了一个由专家用户精心创建的真实世界的协调数据集，用于评估和基准测试。我们的结果表明，所提出的自监督方法在参考指标、视觉质量和主题用户研究方面优于以前的最新方法。代码和数据集将公开提供。

Absolute camera pose regression methods estimate the position and orientation of a camera by only using the captured image. A convolutional backbone with a multi-layer perceptron head is trained with images and pose labels to embed a single reference scene at a time. Recently, this framework was extended for learning multiple scenes with a single model by adding a multi-layer perceptron head per scene. In this work, we propose to learn multi-scene absolute camera pose regression with transformers, where encoders are used to aggregate activation maps with self-attention and decoders transform latent features into candidate pose predictions in parallel, each associated with a different scene. This formulation allows our model to focus on general features that are informative for localization while embedding multiple scenes at once. We evaluate our method on commonly benchmarked indoor and outdoor datasets and show that it surpasses both multi-scene and single-scene absolute pose regressors.

绝对相机姿态回归方法仅使用捕获的图像来估计相机的位置和方向。使用图像和姿势标签训练具有多层次感知器头的卷积主干，以便一次嵌入单个参考场景。最近，通过在每个场景中添加多层感知器头，该框架被扩展用于使用单个模型学习多个场景。在这项工作中，我们建议学习带有变形金刚的多场景绝对相机姿势回归，其中编码器用于聚合具有自我注意的激活贴图，除臭剂编码器将潜在特征并行转换为候选姿势预测，每个预测与不同场景相关联。该公式允许我们的模型在一次嵌入多个场景的同时，将重点放在为本地化提供信息的一般特征上。我们在常用的室内外基准数据集上对我们的方法进行了评估，结果表明，它优于多场景和单场景绝对姿势回归器。

In multi-object detection using neural networks, the fundamental problem is, "How should the network learn a variable number of bounding boxes in different input images?". Previous methods train a multi-object detection network through a procedure that directly assigns the ground truth bounding boxes to the specific locations of the network's output. However, this procedure makes the training of a multi-object detection network too heuristic and complicated. In this paper, we reformulate the multi-object detection task as a problem of density estimation of bounding boxes. Instead of assigning each ground truth to specific locations of network's output, we train a network by estimating the probability density of bounding boxes in an input image using a mixture model. For this purpose, we propose a novel network for object detection called Mixture Density Object Detector (MDOD), and the corresponding objective function for the density-estimation-based training. We applied MDOD to MS COCO dataset. Our proposed method not only deals with multi-object detection problems in a new approach, but also improves detection performances through MDOD. The code is available: <https://github.com/yoojy31/MDOD>.

在使用神经网络的多目标检测中，基本问题是，“网络应该如何在不同的输入图像中学习可变数量的边界框？”。以前的方法通过直接将地面真值边界框指定给网络输出的特定位置的过程来训练多目标检测网络。然而，这个过程使得多目标检测网络的训练过于启发式和复杂。在本文中，我们将多目标检测任务转化为边界盒密度估计问题。我们通过使用混合模型估计输入图像中边界框的概率密度来训练网络，而不是将每个地面真值指定给网络输出的特定位置。为此，我们提出了一种新的目标检测网络，称为混合密度目标检测器（MDOD），以及相应的基于密度估计的训练目标函数。我们将MDOD应用于MS COCO数据集。我们提出的方法不仅以一种新的方法处理多目标检测问题，而且通过MDOD提高了检测性能。代码如下：<https://github.com/yoojy31/MDOD>.

We present a method that takes as input a single dual-pixel image, and simultaneously estimates the image's defocus map---the amount of defocus blur at each pixel---and recovers an all-in-focus image. Our method is inspired from recent works that leverage the dual-pixel sensors available in many consumer cameras to assist with autofocus, and use them for recovery of defocus maps or all-in-focus images. These prior works have solved the two recovery problems independently of each other, and often require large labeled datasets for supervised training. By contrast, we show that it is beneficial to treat these two closely-connected problems simultaneously. To this end, we set up an optimization problem that, by carefully modeling the optics of dual-pixel images, jointly solves both problems. We use data captured with a consumer smartphone camera to demonstrate that, after a one-time calibration step, our approach improves upon prior works for both defocus map estimation and blur removal, despite being entirely unsupervised.

我们提出了一种方法，该方法将单个双像素图像作为输入，同时估计图像的散焦图（每个像素处的散焦模糊量）并恢复全聚焦图像。我们的方法是从最近的工作中得到启发的，这些工作利用了许多消费相机中可用的双像素传感器来辅助自动对焦，并使用它们来恢复散焦贴图或所有对焦图像。这些先前的工作已经独立地解决了这两个恢复问题，并且通常需要大型标记数据集进行监督训练。相比之下，我们表明同时处理这两个密切相关的问题是有益的。为此，我们建立了一个优化问题，通过仔细建模双像素图像

的光学特性，共同解决这两个问题。我们使用消费者智能手机摄像头捕获的数据证明，经过一次性校准步骤后，我们的方法在离焦贴图估计和模糊消除方面都比以前的工作有所改进，尽管完全没有监督。

Human-designed data augmentation strategies have been replaced by automatically learned augmentation policy in the past two years. Specifically, recent works have experimentally shown that the superior performance of the automated methods stems from increasing the diversity of augmented data. However, two factors regarding the diversity of augmented data are still missing: 1) the explicit definition (and thus measurement) of diversity and 2) the quantifiable relationship between diversity and its regularization effects. To fill this gap, we propose a di-versity measure called "Variance Diversity" and theoretically show that the regularization effect of data augmentation is promised by Variance Diversity. We confirm in experiments that the relative gain from automated data augmentation in test accuracy of a given model is highly correlated to Variance Diversity. To improve the search process of automated augmentation, an unsupervised sampling-based framework, DivAug, is designed to directly optimize Variance Diversity and hence strengthen the regularization effect. Without requiring a separate search process, the performance gain from DivAug is comparable with state-of-the-art method with better efficiency. Moreover, under the semi-supervised setting, our framework can further improve the performance of semi-supervised learning algorithms based on RandAugment, making it highly applicable to real-world problems, where labeled data is scarce. The code is available at <https://github.com/warai-Otoko/DivAug>.

在过去两年中，人工设计的数据扩充策略已被自动学习的扩充策略所取代。具体地说，最近的工作实验表明，自动化方法的优越性能来自于增加了增强数据的多样性。然而，关于增强数据多样性的两个因素仍然缺失：1) 多样性的明确定义（以及测量）和2) 多样性及其正则化效应之间的可量化关系。为了填补这一空白，我们提出了一种称为“方差多样性”的多样性度量，并从理论上证明了方差多样性承诺了数据增强的正则化效果。我们在实验中证实，在给定模型的测试精度中，自动数据拟合的相对增益与方差多样性高度相关。为了改进自动增广的搜索过程，设计了一个基于无监督采样的框架DivAug，直接优化变量多样性，从而增强正则化效果。不需要单独的搜索过程，Divagug的性能增益与最先进的方法相当，具有更好的效率。此外，在半监督设置下，我们的框架可以进一步提高基于RandAugment的半监督学习算法的性能，使其非常适用于标记数据稀少的实际问题。该守则可于<https://github.com/warai-Otoko/DivAug>。

We propose the first learning-based approach for fast moving objects detection. Such objects are highly blurred and move over large distances within one video frame. Fast moving objects are associated with a deblurring and matting problem, also called deblatting. We show that the separation of deblatting into consecutive matting and deblurring allows achieving real-time performance, i.e. an order of magnitude speed-up, and thus enabling new classes of application. The proposed method detects fast moving objects as a truncated distance function to the trajectory by learning from synthetic data. For the sharp appearance estimation and accurate trajectory estimation, we propose a matting and fitting network that estimates the blurred appearance without background, followed by an energy minimization based deblurring. The state-of-the-art methods are outperformed in terms of recall, precision, trajectory estimation, and sharp appearance reconstruction. Compared to other methods, such as deblatting, the inference is of several orders of magnitude faster and allows applications such as real-time fast moving object detection and retrieval in large video collections.

我们提出了第一种基于学习的快速运动目标检测方法。这类对象在一个视频帧内高度模糊并移动很远。快速移动的对象与去模糊和消光问题有关，也称为去模糊。我们表明，将去格化分离为连续的消光和去模糊可以实现实时性能，即一个数量级的加速，从而实现新的应用类别。该方法通过对合成数据的学习，将快速运动目标检测为到轨迹的截断距离函数。对于锐利的外观估计和精确的轨迹估计，我们提出了一种matting和fitting网络来估计无背景的模糊外观，然后基于能量最小化的去模糊。在查全率、查准率、轨迹估计和锐利外观重建方面，最先进的方法都优于传统的方法。与其他方法（如去格化）相比，该推理速度快几个数量级，并允许在大型视频采集中实时快速移动目标检测和检索等应用。

Label distributions in real-world are oftentimes long-tailed and imbalanced, resulting in biased models towards dominant labels. While long-tailed recognition has been extensively studied for image classification tasks, limited effort has been made for video domain. In this paper, we introduce VideoLT, a large-scale long-tailed video recognition dataset, as a step toward real-world video recognition. VideoLT contains 256,218 untrimmed videos, annotated into 1,004 classes with a long-tailed distribution. Through extensive studies, we demonstrate that state-of-the-art methods used for long-tailed image recognition do not perform well in the video domain due to the additional temporal dimension in video data. This motivates us to propose FrameStack, a simple yet effective method for long-tailed video recognition task. In particular, FrameStack performs sampling at the frame-level in order to balance class distributions, and the sampling ratio is dynamically determined using knowledge derived from the network during training. Experimental results demonstrate that FrameStack can improve classification performance without sacrificing overall accuracy. Code and dataset are available at: <https://github.com/17Skye17/VideoLT>.

现实世界中的标签分布通常是长尾和不平衡的，导致模型偏向于主导标签。虽然长尾识别在图像分类任务中得到了广泛的研究，但在视频领域的研究却非常有限。在本文中，我们介绍了大规模长尾视频识别数据集VideoLT，作为实现真实视频识别的一个步骤。VideoLT包含256218个未剪辑的视频，注释为1004个类，具有长尾分布。通过广泛的研究，我们证明了用于长尾图像识别的最先进的方法在视频域中表现不佳，因为视频数据中存在额外的时间维度。这促使我们提出FrameStack，一种简单但有效的长尾视频识别方法。特别是，FrameStack在帧级别执行采样以平衡类分布，并且采样率是使用在训练期间从网络派生的知识动态确定的。实验结果表明，FrameStack可以在不牺牲整体精度的情况下提高分类性能。代码和数据集位于：<https://github.com/17Skye17/VideoLT>.

In this paper, we propose an anchor-free single-stage LiDAR-based 3D object detector -- RangeDet. The most notable difference with previous works is that our method is purely based on the range view representation. Compared with the commonly used voxelized or Bird's Eye View (BEV) representations, the range view representation is more compact and without quantization error. Although there are works adopting it for semantic segmentation, its performance in object detection is largely behind voxelized or BEV counterparts. We first analyze the existing range-view-based methods and find two issues overlooked by previous works: 1) the scale variation between nearby and far away objects; 2) the inconsistency between the 2D range image coordinates used in feature extraction and the 3D Cartesian coordinates used in output. Then we deliberately design three components to address these issues in our RangeDet. We test our RangeDet in the large-scale Waymo Open Dataset (WOD). Our best model achieves 72.9/75.9/65.8 3D AP on vehicle/pedestrian/cyclist. These results outperform other range-view-based methods by a large margin, and are overall comparable with the state-of-the-art multi-view-based methods. Codes will be released at <https://github.com/TuSimple/RangeDet>.

在本文中，我们提出了一种基于无锚单级激光雷达的三维目标探测器——RangeDet。与以前的工作最显著的区别是，我们的方法完全基于范围视图表示。与常用的体素化或鸟瞰视图（BEV）表示法相比，距离视图表示法更紧凑且无量化误差。虽然有一些工作将其用于语义分割，但其在目标检测中的性能在很大程度上落后于体素化或BEV。我们首先分析了现有的基于距离视图的方法，发现了以前工作中忽略的两个问题：1) 附近和远处物体之间的尺度变化；2) 特征提取中使用的二维范围图像坐标与输出中使用的三维笛卡尔坐标不一致。然后，我们在RangeDet中特意设计了三个组件来解决这些问题。我们在大规模Waymo开放数据集（WOD）中测试RangeDet。我们的最佳模型在车辆/行人/自行车上实现72.9/75.9/65.8 3D AP。这些结果大大优于其他基于范围视图的方法，并且总体上与最先进的基于多视图的方法相当。守则将于<https://github.com/TuSimple/RangeDet>。

Universal adversarial perturbation (UAP), i.e. a single perturbation to fool the network for most images, is widely recognized as a more practical attack because the UAP can be generated beforehand and applied directly during the attack stage. One intriguing phenomenon regarding untargeted UAP is that most images are misclassified to a dominant label. This phenomenon has been reported in previous works while lacking a justified explanation, for which our work attempts to provide an alternative explanation. For a more practical universal attack, our investigation of untargeted UAP focuses on alleviating the dependence on the original training samples, from removing the need for sample labels to limiting the sample size. Towards strictly data-free untargeted UAP, our work proposes to exploit artificial jigsaw images as the training samples, demonstrating competitive performance. We further investigate the possibility of exploiting the UAP for a data-free black-box attack which is arguably the most practical yet challenging threat model. We demonstrate that there exists optimization-free repetitive patterns which can successfully attack deep models. Code is available at <https://bit.ly/3y0ZTIC>.

普遍对抗性干扰（UAP），即针对大多数图像欺骗网络的单一干扰，被广泛认为是一种更实际的攻击，因为UAP可以事先生成并在攻击阶段直接应用。关于非目标UAP的一个有趣现象是，大多数图像被错误分类为一个主要标签。这一现象在以前的著作中已有报道，但缺乏合理的解释，对此我们的工作试图提供另一种解释。对于更实际的通用攻击，我们对非目标UAP的研究侧重于减轻对原始训练样本的依赖，从消除对样本标签的需要到限制样本大小。为了实现严格的无数据无目标UAP，我们的工作建议利用人工拼图图像作为训练样本，展示具有竞争力的性能。我们进一步研究了利用UAP进行无数据黑盒攻击的可能性，这可能是最实际但最具挑战性的威胁模型。我们证明了存在无优化的重复模式，可以成功地攻击深度模型。代码可在<https://bit.ly/3y0ZTIC>。

Imaging depth and spectrum have been extensively studied in isolation from each other for decades. Recently, hyperspectral-depth (HS-D) imaging emerges to capture both information simultaneously by combining two different imaging systems; one for depth, the other for spectrum. While being accurate, this combinational approach induces increased form factor, cost, capture time, and alignment/registration problems. In this work, departing from the combinational principle, we propose a compact single-shot monocular HS-D imaging method. Our method uses a diffractive optical element (DOE), the point spread function of which changes with respect to both depth and spectrum. This enables us to reconstruct spectrum and depth from a single captured image. To this end, we develop a differentiable simulator and a neural-network-based reconstruction method that are jointly optimized via automatic differentiation. To facilitate learning the DOE, we present a first HS-D dataset by building a benchtop HS-D imager that acquires high-quality ground truth. We evaluate our method with synthetic and real experiments by building an experimental prototype and achieve state-of-the-art HS-D imaging results.

几十年来，成像深度和光谱已被广泛研究，彼此独立。近年来，超光谱深度成像（HS-D）技术应运而生，它通过组合两种不同的成像系统来同时捕获这两种信息；一个用于深度，另一个用于光谱。这种组合方法虽然准确，但会导致形状因素、成本、捕获时间和对齐/注册问题增加。在这项工作中，我们从组合原理出发，提出了一种紧凑的单镜头单目HS-D成像方法。我们的方法使用衍射光学元件（DOE），其点扩散函数随深度和光谱而变化。这使我们能够从单个捕获的图像重建光谱和深度。为此，我们开发了一个可微模拟器和基于神经网络的重建方法，通过自动微分进行联合优化。为了便于学习DOE，我们通过构建获取高质量地面真相的台式HS-D成像仪，展示了第一个HS-D数据集。我们通过构建一个实验原型，通过合成和真实实验来评估我们的方法，并获得最先进的HS-D成像结果。

Recent advancements in deep neural networks have made remarkable leap-forwards in dense image prediction. However, the issue of feature alignment remains as neglected by most existing approaches for simplicity. Direct pixel addition between upsampled and local features leads to feature maps with misaligned contexts that, in turn, translate to mis-classifications in prediction, especially on object boundaries. In this paper, we propose a feature alignment module that learns transformation offsets of pixels to contextually align upsampled higher-level features; and another feature selection module to emphasize the lower-level features with rich spatial details. We then integrate these two modules in a top-down pyramidal architecture and present the Feature-aligned Pyramid Network (FaPN). Extensive experimental evaluations on four dense prediction tasks and four datasets have demonstrated the efficacy of FaPN, yielding an overall improvement of 1.2 - 2.6 points in AP / mIoU over FPN when paired with Faster / Mask R-CNN. In particular, our FaPN achieves the state-of-the-art of 56.7% mIoU on ADE20K when integrated within Mask-Former. The code is available from <https://github.com/EMI-Group/FaPN>.

深度神经网络的最新进展使密集图像预测取得了显著的飞跃。然而，为了简单起见，大多数现有方法仍然忽略了特征对齐问题。在上采样特征和局部特征之间直接添加像素会导致特征映射的上下文不对齐，进而导致预测中的错误分类，尤其是在对象边界上。在本文中，我们提出了一个特征对齐模块，该模块学习像素的变换偏移量，以上下文对齐上采样的高级特征；另一个特征选择模块强调具有丰富空间细节的底层特征。然后，我们将这两个模块集成到一个自顶向下的金字塔结构中，并提出了特征对齐金字塔网络（FaPN）。对四个密集预测任务和四个数据集的广泛实验评估已经证明了FaPN的有效性，当与Faster/Mask R-CNN配对时，AP/mIoU比FPN的总体改善1.2-2.6个点。特别是，当集成在掩模成型器中时，我们的FaPN在ADE20K上实现了56.7%mIoU的最先进水平。该代码可从以下网址获得：<https://github.com/EMI-Group/FaPN>。

Recent advances in attention-based networks have shown that Vision Transformers can achieve state-of-the-art or near state-of-the-art results on many image classification tasks. This puts transformers in the unique position of being a promising alternative to traditional convolutional neural networks (CNNs). While CNNs have been carefully studied with respect to adversarial attacks, the same cannot be said of Vision Transformers. In this paper, we study the robustness of Vision Transformers to adversarial examples. Our analyses of transformer security is divided into three parts. First, we test the transformer under standard white-box and black-box attacks. Second, we study the transferability of adversarial examples between CNNs and transformers. We show that adversarial examples do not readily transfer between CNNs and transformers. Based on this finding, we analyze the security of a simple ensemble defense of CNNs and transformers. By creating a new attack, the self-attention blended gradient attack, we show that such an ensemble is not secure under a white-box adversary. However, under a black-box adversary, we show that an ensemble can achieve unprecedented robustness without sacrificing clean accuracy. Our analysis for this work is done using six types of white-box attacks and two types of black-box attacks. Our study encompasses multiple Vision Transformers, Big Transfer Models and CNN architectures trained on CIFAR-10, CIFAR-100 and ImageNet.

基于注意的网络的最新进展表明，视觉变换器可以在许多图像分类任务中实现最先进或接近最先进的结果。这使变压器处于独特的地位，成为传统卷积神经网络（CNN）的一个有前途的替代品。虽然CNN在对抗性攻击方面进行了仔细的研究，但在视觉变形金刚方面却不是这样。在本文中，我们研究了视觉变换器对对抗性示例的鲁棒性。我们对变压器安全性的分析分为三个部分。首先，我们在标准的白盒和黑盒攻击下测试变压器。其次，我们研究了CNN和Transformer之间对抗性示例的可转移性。我们表明，对抗性示例不容易在CNN和变压器之间传输。基于这一发现，我们分析了CNN和变压器的简单集成防御的安全性。通过创建一种新的攻击，即自注意混合梯度攻击，我们证明了这样的集成在白盒对手下是不安全的。然而，在黑箱对手的情况下，我们证明了一个集成可以在不牺牲精确性的情况下实现前所未有的健壮性。我们使用六种类型的白盒攻击和两种类型的黑盒攻击来分析这项工作。我们的研究包括在CIFAR-10、CIFAR-100和ImageNet上训练的多个视觉转换器、大传输模型和CNN架构。

Transformers, which are popular for language modeling, have been explored for solving vision tasks recently, e.g., the Vision Transformer (ViT) for image classification. The ViT model splits each image into a sequence of tokens with fixed length and then applies multiple Transformer layers to model their global relation for classification. However, ViT achieves inferior performance to CNNs when trained from scratch on a midsize dataset like ImageNet. We find it is because: 1) the simple tokenization of input images fails to model the important local structure such as edges and lines among neighboring pixels, leading to low training sample efficiency; 2) the redundant attention backbone design of ViT leads to limited feature richness for fixed computation budgets and limited training samples. To overcome such limitations, we propose a new Tokens-To-Token vision Transformer (T2T-ViT), which incorporates 1) a layer-wise Tokens-to-Token (T2T) transformation to progressively structure the image to tokens by recursively aggregating neighboring Tokens into one Token (Tokens-to-Token), such that local structure represented by surrounding tokens can be modeled and tokens length can be reduced; 2) an efficient backbone with a deep-narrow structure for vision transformer motivated by CNN architecture design after empirical study. Notably, T2T-ViT reduces the parameter count and MACs of vanilla ViT by half, while achieving more than 3.0% improvement when trained from scratch on ImageNet. It also outperforms ResNets and achieves comparable performance with MobileNets by directly training on ImageNet. For example, T2T-ViT with comparable size to ResNet50 (21.5M parameters) can achieve 83.3% top1 accuracy in image resolution 384x384 on ImageNet.

变形金刚（Transformers）是语言建模领域的热门工具，近年来被广泛用于解决视觉任务，例如用于图像分类的视觉变形金刚（ViT）。ViT模型将每个图像分割成一系列具有固定长度的标记，然后应用多个转换器层来建模其全局关系以进行分类。然而，当在中型数据集（如ImageNet）上从头开始训练时，ViT的性能不如CNN。我们发现这是因为：1) 输入图像的简单标记化未能对重要的局部结构（如相邻像素之间的边和线）建模，导致训练样本效率低下；2) ViT的冗余注意主干设计导致固定计算预算和有限训练样本的特征丰富性有限。为了克服这些限制，我们提出了一种新的令牌到令牌视觉转换器（T2T ViT），它包含1) 分层令牌到令牌（T2T）转换，通过递归地将相邻令牌聚合为一个令牌（令牌到令牌），逐步将图像结构化为令牌，这样，可以对由周围令牌表示的局部结构进行建模，并且可以减少令牌的长度；2) 通过实证研究，提出了一种基于CNN体系结构设计的、用于视觉转换器的深窄结构的高效主干网。值得注意的是，T2T ViT将香草ViT的参数计数和MAC值降低了一半，而在ImageNet上从头开始训练时，其改善率超过3.0%。通过直接在ImageNet上进行培训，它的性能也优于ResNet，并与MobileNet具有相当的性能。例如，尺寸与ResNet50相当的T2T ViT（21.5M参数）在ImageNet上的图像分辨率384x384可达到83.3%的top1精度。

We present the Teacher-Student Generative Adversarial Network (TS-GAN) to generate depth images from single RGB images in order to boost the performance of face recognition systems. For our method to generalize well across unseen datasets, we design two components in the architecture, a teacher and a student. The teacher, which itself consists of a generator and a discriminator, learns a latent mapping between input RGB and paired depth images in a supervised fashion. The student, which consists of two generators (one shared with the teacher) and a discriminator, learns from new RGB data with no available paired depth information, for improved generalization. The fully trained shared generator can then be used in runtime to hallucinate depth from RGB for downstream applications such as face recognition. We perform rigorous experiments to show the superiority of TS-GAN over other methods in generating synthetic depth images. Moreover, face recognition experiments demonstrate that our hallucinated depth along with the input RGB images boost performance across various architectures when compared to a single RGB modality by average values of +1.2%, +2.6%, and +2.6% for IIIT-D, EURECOM, and LFW datasets respectively. We make our implementation public at: <https://github.com/hardik-uppal/teacher-student-gan.git>.

为了提高人脸识别系统的性能，我们提出了教师-学生成对抗网络（TS-GAN）从单个RGB图像生成深度图像。为了使我们的方法能够很好地在看不见的数据集中推广，我们在体系结构中设计了两个组件，一个教师和一个学生。教师本身由生成器和鉴别器组成，以有监督的方式学习输入RGB和成对深度图像之间的潜在映射。该学生由两个生成器（一个与教师共享）和一个鉴别器组成，从没有可用成对深度信息的新RGB数据中学习，以改进泛化。经过充分训练的共享生成器可以在运行时用于从RGB产生深度幻觉，用于下游应用程序，如人脸识别。我们进行了严格的实验，以证明TS-GAN在生成合成深度图像方面优于其他方法。此外，人脸识别实验表明，当IIIT-D、EURECOM和LFW数据集的平均值分别为+1.2%、+2.6%和+2.6%时，与单个RGB模式相比，我们的幻觉深度以及输入RGB图像在各种架构中提高了性能。我们在以下网站公布实施情况：<https://github.com/hardik-uppal/teacher-student-gan.git>.

Generative Adversarial Networks (GANs) have witnessed prevailing success in yielding outstanding images, however, they are burdensome to deploy on resource-constrained devices due to ponderous computational costs and hulking memory usage. Although recent efforts on compressing GANs have acquired remarkable results, they still exist potential model redundancies and can be further compressed. To solve this issue, we propose a novel online multi-granularity distillation (OMGD) scheme to obtain lightweight GANs, which contributes to generating high-fidelity images with low computational demands. We offer the first attempt to popularize single-stage online distillation for GAN-oriented compression, where the progressively promoted teacher generator helps to refine the discriminator-free based student generator. Complementary teacher generators and network layers provide comprehensive and multi-granularity concepts to enhance visual fidelity from diverse dimensions. Experimental results on four benchmark datasets demonstrate that OMGD succeeds to compress 40xMACs and 82.5xparameters on Pix2Pix and CycleGAN, without loss of image quality. It reveals that OMGD provides a feasible solution for the deployment of real-time image translation on resource-constrained devices. Our code and models are made public at: <https://github.com/bytedance/OMGD>

生成性对抗网络 (GAN) 在生成出色的图像方面取得了巨大的成功，但是，由于沉重的计算成本和庞大的内存使用，它们在资源受限的设备上部署起来很麻烦。尽管最近压缩GAN的努力取得了显著的成果，但它们仍然存在潜在的模型冗余，可以进一步压缩。为了解决这个问题，我们提出了一种新的在线多粒度蒸馏 (OMGD) 方案来获得轻量级的GAN，这有助于以较低的计算需求生成高保真图像。我们首次尝试推广用于GAN定向压缩的单级在线蒸馏，逐步提升的教师生成器有助于改进基于无鉴别器的学生生成器。互补的教师生成器和网络层提供了全面和多粒度的概念，以从不同维度增强视觉逼真度。在四个基准数据集上的实验结果表明，OMGD成功地压缩了Pix2Pix和CycleGAN上的40xMAC和82.5xparameters，而不损失图像质量。这表明OMGD为在资源受限的设备上部署实时图像翻译提供了一个可行的解决方案。我们的代码和模型公开于：<https://github.com/bytedance/OMGD>

In this paper, we propose a balancing training method to address problems in imbalanced data learning. To this end, we derive a new loss used in the balancing training phase that alleviates the influence of samples that cause an overfitted decision boundary. The proposed loss efficiently improves the performance of any type of imbalance learning methods. In experiments on multiple benchmark data sets, we demonstrate the validity of our method and reveal that the proposed loss outperforms the state-of-the-art cost-sensitive loss methods. Furthermore, since our loss is not restricted to a specific task, model, or training method, it can be easily used in combination with other recent re-sampling, meta-learning, and cost-sensitive learning methods for class-imbalance problems. Our code is made available.

在本文中，我们提出了一种平衡训练方法来解决不平衡数据学习中的问题。为此，我们推导了一种用于平衡训练阶段的新损失，该损失可减轻导致过度拟合决策边界的样本的影响。所提出的损失有效地提高了任何类型的不平衡学习方法的性能。在多个基准数据集上的实验中，我们证明了我们的方法的有效性，并揭示了所提出的损失优于最新的成本敏感损失方法。此外，由于我们的损失不限于特定的任务、模型或训练方法，因此它可以很容易地与其他最近的重新抽样、元学习和成本敏感学习方法结合使用，以解决班级不平衡问题。我们的代码是可用的。

One of the fundamental goals of visual perception is to allow agents to meaningfully interact with their environment. In this paper, we take a step towards that long-term goal -- we extract highly localized actionable information related to elementary actions such as pushing or pulling for articulated objects with movable parts. For example, given a drawer, our network predicts that applying a pulling force on the handle opens the drawer. We propose, discuss, and evaluate novel network architectures that given image and depth data, predict the set of actions possible at each pixel, and the regions over articulated parts that are likely to move under the force. We propose a learning-from-interaction framework with an online data sampling strategy that allows us to train the network in simulation (SAPIEN) and generalizes across categories. Check the website for code and data release.

视觉感知的一个基本目标是让代理与环境进行有意义的交互。在本文中，我们朝着这个长期目标迈出了一步——我们提取了与基本动作相关的高度本地化的可操作信息，例如推或拉带有可移动部件的铰接对象。例如，给定一个抽屉，我们的网络预测在把手上施加拉力会打开抽屉。我们提出、讨论和评估新的网络架构，这些架构提供图像和深度数据，预测每个像素上可能的动作集，以及在力作用下可能移动的关节部位上的区域。我们提出了一个在线数据采样策略的交互学习框架，该框架允许我们在模拟中训练网络（SAPIEN）并跨类别进行概括。检查网站上的代码和数据发布。

The abundance and richness of Internet photos of landmarks and cities has led to significant progress in 3D vision over the past two decades, including automated 3D reconstructions of the world's landmarks from tourist photos. However, a major source of information available for these 3D-augmented collections---language, e.g., from image captions---has been virtually untapped. In this work, we present WikiScenes, a new, large-scale dataset of landmark photo collections that contains descriptive text in the form of captions and hierarchical category names. WikiScenes forms a new testbed for multimodal reasoning involving images, text, and 3D geometry. We demonstrate the utility of WikiScenes for learning semantic concepts over images and 3D models. Our weakly-supervised framework connects images, 3D structure and semantics---utilizing the strong constraints provided by 3D geometry---to associate semantic concepts to image pixels and points in 3D space.

在过去的二十年中，地标和城市的互联网照片的丰富和丰富导致了3D视觉的重大进步，包括从旅游照片中自动重建世界地标。然而，这些3D增强集合的一个主要信息来源——语言，例如，来自图像字幕的信息——实际上尚未开发。在这项工作中，我们介绍了WikiScenes，这是一个新的大规模地标照片集数据集，其中包含标题和分级类别名称形式的描述性文本。WikiScenes为涉及图像、文本和3D几何体的多模态推理提供了一个新的测试平台。我们演示了WikiScenes在图像和3D模型上学习语义概念的实用性。我们的弱监督框架将图像、3D结构和语义连接起来，利用3D几何提供的强大约束，将语义概念与3D空间中的图像像素和点关联起来。

Removing noise from scanned pages is a vital step before their submission to optical character recognition (OCR) system. Most available image denoising methods are supervised where the pairs of noisy/clean pages are required. However, this assumption is rarely met in real settings. Besides, there is no single model that can remove various noise types from documents. Here, we propose a unified end-to-end unsupervised deep learning model, for the first time, that can effectively remove multiple types of noise, including salt & pepper noise, blurred and/or faded text, as well as watermarks from documents at various levels of intensity. We demonstrate that the proposed model significantly improves the quality of scanned images and the OCR of the pages on several test datasets.

在扫描页面提交到光学字符识别（OCR）系统之前，去除噪声是至关重要的一步。大多数可用的图像去噪方法都是在需要噪声/干净页面对的情况下进行监督的。然而，这一假设在实际环境中很少得到满足。此外，没有单一的模型可以消除文档中的各种噪声类型。在这里，我们首次提出了一个统一的端到端无监督深度学习模型，该模型可以有效地去除多种类型的噪声，包括椒盐噪声、模糊和/或褪色文本，以及不同强度文档中的水印。我们证明了所提出的模型显著提高了扫描图像的质量，并在多个测试数据集上提高了页面的OCR。

A standard practice of deploying deep neural networks is to apply the same architecture to all the input instances. However, a fixed architecture may not be suitable for different data with high diversity. To boost the model capacity, existing methods usually employ larger convolutional kernels or deeper network layers, which incurs prohibitive computational costs. In this paper, we address this issue by proposing Differentiable Dynamic Wirings (DDW), which learns the instance-aware connectivity that creates different wiring patterns for different instances. 1) Specifically, the network is initialized as a complete directed acyclic graph, where the nodes represent convolutional blocks and the edges represent the connection paths. 2) We generate edge weights by a learnable module, Router, and select the edges whose weights are larger than a threshold, to adjust the connectivity of the neural network structure. 3) Instead of using the same path of the network, DDW aggregates features dynamically in each node, which allows the network to have more representation power. To facilitate effective training, we further represent the network connectivity of each sample as an adjacency matrix. The matrix is updated to aggregate features in the forward pass, cached in the memory, and used for gradient computing in the backward pass. We validate the effectiveness of our approach with several mainstream architectures, including MobileNetV2, ResNet, ResNeXt, and RegNet. Extensive experiments are performed on ImageNet classification and COCO object detection, which demonstrates the effectiveness and generalization ability of our approach.

部署深度神经网络的标准实践是对所有输入实例应用相同的体系结构。但是，固定的体系结构可能不适用于具有高度多样性的不同数据。为了提高模型容量，现有的方法通常采用更大的卷积核或更深的网络层，这会导致高昂的计算成本。在本文中，我们通过提出可微动态布线（DDW）来解决这个问题，DDW学习实例感知连通性，从而为不同实例创建不同的布线模式。1）具体来说，网络初始化为一个完整的有向无环图，其中节点表示卷积块，边表示连接路径。2）我们通过一个可学习的模块，路由器来生成边权值，并选择权值大于阈值的边来调整神经网络结构的连通性。3）DDW不使用网络的同一路径，而是在每个节点中动态聚合功能，从而使网络具有更大的表示能力。为了便于有效的训练，我们进一步将每个样本的网络连通性表示为邻接矩阵。矩阵将更新以聚合前向过程中的特征，缓存在内存中，并用于后向过程中的梯度计算。我们用几个主流体系结构验证了我们的方法的有效性，包括MobileNetV2、ResNet、ResNeXt和RegNet。在ImageNet分类和COCO目标检测上进行了大量实验，证明了该方法的有效性和泛化能力。

Lensless cameras provide a framework to build thin imaging systems by replacing the lens in a conventional camera with an amplitude or phase mask near the sensor. Existing methods for lensless imaging can recover the depth and intensity of the scene, but they require solving computationally-expensive inverse problems. Furthermore, existing methods struggle to recover dense scenes with large depth variations. In this paper, we propose a lensless imaging system that captures a small number of measurements using different patterns on a programmable mask. In this context, we make three contributions. First, we present a fast recovery algorithm to recover textures on a fixed number of depth planes in the scene. Second, we consider the mask design problem, for programmable lensless cameras, and provide a design template for optimizing the mask patterns with the goal of improving depth estimation. Third, we use a refinement network as a post-processing step to identify and remove artifacts in the reconstruction. These modifications are evaluated extensively with experimental results on a lensless camera prototype to showcase the performance benefits of the optimized masks and recovery algorithms over the state of the art.

无透镜相机通过在传感器附近用振幅或相位掩模替换传统相机中的镜头，提供了一个构建薄成像系统的框架。现有的无透镜成像方法可以恢复场景的深度和强度，但需要解决计算量大的反问题。此外，现有方法难以恢复深度变化较大的密集场景。在本文中，我们提出了一种无透镜成像系统，该系统使用可编程掩模上的不同图案捕获少量测量值。在这方面，我们作出三项贡献。首先，我们提出了一种快速恢复算法来恢复场景中固定数量深度平面上的纹理。第二，我们考虑掩模设计问题，可编程无透镜相机，并提供了一个设计模板优化掩模图案，目的是改善深度估计。第三，我们使用细化网络作为后处理步骤来识别和移除重建中的伪影。在无透镜相机原型上的实验结果对这些修改进行了广泛评估，以展示优化的掩模和恢复算法相对于最新技术的性能优势。

Joint Energy-based Model (JEM) is a recently proposed hybrid model that retains strong discriminative power of modern CNN classifiers, while generating samples rivaling the quality of GAN-based approaches. In this paper, we propose a variety of new training procedures and architecture features to improve JEM's accuracy, training stability, and speed altogether. 1) we propose a proximal SGLD to generate samples in the proximity of samples from previous step, which improves the stability. 2) we further treat the approximate maximum likelihood learning of EBM as a multi-step differential game, and extend the YOPO framework to cut out redundant calculations during backpropagation, which accelerates the training substantially. 3) Rather than initializing SGLD chain from random noise, we introduce a new informative initialization that samples from a distribution estimated from training data. 4) This informative initialization allows us to enable batch normalization in JEM, which further releases the power of modern CNN architectures for hybrid modeling.

联合能量模型 (JEM) 是最近提出的一种混合模型，它保留了现代CNN分类器的强大识别能力，同时生成的样本与基于GAN的方法的质量相当。在本文中，我们提出了各种新的训练程序和体系结构特征，以提高JEM的准确性、训练稳定性和速度。1) 我们提出了一种近端SGLD，在前一步的样本附近生成样本，提高了稳定性。2) 我们进一步将EBM的近似最大似然学习视为一个多步微分对策，并扩展了YOPO框架，以减少反向传播过程中的冗余计算，从而大大加快了训练速度。3) 我们没有从随机噪声中初始化SGLD链，而是引入了一种新的信息初始化，该初始化从训练数据估计的分布中采样。4) 这种信息丰富的初始化使我们能够在JEM中实现批处理规范化，这进一步释放了现代CNN体系结构用于混合建模的能力。

Recent Visual Question Answering (VQA) models have shown impressive performance on the VQA benchmark but remain sensitive to small linguistic variations in input questions. Existing approaches address this by augmenting the dataset with question paraphrases from visual question generation models or adversarial perturbations. These approaches use the combined data to learn an answer classifier by minimizing the standard cross-entropy loss. To more effectively leverage augmented data, we build on the recent success in contrastive learning. We propose a novel training paradigm (ConClat) that optimizes both cross-entropy and contrastive losses. The contrastive loss encourages representations to be robust to linguistic variations in questions while the cross-entropy loss preserves the discriminative power of representations for answer prediction. We find that optimizing both losses -- either alternately or jointly -- is key to effective training. On the VQA-Rephrasings benchmark, which measures the VQA model's answer consistency across human paraphrases of a question, ConClat improves Consensus Score by 1.63% over an improved baseline. In addition, on the standard VQA 2.0 benchmark, we improve the VQA accuracy by 0.78% overall. We also show that ConClat is agnostic to the type of data-augmentation strategy used.

最近的可视化问答 (VQA) 模型在VQA基准上表现出了令人印象深刻的性能，但对输入问题中的小的语言变化仍然敏感。现有的方法通过使用来自可视问题生成模型或对抗性干扰的问题解释来扩充数据集来解决这一问题。这些方法使用组合数据通过最小化标准交叉熵损失来学习答案分类器。为了更有效地利用扩充数据，我们在对比学习方面取得了近期的成功。我们提出了一种新的训练范式 (ConClat)，它优化了交叉熵和对比损失。对比损失鼓励表征对问题中的语言变化具有鲁棒性，而交叉熵损失保持表征对答案预测的辨别力。我们发现，优化两种损失——交替或联合——是有效培训的关键。在VQA改写基准上，ConClat将共识分数提高了1.63%，这一基准衡量了VQA模型在问题的人类释义中的答案一致性。此外，在标准VQA 2.0基准上，我们将VQA精度总体提高了0.78%。我们还表明ConClat对所使用的数据增强策略的类型是不可知的。

Scene understanding under low-light conditions is a challenging problem. This is due to the small number of photons captured by the camera and the resulting low signal-to-noise ratio (SNR). Single-photon cameras (SPCs) are an emerging sensing modality that are capable of capturing images with high sensitivity. Despite having minimal read-noise, images captured by SPCs in photon-starved conditions still suffer from strong shot noise, preventing reliable scene inference. We propose photon scale-space -- a collection of high-SNR images spanning a wide range of photons-per-pixel (PPP) levels (but same scene content) as guides to train inference model on low photon flux images. We develop training techniques that push images with different illumination levels closer to each other in feature representation space. The key idea is that having a spectrum of different brightness levels during training enables effective guidance, and increases robustness to shot noise even in extreme noise cases. Based on the proposed approach, we demonstrate, via simulations and real experiments with a SPAD camera, high-performance on various inference tasks such as image classification and monocular depth estimation under ultra low-light, down to <1 PPP.

弱光条件下的场景理解是一个具有挑战性的问题。这是由于摄像机捕获的光子数量较少，因此信噪比 (SNR) 较低。单光子相机 (SPC) 是一种新兴的传感方式，能够以高灵敏度捕捉图像。尽管具有最小的读取噪声，但在光子匮乏条件下由SPC捕获的图像仍然受到强散粒噪声的影响，从而妨碍了可靠的场景推断。我们提出光子尺度空间 (photon scale space) ——一组高信噪比的图像，它们跨越了每像素光子 (PPP) 级别的广泛范围（但场景内容相同），作为在低光子通量图像上训练推断模型的指南。我们开发了训练技术，可以在特征表示空间中使具有不同照明级别的图像更接近彼此。其关键思想是，在训练期间拥有不同亮度水平的光谱能够实现有效的引导，并提高即使在极端噪声情况下对散粒噪声的鲁棒性。基于所提出的方法，我们通过仿真和SPAD相机的实际实验证明，在超微光照下，高性能的推理任务，如图像分类和单目深度估计，低至<1 PPP。

The backbone of traditional CNN classifier is generally considered as a feature extractor, followed by a linear layer which performs the classification. We propose a novel loss function, termed as CAM-loss, to constrain the embedded feature maps with the class activation maps (CAMs) which indicate the spatially discriminative regions of an image for particular categories. CAM-loss drives the backbone to express the features of target category and suppress the features of non-target categories or background, so as to obtain more discriminative feature representations. It can be simply applied in any CNN architecture with neglectable additional parameters and calculations. Experimental results show that CAM-loss is applicable to a variety of network structures and can be combined with mainstream regularization methods to improve the performance of image classification. The strong generalization ability of CAM-loss is validated in the transfer learning and few shot learning tasks. Based on CAM-loss, we also propose a novel CAAM-CAM matching knowledge distillation method. This method directly uses the CAM generated by the teacher network to supervise the CAAM generated by the student network, which effectively improves the accuracy and convergence rate of the student network.

传统的CNN分类器的主干通常被认为是一个特征提取器，然后是一个执行分类的线性层。我们提出了一种新的损失函数，称为CAM损失，用类激活映射（CAM）来约束嵌入的特征映射，该类激活映射表示特定类别图像的空间分辨区域。CAM丢失驱动主干表达目标类别的特征，抑制非目标类别或背景的特征，从而获得更具区分性的特征表示。它可以简单地应用于任何CNN体系结构中，并具有可忽略的附加参数和计算。实验结果表明，CAM-loss适用于多种网络结构，可以与主流正则化方法相结合，提高图像分类性能。在迁移学习和少镜头学习任务中，验证了CAM损失的强泛化能力。基于CAM损失，提出了一种新的CAAM-CAM匹配知识提取方法。该方法直接利用教师网络生成的CAM对学生网络生成的CAAM进行监控，有效地提高了学生网络的精度和收敛速度。

Data augmentation is vital for deep learning neural networks. By providing massive training samples, it helps to improve the generalization ability of the model. weakly supervised semantic segmentation (WSSS) is a challenging problem that has been deeply studied in recent years, conventional data augmentation approaches for WSSS usually employ geometrical transformations, random cropping, and color jittering. However, merely increasing the same contextual semantic data does not bring much gain to the networks to distinguish the objects, e.g., the correct image-level classification of "aeroplane" may be not only due to the recognition of the object itself but also its co-occurrence context like "sky", which will cause the model to focus less on the object features. To this end, we present a Context Decoupling Augmentation (CDA) method, to change the inherent context in which the objects appear and thus drive the network to remove the dependence between object instances and contextual information. To validate the effectiveness of the proposed method, extensive experiments on PASCAL VOC 2012 dataset with several alternative network architectures demonstrate that CDA can boost various popular WSSS methods to the new state-of-the-art by a large margin.

数据扩充对于深入学习神经网络至关重要。通过提供大量的训练样本，有助于提高模型的泛化能力。弱监督语义分割（WSSS）是近年来被深入研究的一个具有挑战性的问题，传统的WSSS数据增强方法通常采用几何变换、随机裁剪和颜色抖动。然而，仅仅增加相同的上下文语义数据并不会给网络带来太多的收益来区分对象，例如，“飞机”的正确图像级分类可能不仅是因为对对象本身的识别，还因为它与“天空”一样的共现上下文，这将导致模型较少关注对象特征。为此，我们提出了一种上下文解耦增强（CDA）方法，以改变对象出现的固有上下文，从而驱动网络消除对象实例与上下文信息之间的依赖关系。为了验证该方法的有效性，在PASCAL VOC 2012数据集上进行的大量实验表明，CDA可以将各种流行的WSSS方法大幅提升到最新水平。

Training temporal action detection in videos requires large amounts of labeled data, yet such annotation is expensive to collect. Incorporating unlabeled or weakly-labeled data to train action detection model could help reduce annotation cost. In this work, we first introduce the Semi-supervised Action Detection (SSAD) task with a mixture of labeled and unlabeled data and analyze different types of errors in the proposed SSAD baselines which are directly adapted from the semi-supervised classification literature. Identifying that the main source of error is action incompleteness (i.e., missing parts of actions), we alleviate it by designing an unsupervised foreground attention (UFA) module utilizing the conditional independence between foreground and background motion. Then we incorporate weakly-labeled data into SSAD and propose Omni-supervised Action Detection (OSAD) with three levels of supervision. To overcome the accompanying action-context confusion problem in OSAD baselines, an information bottleneck (IB) is designed to suppress the scene information in non-action frames while preserving the action information. We extensively benchmark against the baselines for SSAD and OSAD on our created data splits in THUMOS14 and ActivityNet1.2, and demonstrate the effectiveness of the proposed UFA and IB methods. Lastly, the benefit of our full OSAD-IB model under limited annotation budgets is shown by exploring the optimal annotation strategy for labeled, unlabeled and weakly-labeled data.

训练视频中的时间动作检测需要大量的标记数据，然而这种注释的收集成本很高。将未标记或弱标记的数据合并到训练动作检测模型中有助于降低标注成本。在这项工作中，我们首先介绍了半监督动作检测（SSAD）任务，该任务包含标记和未标记的数据，并分析了直接从半监督分类文献中改编的SSAD基线中不同类型的错误。确定错误的主要来源是动作的不完整性（即动作的缺失部分），我们通过利用前景和背景运动之间的条件独立性设计一个无监督的前景注意（UFA）模块来缓解错误。然后，我们将弱标记数据合并到SSAD中，并提出了三级监督的全监督动作检测（OSAD）。为了克服OSAD基线中伴随的动作上下文混淆问题，设计了一个信息瓶颈（IB）来抑制非动作帧中的场景信息，同时保留动作信息。在THUMOS14和ActivityNet1中创建的数据拆分上，我们广泛地对照SSAD和OSAD的基线进行基准测试。2，并证明所提出的UFA和IB方法的有效性。最后，通过探索标记、未标记和弱标记数据的最佳注释策略，展示了我们的完整OSAD-IB模型在有限注释预算下的优势。

In this paper, we propose a talking face generation method that takes an audio signal as input and a short target video clip as reference, and synthesizes a photo-realistic video of the target face with natural lip motions, head poses, and eye blinks that are in-sync with the input audio signal. We note that the synthetic face attributes include not only explicit ones such as lip motions that have high correlations with speech, but also implicit ones such as head poses and eye blinks that have only weak correlation with the input audio. To model such complicated relationships among different face attributes with input audio, we propose a FACE Implicit Attribute Learning Generative Adversarial Network (FACIAL-GAN), which integrates the phonetics-aware, context-aware, and identity-aware information to synthesize the 3D face animation with realistic motions of lips, head poses, and eye blinks. Then, our Rendering-to-video network takes the rendered face images and the attention map of eye blinks as input to generate the photo-realistic output video frames. Experimental results and user studies show our method can generate realistic talking face videos with not only synchronized lip motions, but also natural head movements and eye blinks, with better qualities than the results of state-of-the-art methods.

在本文中，我们提出了一种谈话人脸生成方法，该方法以音频信号为输入，以短目标视频剪辑为参考，合成目标人脸的照片逼真视频，具有与输入音频信号同步的自然嘴唇运动、头部姿势和眨眼。我们注意到，合成人脸属性不仅包括与语音高度相关的嘴唇运动等显式属性，还包括与输入音频相关性较弱的头部姿势和眨眼等隐式属性。为了用输入音频模拟不同人脸属性之间的复杂关系，我们提出了一种人脸隐式属性学习生成对抗网络（face-GAN），该网络集成了语音感知、上下文感知和身份感知信息，以合成

具有嘴唇、头部姿势和，还有眨眼。然后，我们的渲染到视频网络将渲染的人脸图像和眨眼的注意图作为输入，生成具有照片真实感的输出视频帧。实验结果和用户研究表明，我们的方法不仅可以生成具有同步嘴唇运动、自然头部运动和眨眼的真实对话人脸视频，其质量优于最先进的方法。

Deep convolutional neural networks (CNNs) for video denoising are typically trained with supervision, assuming the availability of clean videos. However, in many applications, such as microscopy, noiseless videos are not available. To address this, we propose an Unsupervised Deep Video Denoiser (UDVD), a CNN architecture designed to be trained exclusively with noisy data. The performance of UDVD is comparable to the supervised state-of-the-art, even when trained only on a single short noisy video. We demonstrate the promise of our approach in real-world imaging applications by denoising raw video, fluorescence-microscopy and electron-microscopy data. In contrast to many current approaches to video denoising, UDVD does not require explicit motion compensation. This is advantageous because motion compensation is computationally expensive, and can be unreliable when the input data are noisy. A gradient-based analysis reveals that UDVD automatically tracks the motion of objects in the input noisy videos. Thus, the network learns to perform implicit motion compensation, even though it is only trained for denoising.

用于视频去噪的深度卷积神经网络 (CNN) 通常在监督的情况下进行训练，前提是可以获得干净的视频。然而，在许多应用中，如显微镜，无噪音的视频是不可用的。为了解决这个问题，我们提出了一种无监督的深度视频去噪器 (UDVD)，这是一种专为训练噪声数据而设计的CNN结构。UDVD的性能可与有监督的最新技术相媲美，即使仅在一个短而嘈杂的视频上进行训练也是如此。我们通过对原始视频、荧光显微镜和电子显微镜数据进行去噪，展示了我们的方法在真实成像应用中的前景。与当前许多视频去噪方法相比，UDVD不需要明确的运动补偿。这是有利的，因为运动补偿在计算上很昂贵，并且在输入数据有噪声时可能不可靠。基于梯度的分析表明，UDVD自动跟踪输入噪声视频中对象的运动。因此，该网络学习执行隐式运动补偿，即使它仅接受去噪训练。

We present an efficient approximate message passing solver for the lifted disjoint paths problem (LDP), a natural but NP-hard model for multiple object tracking (MOT). Our tracker scales to very large instances that come from long and crowded MOT sequences. Our approximate solver enables us to process the MOT15/16/17 benchmarks without sacrificing solution quality and allows for solving MOT20, which has been out of reach up to now for LDP solvers due to its size and complexity. On all these four standard MOT benchmarks we achieve performance comparable or better than current state-of-the-art methods including a tracker based on an optimal LDP solver.

针对提升不相交路径问题 (LDP)，我们提出了一种有效的近似消息传递求解器，LDP是一种自然但NP难的多目标跟踪模型 (MOT)。我们的跟踪器可以扩展到非常大的实例，这些实例来自长而拥挤的MOT序列。我们的近似解算器使我们能够在不牺牲解决方案质量的情况下处理MOT15/16/17基准，并允许解算MOT20，由于其规模和复杂性，目前LDP解算器无法解算MOT20。在所有这四个标准MOT基准上，我们实现了与当前最先进的方法相当或更好的性能，包括基于最优LDP解算器的跟踪器。

Training vision-based Autonomous driving models is a challenging problem with enormous practical implications. one of the main challenges is the requirement of storage and processing of vast volumes of (possibly redundant) driving video data. In this paper, we study the problem of data-efficient training of autonomous driving systems. We argue that in the context of an edge-device deployment, multi-criteria online video frame subset selection is an appropriate technique for developing such frameworks. We study existing convex optimization based solutions and show that they are unable to provide solution with high weightage to loss of selected video frames. We design a novel multi-criteria online subset selection algorithm, TMCOSS, which uses a thresholded concave function of selection variables. Extensive experiments using driving simulator CARLA show that we are able to drop 80% of the frames, while succeeding to complete 100% of the episodes. We also show that TMCOSS improves performance on the crucial affordance 'Relative Angle' during turns, on inclusion of bucket-specific relative angle loss (BL), leading to selection of more frames in those parts. TMCOSS also achieves an 80% reduction in number of training video frames, on real-world videos from the standard BDD and Cityscapes datasets, for the tasks of drivable area segmentation, and semantic segmentation.

训练基于视觉的自动驾驶模型是一个具有巨大实际意义的挑战性问题。主要挑战之一是需要存储和处理大量（可能是冗余的）驱动视频数据。本文研究了自动驾驶系统的数据高效训练问题。我们认为，在边缘设备部署的环境中，多标准在线视频帧子集选择是开发此类框架的合适技术。我们研究了现有的基于凸优化的解决方案，结果表明，它们无法提供对所选视频帧丢失具有高权重的解决方案。我们设计了一种新的多准则在线子集选择算法TMCOSS，该算法使用选择变量的阈值凹函数。使用驾驶模拟器卡拉进行的大量实验表明，我们能够减少80%的画面，同时成功地完成100%的情节。我们还表明，TMCOS在转弯时的关键可承受性“相对角度”上提高了性能，包括了特定于铲斗的相对角度损失（BL），从而在这些部分选择了更多的帧。TMCOSS还可将标准BDD和Cityscapes数据集的真实视频中的训练视频帧数量减少80%，用于执行驾驶区域分割和语义分割任务。

Video objection detection is a challenging task because isolated video frames may encounter appearance deterioration, which introduces great confusion for detection. One of the popular solutions is to exploit the temporal information and enhance per-frame representation through aggregating features from neighboring frames. Despite achieving improvements in detection, existing methods focus on the selection of higher-level video frames for aggregation rather than modeling lower-level temporal relations to increase the feature representation. To address this limitation, we propose a novel solution named TF-Blender, which includes three modules: 1) Temporal relation models the relations between the current frame and its neighboring frames to preserve spatial information. 2). Feature adjustment enriches the representation of every neighboring feature map; 3) Feature blender combines outputs from the first two modules and produces stronger features for the later detection tasks. For its simplicity, TF-Blender can be effortlessly plugged into any detection network to improve detection behavior. Extensive evaluations on ImageNet VID and YouTube-VIS benchmarks indicate the performance guarantees of using TF-Blender on recent state-of-the-art methods.

视频目标检测是一项具有挑战性的任务，因为孤立的视频帧可能会出现外观退化，这给检测带来很大的混乱。一种流行的解决方案是利用时间信息，通过聚集相邻帧的特征来增强每帧表示。尽管在检测方面取得了改进，但现有的方法侧重于选择较高级别的视频帧进行聚合，而不是建模较低级别的时间关系以增加特征表示。针对这一局限性，我们提出了一种新的解决方案TF-Blender，它包括三个模块：1) 时间关系模型对当前帧与其相邻帧之间的关系进行建模，以保留空间信息。2). 特征调整丰富了相邻特征图的表示；3) 特征混合器结合前两个模块的输出，为以后的检测任务生成更强的特征。为了简单起见，

TF Blender可以轻松地插入任何检测网络，以改善检测行为。对ImageNet VID和YouTube VIS基准的广泛评估表明，在最新的最先进方法上使用TF Blender可以保证性能。

Active speaker detection requires a solid integration of multi-modal cues. While individual modalities can approximate a solution, accurate predictions can only be achieved by explicitly fusing the audio and visual features and modeling their temporal progression. Despite its inherent multi-modal nature, current methods still focus on modeling and fusing short-term audiovisual features for individual speakers, often at frame level. In this paper we present a novel approach to active speaker detection that directly addresses the multi-modal nature of the problem, and provides a straightforward strategy where independent visual features from potential speakers in the scene are assigned to a previously detected speech event. Our experiments show that, an small graph data structure built from local information, allows to approximate an instantaneous audio-visual assignment problem. Moreover, the temporal extension of this initial graph achieves a new state-of-the-art performance on the AVA-ActiveSpeaker dataset with a mAP of 88.8%.

主动说话人检测需要多模态线索的可靠集成。虽然单个模式可以近似解决方案，但准确的预测只能通过明确融合音频和视频特征并建模其时间进程来实现。尽管其固有的多模态特性，当前的方法仍然侧重于建模和融合单个扬声器的短期视听特征，通常是在帧级别。在本文中，我们提出了一种新的主动说话人检测方法，该方法直接解决了问题的多模态性质，并提供了一种简单的策略，将场景中潜在说话人的独立视觉特征分配给先前检测到的语音事件。我们的实验表明，基于局部信息构建的小型图形数据结构可以近似处理瞬时视听分配问题。此外，该初始图的时间扩展在AVA ActiveSpeaker数据集上实现了新的最新性能，映射率为88.8%。

To reduce annotation labor associated with object detection, an increasing number of studies focus on transferring the learned knowledge from a labeled source domain to another unlabeled target domain. However, existing methods assume that the labeled data are sampled from a single source domain, which ignores a more generalized scenario, where labeled data are from multiple source domains. For the more challenging task, we propose a unified Faster RCNN based framework, termed Divide-and-Merge Spindle Network (DMSN), which can simultaneously enhance domain invariance and preserve discriminative power. Specifically, the framework contains multiple source subnets and a pseudo target subnet. First, we propose a hierarchical feature alignment strategy to conduct strong and weak alignments for low- and high-level features, respectively, considering their different effects for object detection. Second, we develop a novel pseudo subnet learning algorithm to approximate optimal parameters of pseudo target subset by weighted combination of parameters in different source subnets. Finally, a consistency regularization for region proposal network is proposed to facilitate each subnet to learn more abstract invariances. Extensive experiments on different adaptation scenarios demonstrate the effectiveness of the proposed model.

为了减少与目标检测相关的注释工作，越来越多的研究关注于将所学知识从标记的源域转移到另一个未标记的目标域。然而，现有的方法假设标记的数据是从单个源域中采样的，这忽略了更一般的场景，其中标记的数据来自多个源域。对于更具挑战性的任务，我们提出了一个统一的、更快的基于RCNN的框架，称为分割合并纺锤网络（DMSN），它可以同时增强域不变性和保持区分能力。具体来说，该框架包含多个源子网和一个伪目标子网。首先，我们提出了一种分层特征对齐策略，分别对低层和高层特征进行强对齐和弱对齐，以考虑它们对目标检测的不同影响。其次，我们提出了一种新的伪子网学习算法，通过对不同源子网的参数进行加权组合来逼近伪目标子集的最优参数。最后，提出了区域建议网络的一致性正则化方法，以便于每个子网学习更多的抽象不变性。对不同适应场景的大量实验证明了该模型的有效性。

RGB-D semantic segmentation has attracted increasing attention over the past few years. Existing methods mostly employ homogeneous convolution operators to consume the RGB and depth features, ignoring their intrinsic differences. In fact, the RGB values capture the photometric appearance properties in the projected image space, while the depth feature encodes both the shape of a local geometry as well as the base (whereabout) of it in a larger context. Compared with the base, the shape probably is more inherent and has a stronger connection to the semantics, and thus is more critical for segmentation accuracy. Inspired by this observation, we introduce Shape-aware Convolutional layer (ShapeConv) for processing the depth feature, where the depth feature is firstly decomposed into a shape-component and a base-component, next two learnable weights are introduced to cooperate with them independently, and finally a convolution is applied on the re-weighted combination of these two components. ShapeConv is model-agnostic and can be easily integrated into most CNNs to replace vanilla convolutional layers for semantic segmentation. Extensive experiments on three challenging indoor RGB-D semantic segmentation benchmarks, i.e., NYU-Dv2(-13, -40), SUN RGB-D, and SID, demonstrate the effectiveness of our ShapeConv when employing it over five popular architectures. Moreover, the performance of CNNs with ShapeConv is boosted without introducing any computation and memory increase in the inference phase. The reason is that the learnt weights for balancing the importance between the shape and base components in ShapeConv become constants in the inference phase, and thus can be fused into the following convolution, resulting in a network that is identical to one with vanilla convolutional layers.

RGB-D语义分割在过去的几年中引起了越来越多的关注。现有的方法大多采用齐次卷积算子来消耗RGB和深度特征，忽略了它们的内在差异。事实上，RGB值捕获投影图像空间中的光度外观属性，而深度特征在更大的上下文中对局部几何体的形状以及其基础（位置）进行编码。与基础相比，形状可能更固有，与语义的联系更强，因此对分割精度更为关键。受这一观察结果的启发，我们引入了形状感知卷积层（ShapeConv）来处理深度特征，其中深度特征首先分解为形状分量和基础分量，然后引入两个可学习权重来独立地与它们协作，最后对这两个分量的重加权组合进行卷积。ShapeConv不依赖于模型，可以很容易地集成到大多数CNN中，以取代用于语义分割的普通卷积层。在NYU-Dv2 (-13, -40)、SUN RGB-D和SID这三个具有挑战性的室内RGB-D语义分割基准上进行的大量实验证明了我们的ShapeConv在五种流行架构上的有效性。此外，在不增加推理阶段的计算量和内存的情况下，使用ShapeConv的CNN的性能得到了提高。原因是，用于平衡ShapeConv中形状和基本组件之间重要性的学习权重在推理阶段变为常数，因此可以融合到以下卷积中，从而形成与普通卷积层相同的网络。

In recent years, the growing number of medical imaging studies is placing an ever-increasing burden on radiologists. Deep learning provides a promising solution for automatic medical image analysis and clinical decision support. However, large-scale manually labeled datasets required for training deep neural networks are difficult and expensive to obtain for medical images. The purpose of this work is to develop label-efficient multimodal medical imaging representations by leveraging radiology reports. Specifically, we propose an attention-based framework (GLORIA) for learning global and local representations by contrasting image sub-regions and words in the paired report. In addition, we propose methods to leverage the learned representations for various downstream medical image recognition tasks with limited labels. Our results demonstrate high-performance and label-efficiency for image-text retrieval, classification (finetuning and zeros-shot settings), and segmentation on different datasets.

近年来，越来越多的医学影像学研究给放射科医生带来了越来越大的负担。深度学习为自动医学图像分析和临床决策支持提供了一个有希望的解决方案。然而，训练深度神经网络所需的大规模人工标记数据集对于获取医学图像来说既困难又昂贵。这项工作的目的是通过利用放射学报告来开发标签有效的多模式医学成像表示。具体来说，我们提出了一个基于注意的框架（GLORIA），通过对配对报告中的图像子区域和单词来学习全局和局部表征。此外，我们还提出了一些方法来利用学到的表示来完成具有有

限标签的各种下游医学图像识别任务。我们的结果证明了在不同数据集上图像文本检索、分类（微调和零镜头设置）和分割的高性能和标签效率。

Humans perform co-saliency detection by first summarizing the consensus knowledge in the whole group and then searching corresponding objects in each image. Previous methods usually lack robustness, scalability, or stability for the first process and simply fuse consensus features with image features for the second process. In this paper, we propose a novel consensus-aware dynamic convolution model to explicitly and effectively perform the "summarize and search" process. To summarize consensus image features, we first summarize robust features for every single image using an effective pooling method and then aggregate cross-image consensus cues via the self-attention mechanism. By doing this, our model meets the scalability and stability requirements. Next, we generate dynamic kernels from consensus features to encode the summarized consensus knowledge. Two kinds of kernels are generated in a supplementary way to summarize fine-grained image-specific consensus object cues and the coarse group-wise common knowledge, respectively. Then, we can effectively perform object searching by employing dynamic convolution at multiple scales. Besides, a novel and effective data synthesis method is also proposed to train our network. Experimental results on four benchmark datasets verify the effectiveness of our proposed method. Our code and saliency maps are available at <https://github.com/nzhang/CADC>.

人类通过首先总结整个群体中的一致性知识，然后在每个图像中搜索相应的对象来执行共显著性检测。以前的方法通常在第一个过程中缺乏健壮性、可伸缩性或稳定性，而在第二个过程中只是简单地将一致性特征与图像特征融合。在本文中，我们提出了一种新的共识感知动态卷积模型，以明确有效地执行“总结和搜索”过程。为了总结一致性图像特征，我们首先使用有效的池方法总结每个图像的鲁棒性特征，然后通过自我注意机制聚合跨图像一致性线索。通过这样做，我们的模型满足了可伸缩性和稳定性的要求。接下来，我们从一致性特征生成动态核，对总结的一致性知识进行编码。以一种补充的方式生成两种内核，分别用于总结细粒度图像特定的一致性对象线索和粗糙的群体公共知识。然后，我们可以在多个尺度上利用动态卷积有效地进行目标搜索。此外，还提出了一种新颖有效的数据合成方法来训练我们的网络。在四个基准数据集上的实验结果验证了该方法的有效性。我们的代码和显著性地图可在<https://github.com/nzhang/CADC>.

Change captioning is the task of identifying the change and describing it with a concise caption. Despite recent advancements, filtering out insignificant changes still remains as a challenge. Namely, images from different camera perspectives can cause issues; a mere change in viewpoint should be disregarded while still capturing the actual changes. In order to tackle this problem, we present a new Viewpoint-Agnostic change captioning network with Cycle Consistency (VACC) that requires only one image each for the before and after scene, without depending on any other information. We achieve this by devising a new difference encoder module which can encode viewpoint information and model the difference more effectively. In addition, we propose a cycle consistency module that can potentially improve the performance of any change captioning networks in general by matching the composite feature of the generated caption and before image with the after image feature. We evaluate the performance of our proposed model across three datasets for change captioning, including a novel dataset we introduce here that contains images with changes under extreme viewpoint shifts. Through our experiments, we show the excellence of our method with respect to the CIDEr, BLEU-4, METEOR and SPICE scores. Moreover, we demonstrate that attaching our proposed cycle consistency module yields a performance boost for existing change captioning networks, even with varying image encoding mechanisms.

更改标题是识别更改并用简洁的标题描述它的任务。尽管最近取得了一些进展，筛选出无关紧要的变化仍然是一项挑战。即，来自不同相机视角的图像可能会导致问题；在捕捉实际变化的同时，应忽略仅仅是观点上的变化。为了解决这个问题，我们提出了一种新的视点不可知的循环一致性变化字幕网络

(VACC) ，该网络在不依赖任何其他信息的情况下，前后场景只需要一幅图像。我们通过设计一个新的差分编码器模块来实现这一点，该模块可以对视点信息进行编码，并更有效地对差分进行建模。此外，我们还提出了一个循环一致性模块，通过将生成的字幕和前图像的复合特征与后图像特征相匹配，该模块可以潜在地提高任何更改字幕网络的性能。我们评估了我们提出的模型在三个变更字幕数据集上的性能，包括我们在这里介绍的一个新的数据集，该数据集包含在极端视点移动下发生变化的图像。通过我们的实验，我们展示了我们的方法在苹果酒、BLEU-4、流星和香料评分方面的优越性。此外，我们还证明了附加我们提出的循环一致性模块可以提高现有更改字幕网络的性能，即使使用不同的图像编码机制。

Video portraits relighting is critical in user-facing human photography, especially for immersive VR/AR experience. Recent advances still fail to recover consistent relit result under dynamic illuminations from monocular RGB stream, suffering from the lack of video consistency supervision. In this paper, we propose a neural approach for real-time, high-quality and coherent video portrait relighting, which jointly models the semantic, temporal and lighting consistency using a new dynamic OLAT dataset. We propose a hybrid structure and lighting disentanglement in an encoder-decoder architecture, which combines a multi-task and adversarial training strategy for semantic-aware consistency modeling. We adopt a temporal modeling scheme via flow-based supervision to encode the conjugated temporal consistency in a cross manner. We also propose a lighting sampling strategy to model the illumination consistency and mutation for natural portrait light manipulation in real-world. Extensive experiments demonstrate the effectiveness of our approach for consistent video portrait light-editing and relighting, even using mobile computing.

视频肖像重新照明在面向用户的人体摄影中至关重要，特别是对于沉浸式VR/AR体验。由于缺乏视频一致性监控，最近的进展仍然无法在单目RGB流的动态照明下恢复一致的relit结果。在本文中，我们提出了一种实时、高质量和连贯的视频肖像重新照明的神经方法，该方法使用一个新的动态OLAT数据集联合建模语义、时间和照明一致性。我们提出了一种编码器-解码器结构中的混合结构和光照解纠缠，该结构结合了多任务和对抗性训练策略，用于语义感知一致性建模。我们采用基于流的监控的时态建模方案，以交叉方式对共轭时态一致性进行编码。我们还提出了一种光照采样策略来模拟真实世界中自然人像光操作的光照一致性和突变。大量的实验证明了我们的方法对于一致的视频人像光编辑和重新照明的有效性，即使使用移动计算。

Few-shot semantic segmentation (FSS) is an important task for novel (unseen) object segmentation under the data-scarcity scenario. However, most FSS methods rely on unidirectional feature aggregation, e.g., from support prototypes to get the query prediction, and from high-resolution features to guide the low-resolution ones. This usually fails to fully capture the cross-resolution feature relationships and thus leads to inaccurate estimates of the query objects. To resolve the above dilemma, we propose a cyclic memory network (CMN) to directly learn to read abundant support information from all resolution features in a cyclic manner. Specifically, we first generate N pairs (key and value) of multi-resolution query features guided by the support feature and its mask. Next, we circularly take one pair of these features as the query to be segmented, and the rest N-1 pairs are written into an external memory accordingly, i.e., this leave-one-out process is conducted for N times. In each cycle, the query feature is updated by collaboratively matching its key and value with the memory, which can elegantly cover all the spatial locations from different resolutions. Furthermore, we incorporate the query feature re-adding and the query feature recursive updating mechanisms into the memory reading operation. CMN, equipped with these merits, can thus capture cross-resolution relationships and better handle the object appearance and scale variations in FSS. Experiments on PASCAL-5i and COCO-20i well validate the effectiveness of our model for FSS.

少镜头语义分割 (FSS) 是数据稀缺场景下新的（看不见的）对象分割的一项重要任务。然而，大多数 FSS 方法依赖于单向特征聚合，例如，从支持原型获得查询预测，从高分辨率特征指导低分辨率特征。这通常无法完全捕获跨分辨率特征关系，从而导致对查询对象的估计不准确。为了解决上述难题，我们提出了一种循环存储网络 (CMN)，以直接学习以循环方式从所有分辨率特征中读取丰富的支持信息。具体来说，我们首先在支持特征及其掩码的指导下生成N对多分辨率查询特征（键和值）。接下来，我们循环地将这些特征中的一对作为要分割的查询，其余的N-1对被相应地写入外部存储器中，也就是说，这一遗漏过程被执行N次。在每个周期中，通过将查询特征的键和值与内存协同匹配来更新查询特征，内存可以优雅地覆盖不同分辨率的所有空间位置。此外，我们还将查询特征重新添加和查询特征递归更新机制合并到内存读取操作中。CMN具有这些优点，因此可以捕获交叉分辨率关系，更好地处理FSS中的对象外观和比例变化。在PASCAL-5i和COCO-20i上的实验很好地验证了我们的FSS模型的有效性。

We address the problem of learning to segment actions from weakly-annotated videos, i.e., videos accompanied by transcripts (ordered list of actions). We propose a framework in which we model actions with a union of low-dimensional subspaces, learn the subspaces using transcripts and refine video features that tend themselves to action subspaces. To do so, we design an architecture consisting of a Union-of-Subspace Network, which is an ensemble of autoencoders, each modeling a low-dimensional action subspace and can capture variations of an action within and across videos. For learning, at each iteration, we generate positive and negative soft alignment matrices using the segmentations from the previous iteration, which we use for discriminative training of our model. To regularize the learning, we introduce a constraint loss that prevents imbalanced segmentations and enforces relatively similar duration of each action across videos. To have a real-time inference, we develop a hierarchical segmentation framework that uses subset selection to find representative transcripts and hierarchically align a test video with increasingly refined representative transcripts. Our experiments on three datasets show that our method improves the state-of-the-art action segmentation and alignment, while speeding up the inference time by a factor of 4 to 13.

我们解决了学习从弱注释视频中分割动作的问题，即带有转录本的视频（有序动作列表）。我们提出了一个框架，在该框架中，我们用低维子空间的结合来建模动作，使用转录本学习子空间，并细化适合动作子空间的视频特征。为此，我们设计了一个由子空间网络的联合组成的体系结构，它是一个自动编码器的集合，每个自动编码器都建模一个低维动作子空间，并且可以捕获视频中和视频之间动作的变化。

对于学习，在每次迭代中，我们使用前一次迭代的分段生成正和负软对齐矩阵，我们使用这些分段对模型进行区分性训练。为了规范学习，我们引入了一个约束损失，它可以防止不平衡的分割，并在视频中强制执行每个动作相对相似的持续时间。为了进行实时推断，我们开发了一个分层分割框架，该框架使用子集选择来查找代表性转录本，并将测试视频与日益细化的代表性转录本分层对齐。我们在三个数据集上的实验表明，我们的方法改进了最先进的动作分割和对齐，同时将推理时间加快了4到13倍。

Most recent transformer-based models show impressive performance on vision tasks, even better than Convolution Neural Networks (CNN). In this work, we present a novel, flexible, and effective transformer-based model for high-quality instance segmentation. The proposed method, Segmenting objects with TRAnsformers (SOTR), simplifies the segmentation pipeline, building on an alternative CNN backbone appended with two parallel subtasks: (1) predicting per-instance category via transformer and (2) dynamically generating segmentation mask with the multi-level upsampling module. SOTR can effectively extract lower-level feature representations and capture long-range context dependencies by Feature Pyramid Network (FPN) and twin transformer, respectively. Meanwhile, compared with the original transformer, the proposed twin transformer is timeand resource-efficient since only a row and a column attention are involved to encode pixels. Moreover, SOTR is easy to be incorporated with various CNN backbones and transformer model variants to make considerable improvements for the segmentation accuracy and training convergence. Extensive experiments show that our SOTR performs well on the MS COCO dataset and surpasses state-of-the-art instance segmentation approaches. We hope our simple but strong framework could serve as a preferment baseline for instance-level recognition. Our code is available at <https://github.com/easton-cau/SOTR>.

最新的基于变压器的模型在视觉任务上表现出令人印象深刻的性能，甚至比卷积神经网络（CNN）更好。在这项工作中，我们提出了一种新颖、灵活、有效的基于转换器的高质量实例分割模型。所提出的利用转换器分割对象（SOTR）的方法简化了分割管道，建立在一个附加有两个并行子任务的CNN主干上：（1）通过转换器预测每个实例的类别；（2）使用多级上采样模块动态生成分割掩码。SOTR可以分别通过特征金字塔网络（FPN）和twin transformer有效地提取底层特征表示，并捕获长期上下文依赖。同时，与原转换器相比，该双转换器只需注意一行和一列就可以对像素进行编码，因此具有时间和资源效率。此外，SOTR易于与各种CNN主干和变压器模型变体结合，从而大大提高分割精度和训练收敛性。大量实验表明，我们的SOTR在MS COCO数据集上表现良好，超过了最先进的实例分割方法。我们希望我们简单但强大的框架可以作为实例级识别的首选基线。我们的代码可在<https://github.com/easton-cau/SOTR>。

Hypergraph matching is a useful tool to find feature correspondence by considering higher-order structural information. Recently, the employment of deep learning has made great progress in the matching of graphs, suggesting its potential for hypergraphs. Hence, in this paper, we present the first, to our best knowledge, unified hypergraph neural network (HNN) solution for hypergraph matching. Specifically, given two hypergraphs to be matched, we first construct an association hypergraph over them and convert the hypergraph matching problem into a node classification problem on the association hypergraph. Then, we design a novel hypergraph neural network to effectively solve the node classification problem. Being end-to-end trainable, our proposed method, named HNN-HM, jointly learns all its components with improved optimization. For evaluation, HNN-HM is tested on various benchmarks and shows a clear advantage over state-of-the-arts.

超图匹配是一种利用高阶结构信息寻找特征对应关系的有效工具。近年来，深度学习的应用在图的匹配方面取得了很大的进展，这表明了它在超图方面的潜力。因此，在本文中，我们提出了第一个，据我们所知，超图匹配的统一超图神经网络（HNN）解决方案。具体地说，给定两个要匹配的超图，我们首先在其上构造一个关联超图，并将超图匹配问题转化为关联超图上的节点分类问题。然后，我们设计了一

种新的超图神经网络来有效地解决节点分类问题。由于具有端到端的可训练性，我们提出的方法HNN-HM通过改进的优化联合学习其所有组件。对于评估，HNN-HM在各种基准上进行了测试，并显示出明显优于现有技术的优势。

Reconstructing delicate geometric details with consumer RGB-D sensors is challenging due to sensor depth and poses uncertainties. To tackle this problem, we propose a unique geometry-guided fusion framework: 1) First, we characterize fusion correspondences with the geodesic curves derived from the mass transport problem, also known as the Monge-Kantorovich problem. Compared with the depth map back-projection methods, the geodesic curves reveal the geometric structures of the local surface. 2) Moving the points along the geodesic curves is the core of our fusion approach, guided by local geometric properties, i.e., Gaussian curvature and mean curvature. Compared with the state-of-the-art methods, our novel geometry-guided displacement interpolation fully utilizes the meaningful geometric features of the local surface. It makes the reconstruction accuracy and completeness improved. Finally, a significant number of experimental results on real object data verify the superior performance of the proposed method. Our technique achieves the most delicate geometric details on thin objects for which the original depth map back-projection fusion scheme suffers from severe artifacts (See Fig.1).

由于传感器深度和不确定性，使用消费RGB-D传感器重建精细的几何细节具有挑战性。为了解决这个问题，我们提出了一个独特的几何引导的融合框架：1) 首先，我们用从质量传输问题（也称为Monge-Kantorovich问题）导出的测地曲线来描述融合对应关系。与深度图反投影方法相比，测地线曲线揭示了局部曲面的几何结构。2) 沿测地线曲线移动点是我们融合方法的核心，由局部几何特性（即高斯曲率和平均曲率）指导。与最新的方法相比，我们的新几何引导位移插值充分利用了局部曲面有意义的几何特征。提高了重建的精度和完整性。最后，对真实对象数据的大量实验结果验证了该方法的优越性能。我们的技术在薄对象上实现了最精细的几何细节，原始深度贴图反投影融合方案存在严重伪影（见图1）。

In this paper, we propose a frequency-aware spatiotemporal transformers for deep In this paper, we propose a Frequency-Aware Spatiotemporal Transformer (FAST) for video inpainting detection, which aims to simultaneously mine the traces of video inpainting from spatial, temporal, and frequency domains. Unlike existing deep video inpainting detection methods that usually rely on hand-designed attention modules and memory mechanism, the proposed FAST have innate global self-attention mechanisms to capture the long-range relations. While existing video inpainting methods usually explore the spatial and temporal connections in a video, our method employs a spatiotemporal transformer framework to detect the spatial connections between patches and temporal dependency between frames. As the inpainted videos usually lack high frequency details, the proposed FAST simultaneously exploits the frequency domain information with a specifically designed decoder. Extensive experimental results demonstrate that our approach achieves very competitive performance and generalizes well.

在本文中，我们提出了一种频率感知时空变换器，用于视频修复检测。在本文中，我们提出了一种用于视频修复检测的频率感知时空变换器（FAST），其目的是从空间、时间和频率域同时挖掘视频修复痕迹。与现有的深度视频修复检测方法通常依赖于手工设计的注意模块和记忆机制不同，本文提出的FAST具有先天的全局自我注意机制来捕获长程关系。虽然现有的视频修复方法通常探索视频中的空间和时间连接，但我们的方法采用时空变换框架来检测补丁之间的空间连接和帧之间的时间依赖性。由于修复后的视频通常缺乏高频细节，本文提出的快速算法通过专门设计的解码器同时利用频域信息。大量的实验结果表明，我们的方法取得了非常有竞争力的性能和推广良好。

Most existing human matting algorithms tried to separate pure human-only foreground from the background. In this paper, we propose a Virtual Multi-modality Foreground Matting (VMFM) method to learn human-object interactive foreground (human and objects interacted with him or her) from a raw RGB image. The VMFM method requires no additional inputs, e.g. trimap or known background. We reformulate foreground matting as a self-supervised multi-modality problem: factor each input image into estimated depth map, segmentation mask, and interaction heatmap using three auto-encoders. In order to fully utilize the characteristics of each modality, we first train a dual encoder-to-decoder network to estimate the same alpha matte. Then we introduce a self-supervised method: Complementary Learning(CL) to predict deviation probability map and exchange reliable gradients across modalities without label. We conducted extensive experiments to analyze the effectiveness of each modality and the significance of different components in complementary learning. We demonstrate that our model outperforms the state-of-the-art methods.

大多数现有的人像抠图算法都试图将纯人像前景与背景分开。在本文中，我们提出了一种虚拟多模态前景抠图（VMFM）方法，从原始RGB图像中学习人机交互前景（人和与他或她交互的对象）。VMFM方法不需要额外输入，例如trimap或已知背景。我们将前景抠图转化为一个自我监督的多模态问题：使用三个自动编码器将每个输入图像分解为估计深度图、分割遮罩和交互热图。为了充分利用每个模态的特性，我们首先训练一个双编码器-解码器网络来估计相同的alpha蒙版。然后，我们引入了一种自我监督的方法：互补学习（CL）来预测偏差概率图，并在没有标签的模式之间交换可靠的梯度。我们进行了广泛的实验来分析每种形式的有效性以及互补学习中不同成分的重要性。我们证明了我们的模型优于最先进的方法。

The classification and regression head are both indispensable components to build up a dense object detector, which are usually supervised by the same training samples and thus expected to have consistency with each other for detecting objects accurately in final detection pipelines. In this paper, we break the convention of the same training samples for these two heads in dense detectors and explore a novel supervisory paradigm, termed as Mutual Supervision (MuSu), to respectively and mutually assign training samples for the classification and regression head to ensure this consistency. MuSu defines training samples for the regression head mainly based on classification predicting scores and in turn, defines samples for the classification head based on localization scores from the regression head. Experimental results show that the convergence of detectors trained by this mutual supervision is guaranteed and the effectiveness of the proposed method is verified on the challenging MS COCO benchmark. We also find that tiling more anchors at the same location benefits detectors and leads to further improvements under this training scheme. We hope this work can inspire further researches on the interaction of the classification and regression task in detection and the supervision paradigm for detectors, especially separately for these two heads.

分类和回归头都是构建密集目标检测器必不可少的组成部分，它们通常由相同的训练样本进行监督，因此在最终的检测管道中，为了准确地检测目标，期望彼此具有一致性。在本文中，我们打破了密集检测器中两个头的训练样本相同的惯例，探索了一种新的监督范式，称为相互监督（MuSu），分别并相互分配分类和回归头的训练样本，以确保这种一致性。MuSu主要根据分类预测分数为回归头定义训练样本，然后根据回归头的本地化分数为分类头定义样本。实验结果表明，通过这种相互监督训练的检测器的收敛性得到了保证，并且在具有挑战性的MS-COCO基准上验证了该方法的有效性。我们还发现，在同一位置铺设更多锚有利于探测器，并导致在该培训计划下的进一步改进。我们希望这项工作能够对进一步研究检测中的分类和回归任务与检测器监督范式之间的相互作用，特别是对这两个头部的相互作用起到启发作用。

In this paper we consider the epipolar geometry between orthographic and perspective cameras. We generalize many of the classical results for the perspective essential matrix to this setting and derive novel minimal solvers, not only for the calibrated case, but also for partially calibrated and non-central camera setups. While orthographic cameras might seem exotic, they occur naturally in many applications. They can e.g. model 2D maps (such as floor plans), aerial/satellite photography and even approximate narrow field-of-view cameras (e.g. from telephoto lenses). In our experiments we highlight various applications of the developed theory and solvers, including Radar-Camera calibration and aligning Structure-from-Motion models to aerial or satellite images.

在本文中，我们考虑正极和透视相机之间的极几何。我们将透视基本矩阵的许多经典结果推广到此设置，并推导出新的最小解算器，不仅适用于校准情况，也适用于部分校准和非中心相机设置。虽然正交相机可能看起来很奇特，但它们在许多应用中自然出现。例如，它们可以模拟2D地图（如平面图）、航空/卫星摄影，甚至可以模拟窄视场摄像机（如长焦镜头）。在我们的实验中，我们重点介绍了开发的理论和求解器的各种应用，包括雷达相机校准和从运动模型到航空或卫星图像的结构对齐。

Annotation burden has become one of the biggest barriers to semantic segmentation. Approaches based on click-level annotations have therefore attracted increasing attention due to their superior trade-off between supervision and annotation cost. In this paper, we propose seminar learning, a new learning paradigm for semantic segmentation with click-level supervision. The fundamental rationale of seminar learning is to leverage the knowledge from different networks to compensate for insufficient information provided in click-level annotations. Mimicking a seminar, our seminar learning involves a teacher-student and a student-student module, where a student can learn from both skillful teachers and other students. The teacher-student module uses a teacher network based on the exponential moving average to guide the training of the student network. In the student-student module, heterogeneous pseudo-labels are proposed to bridge the transfer of knowledge among students to enhance each other's performance. Experimental results demonstrate the effectiveness of seminar learning, which achieves the new state-of-the-art performance of 72.51% (mIoU), surpassing previous methods by a large margin of up to 16.88% on the Pascal VOC 2012 dataset.

注释负担已成为语义分割的最大障碍之一。因此，基于点击级注释的方法由于其在监督和注释成本之间的优势权衡而越来越受到关注。在本文中，我们提出了研讨会学习，一种新的学习范式的语义分割与点击级监督。研讨会学习的基本原理是利用来自不同网络的知识来弥补单击级别注释中提供的信息不足。模拟研讨会，我们的研讨会学习包括师生和学生模块，学生可以从熟练的老师和其他学生那里学习。师生模块使用基于指数移动平均的教师网络来指导学生网络的培训。在学生-学生模块中，提出了异构伪标签来桥接学生之间的知识转移，以提高彼此的表现。实验结果证明了研讨会学习的有效性，它实现了72.51% (mIoU) 的最新性能，在Pascal VOC 2012数据集上比以前的方法大幅度提高了16.88%。

We present Retrieve in Style (RIS), an unsupervised framework for facial feature transfer and retrieval on real images. Recent work shows capabilities of transferring local facial features by capitalizing on the disentanglement property of the StyleGAN latent space. RIS improves existing art on the following: 1) Introducing more effective feature disentanglement to allow for challenging transfers (i.e., hair, pose) that were not shown possible in SoTA methods. 2) Eliminating the need for per-image hyperparameter tuning, and for computing a catalog over a large batch of images. 3) Enabling fine-grained face retrieval using disentangled facial features (e.g., eyes). To our best knowledge, this is the first work to retrieve face images at this fine level. 4) Demonstrating robust, natural editing on real images. Our qualitative and quantitative analyses show RIS achieves both high-fidelity feature transfers and accurate fine-grained retrievals on real images. We also discuss the responsible applications of RIS. Our code is available at <https://github.com/mchong6/RetrieveInStyle>.

我们提出了风格检索 (RIS) , 一个无监督的框架, 人脸特征转移和检索的真实图像。最近的研究表明, 利用StyleGAN潜在空间的解纠缠特性, 可以转移局部面部特征。RIS在以下方面改进了现有技术: 1) 引入更有效的特征分离, 以允许SoTA方法中无法显示的具有挑战性的转移(即头发、姿势)。2) 无需对每幅图像进行超参数调整, 也无需对大量图像计算目录。3) 使用分离的面部特征(例如眼睛)实现细粒度面部检索。据我们所知, 这是第一个在这种精细级别上检索人脸图像的工作。4) 演示对真实图像进行稳健、自然的编辑。我们的定性和定量分析表明, RIS在真实图像上实现了高保真特征传输和精确的细粒度检索。我们还讨论了RIS的负责任应用。我们的代码可在<https://github.com/mchong6/RetrieveInStyle>。

In this paper, we study a novel meta aggregation scheme towards binarizing graph neural networks (GNNs). We begin by developing a vanilla 1-bit GNN framework that binarizes both the GNN parameters and the graph features. Despite the lightweight architecture, we observed that this vanilla framework suffered from insufficient discriminative power in distinguishing graph topologies, leading to a dramatic drop in performance. This discovery motivates us to devise meta aggregators to improve the expressive power of vanilla binarized GNNs, of which the aggregation schemes can be adaptively changed in a learnable manner based on the binarized features. Towards this end, we propose two dedicated forms of meta neighborhood aggregators, an exclusive meta aggregator termed as Greedy Gumbel Neighborhood Aggregator (GNA), and a diffused meta aggregator termed as Adaptable Hybrid Neighborhood Aggregator (ANA). GNA learns to exclusively pick one single optimal aggregator from a pool of candidates, while ANA learns a hybrid aggregation behavior to simultaneously retain the benefits of several individual aggregators. Furthermore, the proposed meta aggregators may readily serve as a generic plugin module into existing full-precision GNNs. Experiments across various domains demonstrate that the proposed method yields results superior to the state of the art.

本文研究了一种新的二值化图神经网络 (GNNs) 元聚集方案。我们首先开发一个普通的1位GNN框架, 该框架对GNN参数和图形特性进行二值化。尽管采用了轻量级架构, 但我们发现这种普通框架在区分图形拓扑方面的辨别能力不足, 导致性能急剧下降。这一发现促使我们设计元聚合器来提高普通二值化GNN的表达能力, 其中的聚合方案可以基于二值化特征以可学习的方式自适应地改变。为此, 我们提出了两种专用的元邻域聚合器, 一种称为贪婪Gumbel邻域聚合器 (GNA) 的专用元聚合器和一种称为自适应混合邻域聚合器 (ANA) 的扩散元聚合器。GNA学习从候选池中专门挑选一个最佳聚合器, 而ANA学习混合聚合行为以同时保留多个单独聚合器的优点。此外, 所提出的元聚合器可以作为现有全精度GNN的通用插件模块。不同领域的实验表明, 所提出的方法产生的结果优于现有技术。

Aiming at discovering and locating most distinctive objects from visual scenes, salient object detection (SOD) plays an essential role in various computer vision systems. Coming to the era of high resolution, SOD methods are facing new challenges. The major limitation of previous methods is that they try to identify the salient regions and estimate the accurate objects boundaries simultaneously with a single regression task at low-resolution. This practice ignores the inherent difference between the two difficult problems, resulting in poor detection quality. In this paper, we propose a novel deep learning framework for high-resolution SOD task, which disentangles the task into a low-resolution saliency classification network (LRSCN) and a high-resolution refinement network (HRRN). As a pixel-wise classification task, LRSCN is designed to capture sufficient semantics at low-resolution to identify the definite salient, background and uncertain image regions. HRRN is a regression task, which aims at accurately refining the saliency value of pixels in the uncertain region to preserve a clear object boundary at high-resolution with limited GPU memory. It is worth noting that by introducing uncertainty into the training process, our HRRN can well address the high-resolution refinement task without using any high-resolution training data. Extensive experiments on high-resolution saliency datasets as well as some widely used saliency benchmarks show that the proposed method achieves superior performance compared to the state-of-the-art methods.

显著目标检测 (SOD) 旨在从视觉场景中发现和定位最具特征的目标，在各种计算机视觉系统中起着至关重要的作用。进入高分辨率时代，超氧化物歧化酶方法面临着新的挑战。以前的方法的主要局限性在于，它们试图在低分辨率下通过单一回归任务同时识别显著区域并估计精确的对象边界。这种做法忽视了这两个难题之间的固有差异，导致检测质量差。在本文中，我们提出了一种新的高分辨率SOD任务深度学习框架，该框架将任务分解为低分辨率显著性分类网络 (LRSCN) 和高分辨率细化网络 (HRRN)。LRSCN作为一种像素级的分类任务，设计用于在低分辨率下获取足够的语义，以识别明确的显著、背景和不确定的图像区域。HRRN是一项回归任务，其目的是在有限的GPU内存下精确细化不确定区域中像素的显著性值，以在高分辨率下保持清晰的对象边界。值得注意的是，通过在训练过程中引入不确定性，我们的HRRN可以很好地解决高分辨率细化任务，而无需使用任何高分辨率训练数据。在高分辨率显著性数据集和一些广泛使用的显著性基准上的大量实验表明，与现有的方法相比，该方法具有更高的性能。

Generalized zero-shot learning (GZSL) has achieved significant progress, with many efforts dedicated to overcoming the problems of visual-semantic domain gaps and seen-unseen bias. However, most existing methods directly use feature extraction models trained on ImageNet alone, ignoring the cross-dataset bias between ImageNet and GZSL benchmarks. Such a bias inevitably results in poor-quality visual features for GZSL tasks, which potentially limits the recognition performance on both seen and unseen classes. In this paper, we propose a simple yet effective GZSL method, termed feature refinement for generalized zero-shot learning (FREE), to tackle the above problem. FREE employs a feature refinement (FR) module that incorporates semantic-visual mapping into a unified generative model to refine the visual features of seen and unseen class samples. Furthermore, we propose a self-adaptive margin center loss (SAMC-loss) that cooperates with a semantic cycle-consistency loss to guide FR to learn class- and semantically-relevant representations, and concatenate the features in FR to extract the fully refined features. Extensive experiments on five benchmark datasets demonstrate the significant performance gain of FREE over current state-of-the-art methods and its baseline. The code is available at <https://github.com/shiming-chen/FREE>.

广义零镜头学习 (GZSL) 已经取得了重大进展，许多努力致力于克服视觉语义领域的差距和看不见的偏见问题。然而，大多数现有方法直接使用仅在ImageNet上训练的特征提取模型，忽略了ImageNet和GZSL基准之间的跨数据集偏差。这种偏见不可避免地会导致GZSL任务的视觉特征质量低下，这可能会限制可见类和不可见类的识别性能。在本文中，我们提出了一种简单而有效的GZSL方法，称为广义零炮学习 (FREE) 的特征精化方法来解决上述问题。FREE采用了一个特征细化 (FR) 模块，该模块将语义视觉映射合并到一个统一的生成模型中，以细化可见和不可见类样本的视觉特征。此外，我们提出了一种自适应边缘中心损失 (SAMC损失)，它与语义循环一致性损失相结合，引导FR学习类和语义相关的表示，并将FR中的特征连接起来以提取完全细化的特征。在五个基准数据集上进行的大量实验表明，与当前最先进的方法及其基线相比，FREE具有显著的性能增益。该守则可于<https://github.com/shimingchen/FREE>.

Global Covariance Pooling (GCP) aims at exploiting the second-order statistics of the convolutional feature. Its effectiveness has been demonstrated in boosting the classification performance of Convolutional Neural Networks (CNNs). Singular Value Decomposition (SVD) is used in GCP to compute the matrix square root. However, the approximate matrix square root calculated using Newton-Schulz iteration [??] outperforms the accurate one computed via SVD [??]. We empirically analyze the reason behind the performance gap from the perspectives of data precision and gradient smoothness. Various remedies for computing smooth SVD gradients are investigated. Based on our observation and analyses, a hybrid training protocol is proposed for SVD-based GCP meta-layers such that competitive performances can be achieved against Newton-Schulz iteration. Moreover, we propose a new GCP meta-layer that uses SVD in the forward pass, and Pade approximants in the backward propagation to compute the gradients. The proposed meta-layer has been integrated into different CNN models and achieves state-of-the-art performances on both large-scale and fine-grained datasets.

全局协方差池 (GCP) 旨在利用卷积特征的二阶统计量。其有效性已被证明在提高卷积神经网络 (CNN) 的分类性能。GCP中使用奇异值分解 (SVD) 来计算矩阵的平方根。但是，使用牛顿-舒尔茨迭代法计算的近似矩阵平方根[? ?]优于通过SVD[? ?]计算的精确值。我们从数据精度和梯度平滑度的角度实证分析了性能差距背后的原因。研究了计算光滑奇异值分解梯度的各种方法。基于我们的观察和分析，针对基于奇异值分解的GCP元层，提出了一种混合训练协议，这样可以在牛顿-舒尔茨迭代中获得有竞争力的性能。此外，我们还提出了一种新的GCP元层，它在前向传递中使用SVD，在后向传播中使用Pade近似来计算梯度。所提出的元层已集成到不同的CNN模型中，并在大规模和细粒度数据集上实现了最先进的性能。

The vast majority of modern consumer-grade cameras employ a rolling shutter mechanism, leading to image distortions if the camera moves during image acquisition. In this paper, we present a novel deep network to solve the generic rolling shutter correction problem with two consecutive frames. Our pipeline is symmetrically designed to predict the global shutter image corresponding to the intermediate time of these two frames, which is difficult for existing methods because it corresponds to a camera pose that differs most from the two frames. First, two time-symmetric dense undistortion flows are estimated by using well-established principles: pyramidal construction, warping, and cost volume processing. Then, both rolling shutter images are warped into a common global shutter one in the feature space, respectively. Finally, a symmetric consistency constraint is constructed in the image decoder to effectively aggregate the contextual cues of two rolling shutter images, thereby recovering the high-quality global shutter image. Extensive experiments with both synthetic and real data from public benchmarks demonstrate the superiority of our proposed approach over the state-of-the-art methods.

绝大多数现代消费级相机采用滚动快门机制，如果相机在图像采集过程中移动，会导致图像失真。在本文中，我们提出了一种新的深度网络来解决两个连续帧的通用滚动快门校正问题。我们的管道对称设计用于预测与这两帧的中间时间相对应的全局快门图像，这对于现有方法来说是困难的，因为它对应于与这两帧最不同的相机姿势。首先，利用成熟的原理估计两次对称密集不畸变流：金字塔结构、翘曲和成本-体积处理。然后，在特征空间中，将两幅滚动快门图像分别扭曲成一幅通用的全局快门图像。最后，在图像解码器中构造对称一致性约束，以有效地聚合两个滚动快门图像的上下文线索，从而恢复高质量的全局快门图像。对来自公共基准的合成数据和真实数据进行的大量实验表明，我们提出的方法优于最先进的方法。

We propose a novel pointwise descriptor, called DWKS, aimed at finding correspondences across two deformable shape collections. Unlike the majority of existing descriptors, rather than capturing local geometry, DWKS captures the deformation around a point within a collection in a multi-scale and informative manner. This, in turn, allows to compute inter-collection correspondences without using landmarks. To this end, we build upon the successful spectral WKS descriptors, but rather than using the Laplace-Beltrami operator, show that a similar construction can be performed on shape difference operators, that capture differences or distortion within a collection. By leveraging the collection information our descriptor facilitates difficult non-rigid shape matching tasks, even in the presence of strong partiality and significant deformations. We demonstrate the utility of our approach across a range of challenging matching problems on both meshes and point clouds. The code for this paper can be found at <https://github.com/RobinMagnet/DWKS>.

我们提出了一种新的点式描述符，称为DWKS，旨在寻找两个可变形形状集合之间的对应关系。与大多数现有描述符不同，DWK不是捕获局部几何图形，而是以多尺度和信息丰富的方式捕获集合中某个点周围的变形。这反过来又允许在不使用地标的情况下计算集合间的对应关系。为此，我们基于成功的光谱WKS描述符，而不是使用拉普拉斯-贝尔特拉米算子，表明可以对形状差异算子执行类似的构造，捕捉集合中的差异或失真。通过利用收集信息，我们的描述符有助于完成困难的非刚性形状匹配任务，即使是在存在强烈偏好和显著变形的情况下。我们展示了我们的方法在网格和点云上的一系列具有挑战性的匹配问题上的实用性。本文的代码可在<https://github.com/RobinMagnet/DWKS>.

Autoregressive models are a class of exact inference approaches with highly flexible functional forms, yielding state-of-the-art density estimates for natural images. Yet, the sequential ordering on the dimensions makes these models computationally expensive and limits their applicability to low-resolution imagery. In this work, we propose Pixel-Pyramids, a block-autoregressive approach employing a lossless pyramid decomposition with scale-specific representations to encode the joint distribution of image pixels. Crucially, it affords a sparser dependency structure compared to fully autoregressive approaches. Our PixelPyramids yield state-of-the-art results for density estimation on various image datasets, especially for high-resolution data. For CelebA-HQ 1024 x 1024, we observe that the density estimates (in terms of bits/dim) are improved to 44% of the baseline despite sampling speeds superior even to easily parallelizable flow-based models.

自回归模型是一类精确推理方法，具有高度灵活的函数形式，可产生最先进的自然图像密度估计。然而，维度上的顺序使得这些模型的计算成本很高，并且限制了它们对低分辨率图像的适用性。在这项工作中，我们提出了像素金字塔，这是一种块自回归方法，采用无损金字塔分解和特定尺度表示来编码图像像素的联合分布。关键的是，与完全自回归方法相比，它提供了更稀疏的依赖结构。我们的像素金字塔为各种图像数据集（尤其是高分辨率数据）的密度估计提供了最先进的结果。对于CelebA HQ 1024 x 1024，我们观察到，尽管采样速度甚至优于易于并行化的基于流的模型，但密度估计（以比特/暗为单位）仍提高到基线的44%。

Existing video super-resolution (SR) algorithms usually assume that the blur kernels in the degradation process are known and do not model the blur kernels in the restoration. However, this assumption does not hold for blind video SR and usually leads to over-smoothed super-resolved frames. In this paper, we propose an effective blind video SR algorithm based on deep convolutional neural networks (CNNs). Our algorithm first estimates blur kernels from low-resolution (LR) input videos. Then, with the estimated blur kernels, we develop an effective image deconvolution method based on the image formation model of blind video SR to generate intermediate latent frames so that sharp image contents can be restored well. To effectively explore the information from adjacent frames, we estimate the motion fields from LR input videos, extract features from LR videos by a feature extraction network, and warp the extracted features from LR inputs based on the motion fields. Moreover, we develop an effective sharp feature exploration method which first extracts sharp features from restored intermediate latent frames and then uses a transformation operation based on the extracted sharp features and warped features from LR inputs to generate better features for HR video restoration. We formulate the proposed algorithm into an end-to-end trainable framework and show that it performs favorably against state-of-the-art methods.

现有的视频超分辨率 (SR) 算法通常假设退化过程中的模糊核是已知的，并且在恢复过程中不对模糊核进行建模。然而，这种假设不适用于盲视频SR，通常会导致超平滑超分辨率帧。本文提出了一种基于深度卷积神经网络 (CNN) 的盲视频SR算法。我们的算法首先估计来自低分辨率 (LR) 输入视频的模糊核。然后，利用估计的模糊核，基于盲视频SR的图像形成模型，提出了一种有效的图像反卷积方法，以生成中间潜在帧，从而很好地恢复锐利的图像内容。为了有效地挖掘相邻帧的信息，我们从LR输入视频中估计运动场，通过特征提取网络从LR视频中提取特征，并基于运动场扭曲从LR输入中提取的特征。此外，我们开发了一种有效的锐化特征探索方法，该方法首先从恢复的中间潜在帧中提取锐化特征，然后使用基于从LR输入中提取的锐化特征和扭曲特征的变换操作来生成更好的HR视频恢复特征。我们将所提出的算法转化为端到端的可训练框架，并证明其性能优于现有的方法。

3D visual grounding aims at grounding a natural language description about a 3D scene, usually represented in the form of 3D point clouds, to the targeted object region. Point clouds are sparse, noisy, and contain limited semantic information compared with 2D images. These inherent limitations make the 3D visual grounding problem more challenging. In this study, we propose 2D Semantics Assisted Training (SAT) that utilizes 2D image semantics in the training stage to ease point-cloud-language joint representation learning and assist 3D visual grounding. The main idea is to learn auxiliary alignments between rich, clean 2D object representations and the corresponding objects or mentioned entities in 3D scenes. SAT takes 2D object semantics, i.e., object label, image feature, and 2D geometric feature, as the extra input in training but does not require such inputs during inference. By effectively utilizing 2D semantics in training, our approach boosts the accuracy on the Nr3D dataset from 37.7% to 49.2%, which significantly surpasses the non-SAT baseline with the identical network architecture and inference input. Our approach outperforms the state of the art by large margins on multiple 3D visual grounding datasets, i.e., +10.4% absolute accuracy on Nr3D, +9.9% on Sr3D, and +5.6% on ScanRef.

3D视觉基础旨在将3D场景的自然语言描述（通常以3D点云的形式表示）基础到目标对象区域。与二维图像相比，点云是稀疏的、有噪声的，并且包含有限的语义信息。这些固有的限制使得三维视觉接地问题更具挑战性。在这项研究中，我们提出了2D语义辅助训练 (SAT) ，该训练在训练阶段利用2D图像语义来简化点云语言的联合表示学习，并辅助3D视觉基础。其主要思想是学习丰富、清晰的2D对象表示与3D场景中相应的对象或提到的实体之间的辅助对齐。SAT将2D对象语义，即对象标签、图像特征和2D几何特征作为训练中的额外输入，但在推理过程中不需要这些输入。通过在训练中有效地利用2D语义，我们的方法将Nr3D数据集的准确率从37.7%提高到49.2%，显著超过了具有相同网络结构和推理输入的

非SAT基线。我们的方法在多个3D视觉基础数据集上表现出了巨大的优势，即Nr3D的绝对精度为+10.4%，Sr3D的绝对精度为+9.9%，ScanRef的绝对精度为+5.6%。

Adversarial robustness of deep models is pivotal in ensuring safe deployment in real world settings, but most modern defenses have narrow scope and expensive costs. In this paper, we propose a self-supervised method to detect adversarial attacks and classify them to their respective threat models, based on a linear model operating on the embeddings from a pre-trained self-supervised encoder. We use a SimCLR encoder in our experiments, since we show the SimCLR embedding distance is a good proxy for human perceptibility, enabling it to encapsulate many threat models at once. We call our method SimCat since it uses SimCLR encoder to catch and categorize various types of adversarial attacks, including L<sub>p</sub> and non-L<sub>p</sub> evasion attacks, as well as data poisonings. The simple nature of a linear classifier makes our method efficient in both time and sample complexity. For example, on SVHN, using only five pairs of clean and adversarial examples computed with a PGD-L<sub>inf</sub> attack, SimCat's detection accuracy is over 85%. Moreover, on ImageNet, using only 25 examples from each threat model, SimCat can classify eight different attack types such as PGD-L\_2, PGD-L<sub>inf</sub>, CW-L\_2, PPGD, LPA, StAdv, ReColor, and JPEG-L<sub>inf</sub>, with over 40% accuracy. On STL10 data, we apply SimCat as a defense against poisoning attacks, such as BP, CP, FC, CLBD, HTBD, halving the success rate while using only twenty total poisons for training. We find that the detectors generalize well to unseen threat models. Lastly, we investigate the performance of our detection method under adaptive attacks and further boost its robustness against such attacks via adversarial training.

深部模型的对抗性健壮性对于确保在真实环境中的安全部署至关重要，但大多数现代防御范围狭窄，成本昂贵。在本文中，我们提出了一种自监督的方法来检测对抗性攻击，并将其分类到各自的威胁模型中，该方法基于一个线性模型，该模型基于预先训练的自监督编码器的嵌入操作。我们在实验中使用了SimCLR编码器，因为我们表明SimCLR嵌入距离是人类感知能力的一个很好的代理，使其能够同时封装许多威胁模型。我们称我们的方法为SimCat，因为它使用SimCLR编码器捕获并分类各种类型的对抗性攻击，包括L<sub>p</sub>和非L<sub>p</sub>规避攻击，以及数据中毒。线性分类器的简单性质使得我们的方法在时间和样本复杂度上都是有效的。例如，在SVHN上，仅使用PGD-L<sub>inf</sub>攻击计算的五对干净和对抗性示例，SimCat的检测精度超过85%。此外，在ImageNet上，仅使用每个威胁模型中的25个示例，SimCat就可以对八种不同的攻击类型进行分类，如PGD-L\_2、PGD-L<sub>inf</sub>、CW-L\_2、PPGD、LPA、StAdv、RECLOR和JPEG-L<sub>inf</sub>，准确率超过40%。在STL10数据上，我们使用SimCat作为对中毒攻击的防御，例如BP、CP、FC、CLBD、HTBD，在只使用20种总毒药进行训练的情况下，成功率降低了一半。我们发现，检测器可以很好地推广到看不见的威胁模型。最后，我们研究了我们的检测方法在自适应攻击下的性能，并通过对抗性训练进一步增强了其对此类攻击的鲁棒性。

A large body of recent work has identified transformations in the latent spaces of generative adversarial networks (GANs) that consistently and interpretably transform generated images. But existing techniques for identifying these transformations rely on either a fixed vocabulary of pre-specified visual concepts, or on unsupervised disentanglement techniques whose alignment with human judgments about perceptual salience is unknown. This paper introduces a new method for building open-ended vocabularies of primitive visual concepts represented in a GAN's latent space. Our approach is built from three components: (1) automatic identification of perceptually salient directions based on their layer selectivity; (2) human annotation of these directions with free-form, compositional natural language descriptions; and (3) decomposition of these annotations into a visual concept vocabulary, consisting of distilled directions labeled with single words. Experiments show that concepts learned with our approach are reliable and composable--generalizing across classes, contexts, and observers, and enabling fine-grained manipulation of image style and content.

最近的大量工作已经确定了生成性对抗网络 (GAN) 潜在空间中的转换，这些网络一致且可解释地转换生成的图像。但是现有的识别这些转换的技术要么依赖于预先指定的视觉概念的固定词汇表，要么依赖于无监督的解纠缠技术，其与人类关于知觉显著性的判断的一致性是未知的。本文介绍了一种新的方法来建立开放式词汇表的原始视觉概念表示在一个甘的潜在空间。我们的方法由三部分组成：(1) 基于层选择性的感知显著方向的自动识别；(2) 用自由形式、合成自然语言描述对这些方向进行人类注释；

(3) 将这些注释分解为视觉概念词汇表，包括用单个单词标记的提炼方向。实验表明，通过我们的方法学习到的概念是可靠和可组合的——可以跨类、上下文和观察者进行概括，并支持对图像样式和内容的细粒度操作。

We propose a novel weakly supervised approach for 3D semantic segmentation on volumetric images. Unlike most existing methods that require voxel-wise densely labeled training data, our weakly-supervised CIVA-Net is the first model that only needs image-level class labels as guidance to learn accurate volumetric segmentation. Our model learns from cross-image co-occurrence for integral region generation, and explores inter-voxel affinity relations to predict segmentation with accurate boundaries. We empirically validate our model on both simulated and real cryo-ET datasets. Our experiments show that CIVA-Net achieves comparable performance to the state-of-the-art models trained with stronger supervision.

我们提出了一种新的基于体图像的弱监督三维语义分割方法。与大多数现有的需要密集标记体素训练数据的方法不同，我们的弱监督CIVA网络是第一个只需要图像级类别标签作为指导来学习精确体积分割的模型。我们的模型学习交叉图像共现来生成积分区域，并探索体素间的亲和力关系来预测具有精确边界的分割。我们在模拟和真实的cryo ET数据集上验证了我们的模型。我们的实验表明，CIVA Net的性能与经过更严格监督培训的最先进模型相当。

Geometric feature extraction is a crucial component of point cloud registration pipelines. Recent work has demonstrated how supervised learning can be leveraged to learn better and more compact 3D features. However, those approaches' reliance on ground-truth annotation limits their scalability. We propose BYOC: a self-supervised approach that learns visual and geometric features from RGB-D video without relying on ground-truth pose or correspondence. Our key observation is that randomly-initialized CNNs readily provide us with good correspondences; allowing us to bootstrap the learning of both visual and geometric features. Our approach combines classic ideas from point cloud registration with more recent representation learning approaches. We evaluate our approach on indoor scene datasets and find that our method outperforms traditional and learned descriptors, while being competitive with current state-of-the-art supervised approaches.

几何特征提取是点云配准的重要组成部分。最近的工作已经证明了如何利用监督学习来学习更好、更紧凑的3D特征。然而，这些方法对地面真相注释的依赖限制了它们的可扩展性。我们提出了BYOC：一种自监督方法，从RGB-D视频中学习视觉和几何特征，而不依赖于地面真实姿势或对应。我们的主要观察结果是，随机初始化的CNN很容易为我们提供良好的对应关系；允许我们引导视觉和几何特征的学习。我们的方法将点云注册的经典思想与最近的表示学习方法相结合。我们在室内场景数据集上评估了我们的方法，发现我们的方法优于传统和学习的描述符，同时与当前最先进的监督方法具有竞争力。

We study a crucial problem in video analysis: human-object relationship detection. The majority of previous approaches are developed only for the static image scenario, without incorporating the temporal dynamics so vital to contextualizing human-object relationships. We propose a model with Intra- and Inter-Transformers, enabling joint spatial and temporal reasoning on multiple visual concepts of objects, relationships, and human poses. We find that applying attention mechanisms among features distributed spatio-temporally greatly improves our understanding of human-object relationships. Our method is validated on two datasets, Action Genome and CAD-120-EVAR, and achieves state-of-the-art performance on both of them.

我们研究了视频分析中的一个关键问题：人-物关系检测。以前的大多数方法都是针对静态图像场景开发的，没有结合时间动态，这对于上下文化人-物关系至关重要。我们提出了一个具有内部和内部转换器的模型，支持对对象、关系和人体姿势的多个视觉概念进行联合空间和时间推理。我们发现，在时空分布的特征中应用注意机制可以极大地提高我们对人-物关系的理解。我们的方法在Action Genome和CAD-120-EVAR两个数据集上进行了验证，并在这两个数据集上实现了最先进的性能。

Inspired by the human learning principle that learning easier concepts first and then gradually paying more attention to harder ones, curriculum learning uses the non-uniform sampling of mini-batches according to the order of examples' difficulty. Just as a teacher adjusts the curriculum according to the learning progress of each student, a proper curriculum should be adapted to the current state of the model. Therefore, in contrast to recent works using a fixed curriculum, we devise a new curriculum learning method, Adaptive Curriculum Learning (Adaptive CL), adapting the difficulty of examples to the current state of the model. Specifically, we make use of the loss of the current model to adjust the difficulty score while retaining previous useful learned knowledge by KL divergence. Moreover, under a non-linear model and binary classification, we theoretically prove that the expected convergence rate of curriculum learning monotonically decreases with respect to the loss of a point regarding the optimal hypothesis, and monotonically increases with respect to the loss of a point regarding the current hypothesis. The analyses indicate that Adaptive CL could improve the convergence properties during the early stages of learning. Extensive experimental results demonstrate the superiority of the proposed approach over existing competitive curriculum learning methods.

受先学习较容易的概念，然后逐渐关注较难的概念这一人类学习原则的启发，课程学习根据示例的难度顺序使用非均匀的小批量抽样。正如教师根据每个学生的学习进度调整课程一样，适当的课程也应该适应当前的模式。因此，与最近使用固定课程的作品相比，我们设计了一种新的课程学习方法，自适应课程学习（Adaptive CL），使示例的难度适应模型的当前状态。具体来说，我们利用当前模型的损失来调整难度分数，同时通过KL散度保留以前有用的学习知识。此外，在非线性模型和二元分类下，我们从理论上证明了课程学习的预期收敛速度在最优假设下随着一个点的丢失而单调减小，在当前假设下随着一个点的丢失而单调增大。分析表明，在学习的早期阶段，自适应CL可以改善收敛性能。大量的实验结果表明，该方法优于现有的竞争性课程学习方法。

Top-k multi-label learning, which returns the top-k predicted labels from an input, has many practical applications such as image annotation, document analysis, and web search engine. However, the vulnerabilities of such algorithms with regards to dedicated adversarial perturbation attacks have not been extensively studied previously. In this work, we develop methods to create adversarial perturbations that can be used to attack top-k multi-label learning-based image annotation systems (T\_KML-AP). Our methods explicitly consider the top-k ranking relation and are based on novel loss functions. Experimental evaluations on large-scale benchmark datasets including PASCAL VOC and MS COCO demonstrate the effectiveness of our methods in reducing the performance of state-of-the-art top-k multi-label learning methods, under both untargeted and targeted attacks.

Top-k多标签学习是一种从输入中返回Top-k预测标签的学习方法，在图像标注、文档分析和web搜索引擎等领域有着广泛的应用。然而，此类算法在专用对抗性干扰攻击方面的漏洞尚未得到广泛研究。在这项工作中，我们开发了创建对抗性干扰的方法，可用于攻击基于top-k多标签学习的图像标注系统（T\_KML-AP）。我们的方法明确地考虑top-k排名关系，并基于新的损失函数。对大规模基准数据集（包括PASCAL VOC和MS COCO）的实验评估表明，我们的方法在非目标攻击和目标攻击下都能有效降低最先进的top-k多标签学习方法的性能。

Object localisation, in the context of regular images, often depicts objects like people or cars. In these images, there is typically a relatively small number of objects per class, which usually is manageable to annotate. However, outside the setting of regular images, we are often confronted with a different situation. In computational pathology, digitised tissue sections are extremely large images, whose dimensions quickly exceed 250'000x250'000 pixels, where relevant objects, such as tumour cells or lymphocytes can quickly number in the millions. Annotating them all is practically impossible and annotating sparsely a few, out of many more, is the only possibility. Unfortunately, learning from sparse annotations, or sparse-shot learning, clashes with standard supervised learning because what is not annotated is treated as a negative. However, assigning negative labels to what are true positives leads to confusion in the gradients and biased learning. To this end, we present exclusive cross-entropy, which slows down the biased learning by examining the second-order loss derivatives in order to drop the loss terms corresponding to likely biased terms. Experiments on nine datasets and two different localisation tasks, detection with YOLLO and segmentation with Unet, show that we obtain considerable improvements compared to cross-entropy or focal loss, while often reaching the best possible performance for the model with only 10-40% of annotations.

在常规图像的背景下，对象定位通常描述诸如人或汽车之类的对象。在这些图像中，每个类的对象数量通常相对较少，通常可以进行注释。然而，在常规图像的背景之外，我们经常会遇到不同的情况。在计算病理学中，数字化组织切片是非常大的图像，其尺寸迅速超过250'000x250'000像素，其中相关对象（如肿瘤细胞或淋巴细胞）可以迅速增加到数百万。把它们全部注释掉几乎是不可能的，在更多的注释中，只注释很少的一部分是唯一的可能。不幸的是，从稀疏注释或稀疏快照学习中学习与标准的监督学习相冲突，因为未注释的内容被视为负面。然而，给真正积极的东西贴上负面标签会导致梯度的混乱和有偏见的学习。为此，我们提出了排他性交叉熵，它通过检查二阶损失导数来降低与可能的有偏项对应的损失项，从而减慢有偏学习。在九个数据集和两个不同的定位任务（使用YOLLO进行检测和使用Unet进行分割）上的实验表明，与交叉熵或焦点损失相比，我们获得了相当大的改进，而对于只有10-40%注释的模型，我们通常可以达到最佳性能。

In this paper, we propose an end-to-end learning framework for event-based motion deblurring in a self-supervised manner, where real-world events are exploited to alleviate the performance degradation caused by data inconsistency. To achieve this end, optical flows are predicted from events, with which the blurry consistency and photometric consistency are exploited to enable self-supervision on the deblurring network with real-world data. Furthermore, a piece-wise linear motion model is proposed to take into account motion non-linearities and thus leads to an accurate model for the physical formation of motion blurs in the real-world scenario. Extensive evaluation on both synthetic and real motion blur datasets demonstrates that the proposed algorithm bridges the gap between simulated and real-world motion blurs and shows remarkable performance for event-based motion deblurring in real-world scenarios.

在本文中，我们提出了一个基于事件的运动去模糊自监督的端到端学习框架，其中利用真实世界的事件来缓解数据不一致导致的性能下降。为了达到这一目的，从事件中预测光流，利用模糊一致性和光度一致性，利用真实数据对去模糊网络进行自我监控。此外，提出了一种分段线性运动模型，以考虑运动非线性，从而为真实场景中运动模糊的物理形成提供精确的模型。对合成和真实运动模糊数据集的广泛评估表明，该算法弥补了模拟和真实运动模糊之间的差距，并在真实场景中显示了基于事件的运动去模糊的显著性能。

Existing state-of-the-art saliency detection methods heavily rely on CNN-based architectures. Alternatively, we rethink this task from a convolution-free sequence-to-sequence perspective and predict saliency by modeling long-range dependencies, which can not be achieved by convolution. Specifically, we develop a novel unified model based on a pure transformer, namely, visual Saliency Transformer (VST), for both RGB and RGB-D salient object detection (SOD). It takes image patches as inputs and leverages the transformer to propagate global contexts among image patches. Unlike conventional architectures used in Vision Transformer (ViT), we leverage multi-level token fusion and propose a new token upsampling method under the transformer framework to get high-resolution detection results. We also develop a token-based multi-task decoder to simultaneously perform saliency and boundary detection by introducing task-related tokens and a novel patch-task-attention mechanism. Experimental results show that our model outperforms existing methods on both RGB and RGB-D SOD benchmark datasets. Most importantly, our whole framework not only provides a new perspective for the SOD field but also shows a new paradigm for transformer-based dense prediction models. Code is available at <https://github.com/nnizhang/VST>.

现有的最先进的显著性检测方法严重依赖于基于CNN的体系结构。或者，我们从一个无卷积序列到另一个序列的角度重新考虑这项任务，并通过建模长期依赖关系来预测显著性，而这是卷积无法实现的。具体而言，我们开发了一种基于纯变换器的新的统一模型，即视觉显著性变换器 (VST) ，用于RGB和RGB-D显著性目标检测 (SOD) 。它将图像修补程序作为输入，并利用变换器在图像修补程序之间传播全局上下文。与视觉变换器 (ViT) 中使用的传统结构不同，我们利用多级令牌融合，并在变换器框架下提出了一种新的令牌上采样方法，以获得高分辨率的检测结果。我们还开发了一个基于令牌的多任务解码器，通过引入任务相关令牌和一种新的补丁任务注意机制，同时执行显著性和边界检测。实验结果表明，我们的模型在RGB和RGB-D SOD基准数据集上都优于现有的方法。最重要的是，我们的整个框架不仅为超氧化物歧化酶领域提供了新的视角，而且为基于变压器的密集预测模型提供了新的范例。代码可在<https://github.com/nnizhang/VST>。

Event cameras, which output events by detecting spatio-temporal brightness changes, bring a novel paradigm to image sensors with high dynamic range and low latency. Previous works have achieved impressive performances on event-based video reconstruction by introducing convolutional neural networks (CNNs). However, intrinsic locality of convolutional operations is not capable of modeling long-range dependency, which is crucial to many vision tasks. In this paper, we present a hybrid CNN-Transformer network for event-based video reconstruction (ET-Net), which merits the fine local information from CNN and global contexts from Transformer. In addition, we further propose a Token Pyramid Aggregation strategy to implement multi-scale token integration for relating internal and intersected semantic concepts in the token-space. Experimental results demonstrate that our proposed method achieves superior performance over state-of-the-art methods on multiple real-world event datasets. The code is available at <https://github.com/WarranWeng/ET-Net>

事件相机通过检测时空亮度变化来输出事件，为高动态范围、低延迟的图像传感器带来了一种新的模式。以前的工作已经通过引入卷积神经网络（CNN）在基于事件的视频重建方面取得了令人印象深刻的性能。然而，卷积运算的固有局部性无法建模长期依赖性，这对于许多视觉任务来说至关重要。在本文中，我们提出了一种用于基于事件的视频重建（ET-Net）的混合CNN变换网络，它利用了CNN的良好局部信息和变换器的全局上下文。此外，我们还提出了一种令牌金字塔聚合策略，以实现多尺度令牌集成，从而在令牌空间中关联内部和交叉语义概念。实验结果表明，我们提出的方法在多个真实事件数据集上取得了优于现有方法的性能。该守则可于<https://github.com/WarranWeng/ET-Net>

Most prior works on physical adversarial attacks mainly focus on the attack performance but seldom enforce any restrictions over the appearance of the generated adversarial patches. This leads to conspicuous and attention-grabbing patterns for the generated patches which can be easily identified by humans. To address this issue, we propose a method to craft physical adversarial patches for object detectors by leveraging the learned image manifold of a pretrained generative adversarial network (GAN) (e.g., BigGAN and StyleGAN) upon real-world images. Through sampling the optimal image from the GAN, our method can generate natural looking adversarial patches while maintaining high attack performance. With extensive experiments on both digital and physical domains and several independent subjective surveys, the results show that our proposed method produces significantly more realistic and natural looking patches than several state-of-the-art baselines while achieving competitive attack performance.

以前关于物理对抗攻击的大多数工作主要关注攻击性能，但很少对生成的对抗补丁的外观实施任何限制。这导致生成的斑块具有明显的、引人注目的图案，人类可以很容易地识别这些图案。为了解决这个问题，我们提出了一种方法，通过利用预训练生成对抗网络（GAN）（例如BigGAN和StyleGAN）的学习图像流形，为目标探测器制作物理对抗补丁。通过从GAN中采样最佳图像，我们的方法可以在保持高攻击性能的同时生成外观自然的对抗补丁。通过在数字和物理领域的大量实验以及一些独立的主观调查，结果表明，我们提出的方法在获得具有竞争力的攻击性能的同时，比一些最先进的基线生成更真实和自然的补丁。

Estimating the motion of the camera together with the 3D structure of the scene from a monocular vision system is a complex task that often relies on the so-called scene rigidity assumption. When observing a dynamic environment, this assumption is violated which leads to an ambiguity between the ego-motion of the camera and the motion of the objects. To solve this problem, we present a self-supervised learning framework for 3D object motion field estimation from monocular videos. Our contributions are two-fold. First, we propose a two-stage projection pipeline to explicitly disentangle the camera ego-motion and the object motions with dynamics attention module, called DAM. Specifically, we design an integrated motion model that estimates the motion of the camera and object in the first and second warping stages, respectively, controlled by the attention module through a shared motion encoder. Second, we propose an object motion field estimation through contrastive sample consensus, called CSAC, taking advantage of weak semantic prior (bounding box from an object detector) and geometric constraints (each object respects the rigid body motion model). Experiments on KITTI, Cityscapes, and Waymo Open Dataset demonstrate the relevance of our approach and show that our method outperforms state-of-the-art algorithms for the tasks of self-supervised monocular depth estimation, object motion segmentation, monocular scene flow estimation, and visual odometry.

从单目视觉系统估计摄像机的运动以及场景的三维结构是一项复杂的任务，通常依赖于所谓的场景刚性假设。当观察动态环境时，这一假设被违反，导致相机的自我运动和物体的运动之间存在歧义。为了解决这个问题，我们提出了一个自监督学习框架，用于单目视频的三维物体运动场估计。我们的贡献是双重的。首先，我们提出了一个两阶段投影管道，用动态注意模块DAM明确地分离相机自我运动和物体运动。具体来说，我们设计了一个集成的运动模型，分别在第一和第二扭曲阶段估计相机和对象的运动，由注意力模块通过共享运动编码器控制。其次，我们利用弱语义先验（对象检测器的边界框）和几何约束（每个对象都尊重刚体运动模型），通过对比样本一致性提出了一种对象运动场估计方法，称为CSAC。在KITTI、Cityscapes和Waymo开放数据集上的实验证明了我们方法的相关性，并表明我们的方法在自我监督的单目深度估计、对象运动分割、单目场景流估计和视觉里程测量任务方面优于最先进的算法。

In this paper, we propose a novel normalization method called gradient normalization (GN) to tackle the training instability of Generative Adversarial Networks (GANs) caused by the sharp gradient space. Unlike existing work such as gradient penalty and spectral normalization, the proposed GN only imposes a hard 1-Lipschitz constraint on the discriminator function, which increases the capacity of the discriminator. Moreover, the proposed gradient normalization can be applied to different GAN architectures with little modification. Extensive experiments on four datasets show that GANs trained with gradient normalization outperform existing methods in terms of both Frechet Inception Distance and Inception Score.

本文提出了一种称为梯度归一化（GN）的新的归一化方法，以解决由尖锐梯度空间引起的生成性对抗网络（GAN）的训练不稳定性问题。与现有的梯度惩罚和谱归一化等工作不同，该方法只对鉴别器函数施加硬1-Lipschitz约束，从而提高了鉴别器的容量。此外，所提出的梯度归一化方法可以应用于不同的GAN结构，只需稍加修改。在四个数据集上的大量实验表明，梯度归一化训练的GANs在Frechet起始距离和起始分数方面都优于现有方法。

3D object grounding aims to locate the most relevant target object in a raw point cloud scene based on a free-form language description. Understanding complex and diverse descriptions, and lifting them directly to a point cloud is a new and challenging topic due to the irregular and sparse nature of point clouds. There are three main challenges in 3D object grounding: to find the main focus in the complex and diverse description; to understand the point cloud scene; and to locate the target object. In this paper, we address all three challenges.

Firstly, we propose a language scene graph module to capture the rich structure and long-distance phrase correlations. Secondly, we introduce a multi-level 3D proposal relation graph module to extract the object-object and object-scene co-occurrence relationships, and strengthen the visual features of the initial proposals. Lastly, we develop a description guided 3D visual graph module to encode global contexts of phrases and proposals by a nodes matching strategy.

Extensive experiments on challenging benchmark datasets (ScanRefer and Nr3D) show that our algorithm outperforms existing state-of-the-art. Our code is available at <https://github.com/PNxD/FFL-3DOG>.

3D对象定位的目的是基于自由形式的语言描述，在原始点云场景中定位最相关的目标对象。由于点云的不规则性和稀疏性，理解复杂多样的描述并将其直接提升到点云是一个新的、具有挑战性的主题。三维物体基础的三个主要挑战是：在复杂多样的描述中找到主要焦点；了解点云场景；以及定位目标对象。在本文中，我们将解决所有三个挑战。首先，我们提出了一个语言场景图模块来捕获丰富的结构和长距离短语相关性。其次，我们引入了一个多层次的三维提案关系图模块，提取出对象与场景的共生关系，并增强了初始提案的视觉特征。最后，我们开发了一个描述导向的三维可视化图形模块，通过节点匹配策略对短语和建议的全局上下文进行编码。在具有挑战性的基准数据集（ScanRefer和Nr3D）上进行的大量实验表明，我们的算法优于现有的最新技术。我们的代码可在<https://github.com/PNxD/FFL-3DOG>。

Low-cost monocular 3D object detection plays a fundamental role in autonomous driving, whereas its accuracy is still far from satisfactory. Our objective is to dig into the 3D object detection task and reformulate it as the sub-tasks of object localization and appearance perception, which benefits to a deep excavation of reciprocal information underlying the entire task. We introduce a Dynamic Feature Reflecting Network, named DFR-Net, which contains two novel standalone modules: (i) the Appearance-Localization Feature Reflecting module (ALFR) that first separates task-specific features and then self-mutually reflects the reciprocal features; (ii) the Dynamic Intra-Trading module (DIT) that adaptively realigns the training processes of various sub-tasks via a self-learning manner. Extensive experiments on the challenging KITTI dataset demonstrate the effectiveness and generalization of DFR-Net. We rank 1st among all the monocular 3D object detectors in the KITTI test set (till March 16th, 2021). The proposed method is also easy to be plug-and-play in many cutting-edge 3D detection frameworks at negligible cost to boost performance. The code will be made publicly available.

低成本的单目三维目标检测在自动驾驶中起着基础性的作用，但其精度仍远不能令人满意。我们的目标是深入研究三维目标检测任务，并将其转化为目标定位和外观感知的子任务，这有助于深入挖掘整个任务背后的交互信息。我们介绍了一种动态特征反射网络，称为DFR网络，它包含两个独立的模块：(i) 外观定位特征反射模块 (ALFR)，该模块首先分离任务特定的特征，然后自互反射交互特征；(ii) 动态内部交易模块 (DIT)，通过自学习方式自适应地重新调整各子任务的培训过程。在具有挑战性的KITTI数据集上的大量实验证明了DFR网络的有效性和泛化性。我们在KITTI测试集中的所有单目3D物体探测器中排名第一（截至2021年3月16日）。所提出的方法也很容易在许多尖端的3D检测框架中即插即用，成本可以忽略不计，以提高性能。该守则将公开发布。

This paper presents a simple and effective unsupervised adaptation method for Robust Object Detection (SimROD). To overcome the challenging issues of domain shift and pseudo-label noise, our method integrates a novel domain-centric data augmentation, a gradual self-labeling adaptation procedure, and a teacher-guided fine-tuning mechanism. Using our method, target domain samples can be leveraged to adapt object detection models without changing the model architecture or generating synthetic data. When applied to image corruptions and high-level cross-domain adaptation benchmarks, our method outperforms prior baselines on multiple domain adaptation benchmarks. SimROD achieves new state-of-the-art on standard real-to-synthetic and cross-camera setup benchmarks. On the image corruption benchmark, models adapted with our method achieved a relative robustness improvement of 15-25% AP50 on Pascal-C and 5-6% AP on COCO-C and Cityscapes-C. On the cross-domain benchmark, our method outperformed the best baseline performance by up to 8% and 4% AP50 on Comic and Watercolor respectively.

提出了一种简单有效的鲁棒目标检测的无监督自适应方法 (SimROD)。为了克服域移位和伪标签噪声的挑战性问题，我们的方法集成了一种新的以域为中心的数据扩充、一种渐进的自标签自适应过程和一种教师指导的微调机制。使用我们的方法，可以利用目标域样本来调整对象检测模型，而无需改变模型结构或生成合成数据。当应用于图像损坏和高级跨域适配基准时，我们的方法在多域适配基准上优于先前的基线。SimROD在标准真实到合成和跨摄像头设置基准上实现了最新水平。在图像损坏基准上，采用我们的方法的模型在Pascal-C上实现了15-25%的AP50相对鲁棒性改进，在COCO-C和Cityscapes-C上实现了5-6%的AP50相对鲁棒性改进。在跨域基准上，我们的方法在漫画和水彩上分别比最佳基准性能高出8%和4%。

Today's state-of-the-art methods for 3D object detection are based on lidar, stereo, or monocular cameras. Lidar-based methods achieve the best accuracy, but have a large footprint, high cost, and mechanically-limited angular sampling rates, resulting in low spatial resolution at long ranges. Recent approaches using low-cost monocular or stereo cameras promise to overcome these limitations but struggle in low-light or low-contrast regions as they rely on passive CMOS sensors. We propose a novel 3D object detection modality that exploits temporal illumination cues from a low-cost monocular gated imager. We introduce a novel deep detection architecture, Gated3D, that is tailored to temporal illumination cues in gated images. This modality allows us to exploit mature 2D object feature extractors that guide the 3D predictions through a frustum segment estimation. We assess the proposed method experimentally on a 3D detection dataset that includes gated images captured over 10,000 km of driving data. We validate that our method outperforms state-of-the-art monocular and stereo methods, opening up a new sensor modality as an avenue to replace lidar in autonomous driving.

<https://light.princeton.edu/gated3d>

当今最先进的三维物体检测方法是基于激光雷达、立体或单目相机。基于激光雷达的方法实现了最佳精度，但占地面积大，成本高，且机械角度采样率有限，导致远距离的空间分辨率较低。最近使用低成本单目或立体相机的方法有望克服这些限制，但由于依赖于无源CMOS传感器，因此在低光或低对比度区域难以实现。我们提出了一种新的三维物体检测模式，利用低成本单目门控成像仪的时间照明线索。我们介绍了一种新的深度检测体系结构Gated3D，该体系结构针对门控图像中的时间照明线索进行定制。这种模式使我们能够利用成熟的2D对象特征提取器，通过平截头体段估计来指导3D预测。我们在一个3D检测数据集上对所提出的方法进行了实验评估，该数据集包括10000公里行驶数据中采集的选通图像。我们验证了我们的方法优于最先进的单目和立体方法，开辟了一种新的传感器模式，作为在自动驾驶中取代激光雷达的途径。<https://light.princeton.edu/gated3d>

We show for the first time that a multilayer perceptron (MLP) can serve as the only scene representation in a real-time SLAM system for a handheld RGB-D camera. Our network is trained in live operation without prior data, building a dense, scene-specific implicit 3D model of occupancy and colour which is also immediately used for tracking. Achieving real-time SLAM via continual training of a neural network against a live image stream requires significant innovation. Our iMAP algorithm uses a keyframe structure and multi-processing computation flow, with dynamic information-guided pixel sampling for speed, with tracking at 10 Hz and global map updating at 2 Hz. The advantages of an implicit MLP over standard dense SLAM techniques include efficient geometry representation with automatic detail control and smooth, plausible filling-in of unobserved regions such as the back surfaces of objects.

我们首次展示了多层感知器（MLP）可以作为手持RGB-D相机实时SLAM系统中唯一的场景表示。我们的网络在没有事先数据的情况下进行现场操作培训，建立一个密集的、特定于场景的隐式3D占用和颜色模型，该模型也可立即用于跟踪。通过对实时图像流持续训练神经网络来实现实时SLAM需要重大创新。我们的iMAP算法使用关键帧结构和多处理计算流，动态信息引导像素采样以提高速度，跟踪频率为10 Hz，全局地图更新频率为2 Hz。与标准密集SLAM技术相比，隐式MLP的优点包括具有自动细节控制的高效几何表示和平滑、合理地填充未观察区域（如对象的后表面）。

In click-based interactive segmentation, the mask extraction process is dictated by positive/negative user clicks; however, most existing methods do not fully exploit the user cues, requiring excessive numbers of clicks for satisfactory results. We propose Conditional Diffusion Network (CDNet), which propagates labeled representations from clicks to conditioned destinations with two levels of affinities: Feature Diffusion Module (FDM) spreads features from clicks to potential target regions with global similarity; Pixel Diffusion Module (PDM) diffuses the predicted logits of clicks within locally connected regions. Thus, the information inferred by user clicks could be generalized to proper destinations. In addition, we put forward Diversified Training (DT), which reduces the optimization ambiguity caused by click simulation. With FDM, PDM and DT, CDNet could better understand user's intentions and make better predictions with limited interactions. CDNet achieves state-of-the-art performance on several benchmarks.

在基于点击的交互式分割中，掩码提取过程由用户的正/负点击决定；然而，大多数现有的方法没有充分利用用户提示，需要过多的点击才能获得满意的结果。我们提出了条件扩散网络（CDNet），该网络将标签表示从点击传播到具有两级亲和力的条件目的地：特征扩散模块（FDM）将特征从点击传播到具有全局相似性的潜在目标区域；像素扩散模块（PDM）在本地连接的区域内扩散预测的点击次数。因此，通过用户点击推断出的信息可以推广到适当的目的地。此外，我们还提出了多样化训练（DT），减少了点击模拟造成的优化模糊。通过FDM、PDM和DT，CDNet可以更好地理解用户的意图，并在有限的交互中做出更好的预测。CDNet在几个基准上实现了最先进的性能。

The performance of computer vision models significantly improves with more labeled data. However, the acquisition of labeled data is limited by the high cost. To mitigate the reliance on large labeled datasets, active learning (AL) and semi-supervised learning (SSL) are frequently adopted. Although current mainstream methods begin to combine SSL and AL (SSL-AL) to excavate the diverse expressions of unlabeled samples, these methods' fully supervised task models are still trained only with labeled data. Besides, these method's SSL-AL frameworks suffer from mismatch problems. Here, we propose a graph-based SSL-AL framework to unleash the SSL task models' power and make an effective SSL-AL interaction. In the framework, SSL leverages graph-based label propagation to deliver virtual labels to unlabeled samples, rendering AL samples' structural distribution and boosting AL. AL finds samples near the clusters' boundary to help SSL perform better label propagation by exploiting adversarial examples. The information exchange in the closed-loop realizes mutual enhancement of SSL and AL. Experimental results show that our method outperforms the state-of-the-art methods against classification and segmentation benchmarks.

随着标记数据的增多，计算机视觉模型的性能显著提高。然而，高成本限制了标记数据的获取。为了减少对大型标记数据集的依赖，经常采用主动学习（AL）和半监督学习（SSL）。虽然目前主流的方法开始结合SSL和AL（SSL-AL）来挖掘未标记样本的不同表达，但这些方法的完全监督任务模型仍然只使用标记数据进行训练。此外，这些方法的SSL-AL框架存在不匹配问题。在这里，我们提出了一个基于图形的SSL-AL框架来释放SSL任务模型的威力，并进行有效的SSL-AL交互。在该框架中，SSL利用基于图形的标签传播将虚拟标签传递给未标记的样本，呈现AL样本的结构分布，并增强AL。AL在集群边界附近发现样本，以帮助SSL利用对抗性示例执行更好的标签传播。闭环中的信息交换实现了SSL和AL的相互增强。实验结果表明，相对于分类和分割基准，我们的方法优于现有的方法。

In open set recognition, a classifier has to detect unknown classes that are not known at training time. In order to recognize new categories, the classifier has to project the input samples of known classes in very compact and separated regions of the features space for discriminating samples of unknown classes. Recently proposed Capsule Networks have shown to outperform alternatives in many fields, particularly in image recognition, however they have not been fully applied yet to open-set recognition. In capsule networks, scalar neurons are replaced by capsule vectors or matrices, whose entries represent different properties of objects. In our proposal, during training, capsules features of the same known class are encouraged to match a pre-defined gaussian, one for each class. To this end, we use the variational autoencoder framework, with a set of gaussian priors as the approximation for the posterior distribution. In this way, we are able to control the compactness of the features of the same class around the center of the gaussians, thus controlling the ability of the classifier in detecting samples from unknown classes. We conducted several experiments and ablation of our model, obtaining state of the art results on different datasets in the open set recognition and unknown detection tasks.

在开集识别中，分类器必须检测在训练时未知的未知类。为了识别新的类别，分类器必须将已知类别的输入样本投影到特征空间中非常紧凑和分离的区域，以识别未知类别的样本。最近提出的胶囊网络在许多领域，特别是在图像识别方面，表现出了比其他方法更好的性能，但是它们尚未完全应用于开放集识别。在胶囊网络中，标量神经元被胶囊向量或矩阵所代替，其条目代表对象的不同属性。在我们的建议中，在培训期间，鼓励相同已知类别的胶囊特征匹配预定义的高斯分布，每个类别一个。为此，我们使用变分自动编码器框架，以一组高斯先验作为后验分布的近似值。这样，我们就能够控制同一类特征在高斯中心附近的紧致性，从而控制分类器检测未知类样本的能力。我们对我们的模型进行了几次实验和烧蚀，在开放集识别和未知检测任务的不同数据集上获得了最新的结果。

Automatic security inspection using computer vision technology is a challenging task in real-world scenarios due to various factors, including intra-class variance, class imbalance, and occlusion. Most of the previous methods rarely solve the cases that the prohibited items are deliberately hidden in messy objects due to the lack of large-scale datasets, restricted their applications in real-world scenarios. Towards real-world prohibited item detection, we collect a large-scale dataset, named as PIDray, which covers various cases in real-world scenarios for prohibited item detection, especially for deliberately hidden items. With an intensive amount of effort, our dataset contains 12 categories of prohibited items in 47,677 X-ray images with high-quality annotated segmentation masks and bounding boxes. To the best of our knowledge, it is the largest prohibited items detection dataset to date. Meanwhile, we design the selective dense attention network (SDANet) to construct a strong baseline, which consists of the dense attention module and the dependency refinement module. The dense attention module formed by the spatial and channel-wise dense attentions, is designed to learn the discriminative features to boost the performance. The dependency refinement module is used to exploit the dependencies of multi-scale features. Extensive experiments conducted on the collected PIDray dataset demonstrate that the proposed method performs favorably against the state-of-the-art methods, especially for detecting the deliberately hidden items.

由于各种因素，包括类内差异、类间不平衡和遮挡，使用计算机视觉技术的自动安全检查在现实场景中是一项具有挑战性的任务。以往的大多数方法很少解决由于缺乏大规模数据集而故意将禁止项隐藏在杂乱对象中的情况，限制了它们在现实场景中的应用。针对现实世界中的违禁物品检测，我们收集了一个名为PIDray的大规模数据集，该数据集涵盖了现实世界中违禁物品检测场景中的各种情况，尤其是故意隐藏的物品。经过大量的努力，我们的数据集包含了47677张X射线图像中的12类违禁物品，这些图像具有高质量的带注释的分割遮罩和边界框。据我们所知，这是迄今为止最大的违禁品检测数据集。同时，我们设计了选择性密集注意网络 (SDANet) 来构建一个强大的基线，该基线由密集注意模块和依赖项细化模块组成。密集注意模块由空间和通道密集注意组成，用于学习区分特征以提高性能。依赖关系细化模块用于利用多尺度特征的依赖关系。在收集的PIDray数据集上进行的大量实验表明，该方法优于现有的方法，特别是在检测有意隐藏的项目时。

Contaminants such as dust, dirt and moisture adhering to the camera lens can greatly affect the quality and clarity of the resulting image or video. In this paper, we propose a video restoration method to automatically remove these contaminants and produce a clean video. Our approach first seeks to detect attention maps that indicate the regions that need to be restored. In order to leverage the corresponding clean pixels from adjacent frames, we propose a flow completion module to hallucinate the flow of the background scene to the attention regions degraded by the contaminants. Guided by the attention maps and completed flows, we propose a recurrent technique to restore the input frame by fetching clean pixels from adjacent frames. Finally, a multi-frame processing stage is used to further process the entire video sequence in order to enforce temporal consistency. The entire network is trained on a synthetic dataset that approximates the physical lighting properties of contaminant artifacts. This new dataset and our novel framework lead to our method that is able to address different contaminants and outperforms competitive restoration approaches both qualitatively and quantitatively.

粘附在相机镜头上的灰尘、污垢和水分等污染物会极大地影响生成图像或视频的质量和清晰度。在本文中，我们提出了一种视频恢复方法来自动去除这些污染物并生成干净的视频。我们的方法首先试图检测出需要恢复的区域的注意力地图。为了利用相邻帧中相应的干净像素，我们提出了一个流完成模块，将背景场景的流幻觉到被污染物降解的注意区域。在注意图和完整流程的指导下，我们提出了一种通过从相邻帧中提取干净像素来恢复输入帧的递归技术。最后，使用多帧处理阶段进一步处理整个视频序列，以增强时间一致性。整个网络在一个模拟污染工件的物理照明特性的合成数据集上进行训练。这个新的

数据集和我们的新框架使我们的方法能够处理不同的污染物，并且在定性和定量上都优于竞争性恢复方法。

Attribution map visualization has arisen as one of the most effective techniques to understand the underlying inference process of Convolutional Neural Networks. In this task, the goal is to compute a score for each image pixel related to its contribution to the network output. In this paper, we introduce Disentangled Masked Backpropagation (DMBP), a novel gradient-based method that leverages on the piecewise linear nature of ReLU networks to decompose the model function into different linear mappings. This decomposition aims to disentangle the attribution maps into positive, negative and nuisance factors by learning a set of variables masking the contribution of each filter during back-propagation. A thorough evaluation over standard architectures (ResNet50 and VGG16) and benchmark datasets (PASCAL VOC and ImageNet) demonstrates that DMBP generates more visually interpretable attribution maps than previous approaches. Additionally, we quantitatively show that the maps produced by our method are more consistent with the true contribution of each pixel to the final network output.

属性图可视化是理解卷积神经网络推理过程最有效的技术之一。在本任务中，目标是为每个图像像素计算与其对网络输出的贡献相关的分数。在本文中，我们介绍了一种新的基于梯度的方法——解纠缠屏蔽反向传播（DMBP），它利用ReLU网络的分段线性特性将模型函数分解为不同的线性映射。这种分解的目的是通过学习一组变量，掩盖反向传播过程中每个滤波器的贡献，将属性映射分解为积极、消极和有害因素。对标准体系结构（ResNet50和VGG16）和基准数据集（PASCAL VOC和ImageNet）的全面评估表明，DMBP生成的属性图比以前的方法更直观。此外，我们定量地表明，我们的方法生成的贴图更符合每个像素对最终网络输出的真实贡献。

Vessel segmentation is critically essential for diagnosing a series of diseases, e.g., coronary artery disease and retinal disease. However, annotating vessel segmentation maps of medical images is notoriously challenging due to the tiny and complex vessel structures, leading to insufficient available annotated datasets for existing supervised methods and domain adaptation methods. The subtle structures and confusing background of medical images further suppress the efficacy of unsupervised methods. In this paper, we propose a self-supervised vessel segmentation method via adversarial learning. Our method learns vessel representations by training an attention-guided generator and a segmentation generator to simultaneously synthesize fake vessels and segment vessels out of coronary angiograms. To support the research, we also build the first X-ray angiography coronary vessel segmentation dataset, named XCAD. We evaluate our method extensively on multiple vessel segmentation datasets, including the XCAD dataset, the DRIVE dataset, and the STARE dataset. The experimental results show our method suppresses unsupervised methods significantly and achieves competitive performance compared with supervised methods and traditional methods.

血管分割对于诊断一系列疾病至关重要，例如冠状动脉疾病和视网膜疾病。然而，由于血管结构的微小和复杂，对医学图像的血管分割图进行注释是一个众所周知的挑战，导致现有的监督方法和领域适应方法没有足够的可用注释数据集。医学图像的细微结构和混乱背景进一步抑制了无监督方法的效果。本文提出了一种基于对抗学习的自监督血管分割方法。我们的方法通过训练一个注意力引导生成器和一个分割生成器来同时合成假血管和从冠状动脉造影中分割血管来学习血管表示。为了支持这项研究，我们还构建了第一个X射线血管造影冠状血管分割数据集，名为XCAD。我们在多个血管分割数据集上广泛地评估了我们的方法，包括XCAD数据集、DRIVE数据集和STARE数据集。实验结果表明，与有监督方法和传统方法相比，该方法显著抑制了无监督方法，取得了较好的性能。

We present a deep learning pipeline that leverages network self-prior to recover a full 3D model consisting of both a triangular mesh and a texture map from the colored 3D point cloud. Different from previous methods either exploiting 2D self-prior for image editing or 3D self-prior for pure surface reconstruction, we propose to exploit a novel hybrid 2D-3D self-prior in deep neural networks to significantly improve the geometry quality and produce a high-resolution texture map, which is typically missing from the output of commodity-level 3D scanners. In particular, we first generate an initial mesh using a 3D convolutional neural network with 3D self-prior, and then encode both 3D information and color information in the 2D UV atlas, which is further refined by 2D convolutional neural networks with the self-prior. In this way, both 2D and 3D self-priors are utilized for the mesh and texture recovery. Experiments show that, without the need of any additional training data, our method recovers the 3D textured mesh model of high quality from sparse input, and outperforms the state-of-the-art methods in terms of both the geometry and texture quality.

我们提供了一个深度学习管道，该管道在从彩色3D点云恢复由三角形网格和纹理贴图组成的完整3D模型之前利用网络自身。与以往利用2D自先验进行图像编辑或利用3D自先验进行纯表面重建的方法不同，我们建议在深度神经网络中利用一种新的2D-3D混合自先验，以显著提高几何质量并生成高分辨率纹理图，商品级3D扫描仪的输出中通常缺少这一点。特别是，我们首先使用具有三维自先验的三维卷积神经网络生成初始网格，然后在二维UV图谱中编码三维信息和颜色信息，然后通过具有自先验的二维卷积神经网络进一步细化。这样，2D和3D自先验都用于网格和纹理恢复。实验表明，在不需要任何额外的训练数据的情况下，我们的方法从稀疏输入中恢复出高质量的三维纹理网格模型，并且在几何和纹理质量方面都优于现有的方法。

Recently, a novel retina-inspired camera, namely spike camera, has shown great potential for recording high-speed dynamic scenes. Unlike the conventional digital cameras that compact the visual information within the exposure interval into a single snapshot, the spike camera continuously outputs binary spike streams to record the dynamic scenes, yielding a very high temporal resolution. Most of the existing reconstruction methods for spike camera focus on reconstructing images with the same resolution as spike camera. However, as a trade-off of high temporal resolution, the spatial resolution of spike camera is limited, resulting in inferior details of the reconstruction. To address this issue, we develop a spike camera super-resolution framework, aiming to super resolve high-resolution intensity images from the low-resolution binary spike streams. Due to the relative motion between the camera and the objects to capture, the spikes fired by the same sensor pixel no longer describes the same points in the external scene. In this paper, we properly exploit the relative motion and derive the relationship between light intensity and each spike, so as to recover the external scene with both high temporal and high spatial resolution. Experimental results demonstrate that the proposed method can reconstruct pleasant high-resolution images from low-resolution spike streams.

最近，一种新颖的视网膜启发相机，即spike相机，显示出记录高速动态场景的巨大潜力。与将曝光间隔内的视觉信息压缩为单个快照的传统数码相机不同，spike相机连续输出二进制spike流以记录动态场景，产生非常高的时间分辨率。现有的斯派克相机重建方法大多侧重于重建与斯派克相机分辨率相同的图像。然而，作为高时间分辨率的折衷，spike相机的空间分辨率受到限制，导致重建细节较差。为了解决这个问题，我们开发了一个spike摄像机超分辨率框架，旨在从低分辨率二值spike流中超分辨率高分辨率强度图像。由于相机和要捕获的对象之间的相对运动，同一传感器像素发射的尖峰不再描述外部场景中的相同点。在本文中，我们适当地利用了相对运动，推导了光强度和每个尖峰之间的关系，以便以高时间和高空间分辨率恢复外部场景。实验结果表明，该方法能够从低分辨率尖峰流中重建出令人愉悦的高分辨率图像。

Existing methods for multi-modal domain translation learn to embed the input images into a domain-invariant "content" space and a domain-specific "style" space from which novel images can be synthesized. Rather than learning to embed the RGB image from scratch we propose deriving our content representation from conditioning data produced by pretrained off-the-shelf networks. Motivated by the inherent ambiguity of "content", which has different meanings depending on the desired level of abstraction, this approach gives intuitive control over which aspects of content are preserved across domains. We evaluate our method on traditional, well-aligned, datasets such as CelebA-HQ, and propose two novel datasets for evaluation on more complex scenes: ClassicTV and FFHQ-WildCrops. Our approach, which we call Sensorium, enables higher quality domain translation for complex scenes than prior work.

现有的多模态域翻译方法学习将输入图像嵌入到一个域不变的“内容”空间和一个特定于域的“风格”空间中，从中可以合成新图像。与其从头开始学习嵌入RGB图像，我们建议从预训练现成网络生成的条件数据中导出内容表示。受“内容”固有的模糊性（根据所需的抽象级别具有不同的含义）的影响，这种方法可以直观地控制跨域保存内容的哪些方面。我们在传统的、对齐良好的数据集（如CelebA HQ）上评估了我们的方法，并提出了两个新的数据集用于更复杂场景的评估：ClassicTV和FFHQ WildCrops。我们称之为Sensorium的方法能够为复杂场景提供比以前更高质量的域转换。

We present a novel and flexible architecture for point cloud segmentation with dual-representation iterative learning. In point cloud processing, different representations have their own pros and cons. Thus, finding suitable ways to represent point cloud data structure while keeping its own internal physical property such as permutation and scale-invariant is a fundamental problem. Therefore, we propose our work, DRINet, which serves as the basic network structure for dual-representation learning with great flexibility at feature transferring and less computation cost, especially for large-scale point clouds. DRINet mainly consists of two modules called Sparse Point-Voxel Feature Extraction and Sparse Voxel-Point Feature Extraction. By utilizing these two modules iteratively, features can be propagated between two different representations. We further propose a novel multi-scale pooling layer for pointwise locality learning to improve context information propagation. Our network achieves state-of-the-art results for point cloud classification and segmentation tasks on several datasets while maintaining high runtime efficiency. For large-scale outdoor scenarios, our method outperforms state-of-the-art methods with a real-time inference speed of 62ms per frame.

我们提出了一种新颖灵活的基于双表示迭代学习的点云分割体系结构。在点云处理中，不同的表示方式有其优缺点。因此，寻找合适的方法来表示点云数据结构，同时保持其自身的物理特性，如排列和尺度不变，是一个基本问题。因此，我们提出了我们的工作DRINet，它可以作为双表示学习的基本网络结构，在特征传输方面具有极大的灵活性，并且计算量较小，特别是对于大规模点云。DRINet主要由稀疏点体素特征提取和稀疏点体素特征提取两个模块组成。通过迭代地利用这两个模块，可以在两个不同的表示之间传播特征。我们进一步提出了一种新的多尺度池层，用于点式局部学习，以改进上下文信息传播。我们的网络在多个数据集上实现了最先进的点云分类和分割任务，同时保持了较高的运行效率。对于大规模的室外场景，我们的方法优于最先进的方法，实时推理速度为每帧62ms。

Convolutional layers in CNNs implement linear filters which decompose the input into different frequency bands. However, most modern architectures neglect standard principles of filter design when optimizing their model choices regarding the size and shape of the convolutional kernel. In this work, we consider the well-known problem of spectral leakage caused by windowing artifacts in filtering operations in the context of CNNs. We show that the small size of CNN kernels make them susceptible to spectral leakage, which may induce performance-degrading artifacts. To address this issue, we propose the use of larger kernel sizes along with the Hamming window function to alleviate leakage in CNN architectures. We demonstrate improved classification accuracy on multiple benchmark datasets including Fashion-MNIST, CIFAR-10, CIFAR-100 and ImageNet with the simple use of a standard window function in convolutional layers. Finally, we show that CNNs employing the Hamming window display increased robustness against various adversarial attacks.

CNN中的卷积层实现线性滤波器，将输入分解为不同的频带。然而，大多数现代体系结构在优化其关于卷积核的大小和形状的模型选择时忽略了滤波器设计的标准原则。在这项工作中，我们考虑众所周知的问题，频谱泄漏所造成的窗口工件在过滤操作的背景下，CNNs。我们发现，CNN内核的小尺寸使得它们容易受到频谱泄漏的影响，这可能会导致性能下降。为了解决这个问题，我们建议使用更大的内核大小以及Hamming窗口函数来缓解CNN架构中的泄漏。通过在卷积层中简单使用标准窗口函数，我们在多个基准数据集（包括Fashion MNIST、CIFAR-10、CIFAR-100和ImageNet）上展示了改进的分类精度。最后，我们展示了采用汉明窗口显示的CNN增强了对各种对抗性攻击的鲁棒性。

Image completion has made tremendous progress with convolutional neural networks (CNNs), because of their powerful texture modeling capacity. However, due to some inherent properties (eg, local inductive prior, spatial-invariant kernels), CNNs do not perform well in understanding global structures or naturally support pluralistic completion. Recently, transformers demonstrate their power in modeling the long-term relationship and generating diverse results, but their computation complexity is quadratic to input length, thus hampering the application in processing high-resolution images. This paper brings the best of both worlds to pluralistic image completion: appearance prior reconstruction with transformer and texture replenishment with CNN. The former transformer recovers pluralistic coherent structures together with some coarse textures, while the latter CNN enhances the local texture details of coarse priors guided by the high-resolution masked images. The proposed method vastly outperforms state-of-the-art methods in terms of three aspects: 1) large performance boost on image fidelity even compared to deterministic completion methods; 2) better diversity and higher fidelity for pluralistic completion; 3) exceptional generalization ability on large masks and generic dataset, like ImageNet. Code and pre-trained models have been publicly released at <https://github.com/raywzy/ICT>.

卷积神经网络 (CNN) 由于其强大的纹理建模能力，在图像处理方面取得了巨大的进展。然而，由于一些固有的特性（如局部归纳先验、空间不变核），CNN在理解全局结构方面表现不佳，也不能自然地支持多元补全。最近，变压器显示了其在建模长期关系和生成不同结果方面的能力，但其计算复杂度是输入长度的二次方，因此阻碍了其在处理高分辨率图像中的应用。本文将这两个方面的优点结合起来，实现了多元图像的完整性：使用变换器进行外观先验重建和使用CNN进行纹理补充。前一种变换器与一些粗糙纹理一起恢复多元相干结构，而后一种变换器在高分辨率遮罩图像的引导下增强粗糙先验的局部纹理细节。所提出的方法在三个方面远远优于最先进的方法：1) 即使与确定性完成方法相比，图像保真度也有很大的性能提升；2) 多元化完成的更好多样性和更高保真度；3) 在大型遮罩和通用数据集（如ImageNet）上具有出色的泛化能力。代码和预先培训的模型已在<https://github.com/raywzy/ICT>。

This paper proposes a novel weakly supervised approach for anomaly detection, which begins with a relation-aware feature extractor to capture the multi-scale convolutional neural network (CNN) features from a video. Afterwards, self-attention is integrated with conditional random fields (CRFs), the core of the network, to make use of the ability of self-attention in capturing the short-range correlations of the features and the ability of CRFs in learning the inter-dependencies of these features. Such a framework can learn not only the spatio-temporal interactions among the actors which are important for detecting complex movements, but also their short- and long-term dependencies across frames. Also, to deal with both local and non-local relationships of the features, a new variant of self-attention is developed by taking into consideration a set of cliques with different temporal localities. Moreover, a contrastive multi-instance learning scheme is considered to broaden the gap between the normal and abnormal instances, resulting in more accurate abnormal discrimination. Simulations reveal that the new method provides superior performance to the state-of-the-art works on the widespread UCF-Crime and ShanghaiTech datasets.

本文提出了一种新的弱监督异常检测方法，该方法从一个关系感知的特征抽取器开始，从视频中捕获多尺度卷积神经网络（CNN）特征。然后，将自我注意与网络的核心条件随机场（CRF）相结合，利用自我注意捕获特征短期相关性的能力和CRF学习这些特征相互依赖性的能力。这样一个框架不仅可以了解对检测复杂运动非常重要的参与者之间的时空交互，还可以了解他们在帧间的短期和长期依赖关系。此外，为了处理特征的局部和非局部关系，通过考虑一组具有不同时间位置的派系，开发了一种新的自我注意变体。此外，还考虑了一种对比多实例学习方案，以扩大正常和异常实例之间的差距，从而实现更准确的异常识别。仿真结果表明，在广泛分布的UCF犯罪数据集和上海科技大学数据集上，新方法的性能优于最先进的方法。

Text recognition remains a fundamental and extensively researched topic in computer vision, largely owing to its wide array of commercial applications. The challenging nature of the very problem however dictated a fragmentation of research efforts: Scene Text Recognition (STR) that deals with text in everyday scenes, and Handwriting Text Recognition (HTR) that tackles hand-written text. In this paper, for the first time, we argue for their unification -- we aim for a single model that can compete favourably with two separate state-of-the-art STR and HTR models. We first show that cross-utilisation of STR and HTR models trigger significant performance drops due to differences in their inherent challenges. We then tackle their union by introducing a knowledge distillation (KD) based framework. This however is non-trivial, largely due to the variable-length and sequential nature of text sequences, which renders off-the-shelf KD techniques that mostly work with global fixed length data, inadequate. For that, we propose four distillation losses, all of which are specifically designed to cope with the aforementioned unique characteristics of text recognition. Empirical evidence suggests that our proposed unified model performs at par with individual models, even surpassing them in certain cases. Ablative studies demonstrate that naive baselines such as a two-stage framework, multi-task and domain adaption/generalisation alternatives do not work that well, further authenticating our design.

由于其广泛的商业应用，文本识别仍然是计算机视觉中一个基础和广泛研究的课题。然而，这个问题的挑战性决定了研究工作的碎片化：处理日常场景中文本的场景文本识别（STR）和处理手写文本的手写文本识别（HTR）。在这篇论文中，我们第一次主张它们的统一——我们的目标是建立一个能够与两个独立的最先进的STR和HTR模型竞争的单一模型。我们首先表明，由于STR和HTR模型固有挑战的差异，它们的交叉使用会导致性能显著下降。然后，我们通过引入基于知识提炼（KD）的框架来解决它们的结合问题。然而，这并非无关紧要，主要是由于文本序列的可变长度和序列性质，这使得主要用于全局固定长度数据的现成KD技术不充分。为此，我们提出了四种蒸馏损失，所有这些都是专门为处理上述文本识别的独特特征而设计的。经验证据表明，我们提出的统一模型与个别模型相比，在某些情况下甚至超

过了它们。烧蚀研究表明，天真的基线，如两阶段框架、多任务和领域适应/泛化替代方案，并不能很好地发挥作用，进一步验证了我们的设计。

We present a passive non-line-of-sight method that infers the number of people or activity of a person from the observation of a blank wall in an unknown room. Our technique analyzes complex imperceptible changes in indirect illumination in a video of the wall to reveal a signal that is correlated with motion in the hidden part of a scene. We use this signal to classify between zero, one, or two moving people, or the activity of a person in the hidden scene. We train two convolutional neural networks using data collected from 20 different scenes, and achieve an accuracy of approximately 94% for both tasks in unseen test environments and real-time online settings. Unlike other passive non-line-of-sight methods, the technique does not rely on known occluders or controllable light sources, and generalizes to unknown rooms with no recalibration. We analyze the generalization and robustness of our method with both real and synthetic data, and study the effect of the scene parameters on the signal quality.

我们提出了一种被动的非视线方法，该方法通过观察未知房间中的空白墙来推断人员数量或人员活动。我们的技术分析了墙壁视频中间接照明的复杂不可感知变化，以揭示与场景隐藏部分的运动相关的信号。我们使用这个信号来区分零个、一个或两个移动的人，或者隐藏场景中一个人的活动。我们使用从20个不同场景收集的数据来训练两个卷积神经网络，在看不见的测试环境和实时在线设置中，两个任务的准确率都达到了约94%。与其他被动非视线方法不同，该技术不依赖于已知的遮光罩或可控光源，并可推广到未知房间，无需重新校准。我们用真实数据和合成数据分析了该方法的泛化性和鲁棒性，并研究了场景参数对信号质量的影响。

weakly Supervised Semantic Segmentation (WSSS) based on image-level labels has been greatly advanced by exploiting the outputs of Class Activation Map (CAM) to generate the pseudo labels for semantic segmentation. However, CAM merely discovers seeds from a small number of regions, which may be insufficient to serve as pseudo masks for semantic segmentation. In this paper, we formulate the expansion of object regions in CAM as an increase in information. From the perspective of information theory, we propose a novel Complementary Patch (CP) Representation and prove that the information of the sum of the CAMs by a pair of input images with complementary hidden (patched) parts, namely CP Pair, is greater than or equal to the information of the baseline CAM. Therefore, a CAM with more information related to object seeds can be obtained by narrowing down the gap between the sum of CAMs generated by the CP Pair and the original CAM. We propose a CP Network (CPN) implemented by a triplet network and three regularization functions. To further improve the quality of the CAMs, we propose a Pixel-Region Correlation Module (PRCM) to augment the contextual information by using object-region relations between the feature maps and the CAMs. Experimental results on the PASCAL VOC 2012 datasets show that our proposed method achieves a new state-of-the-art in WSSS, validating the effectiveness of our CP Representation and CPN.

基于图像级标签的弱监督语义分割 (WSSS) 通过利用类激活图 (CAM) 的输出生成用于语义分割的伪标签得到了极大的改进。然而，CAM仅从少量区域中发现种子，这可能不足以作为语义分割的伪掩码。在本文中，我们将CAM中对象区域的扩展表述为信息的增加。从信息论的角度，我们提出了一种新的互补补丁 (CP) 表示，并证明了一对具有互补隐藏 (补丁) 部分的输入图像 (即CP对) 的CAM和的信息大于或等于基线CAM的信息。因此，通过缩小CP对生成的凸轮和原始凸轮之间的间隙，可以获得具有更多对象种子相关信息的凸轮。我们提出了一个由三重网络和三个正则化函数实现的CP网络 (CPN)。为了进一步提高CAM的质量，我们提出了一个像素区域相关模块 (PRCM)，通过使用特征映射和CAM之间的对象区域关系来增强上下文信息。在PASCAL VOC 2012数据集上的实验结果表明，我们提出的方法在WSSS中实现了新的技术水平，验证了我们的CP表示和CPN的有效性。

Few-shot semantic segmentation aims at learning to segment a target object from a query image using only a few annotated support images of the target class. This challenging task requires to understand diverse levels of visual cues and analyze fine-grained correspondence relations between the query and the support images. To address the problem, we propose Hypercorrelation Squeeze Networks (HSNet) that leverages multi-level feature correlation and efficient 4D convolutions. It extracts diverse features from different levels of intermediate convolutional layers and constructs a collection of 4D correlation tensors, i.e., hypercorrelations. Using efficient center-pivot 4D convolutions in a pyramidal architecture, the method gradually squeezes high-level semantic and low-level geometric cues of the hypercorrelation into precise segmentation masks in coarse-to-fine manner. The significant performance improvements on standard few-shot segmentation benchmarks of PASCAL-5i, COCO-20i, and FSS-1000 verify the efficacy of the proposed method.

少镜头语义分割的目的是学习仅使用目标类的几个带注释的支持图像从查询图像中分割目标对象。这项具有挑战性的任务需要理解不同层次的视觉线索，并分析查询和支持图像之间的细粒度对应关系。为了解决这个问题，我们提出了超相关压缩网络 (HSNet)，它利用多级特征相关和有效的4D卷积。它从不同层次的中间卷积层中提取不同的特征，并构造一组4D相关张量，即超相关。该方法在金字塔结构中使用高效的中心轴4D卷积，以从粗到精的方式将超相关的高级语义和低级几何线索逐渐压缩到精确的分割模板中。PASCAL-5i、COCO-20i和FSS-1000标准少镜头分割基准的显著性能改进验证了该方法的有效性。

Self-supervised Multi-view stereo (MVS) with a pretext task of image reconstruction has achieved significant progress recently. However, previous methods are built upon intuitions, lacking comprehensive explanations about the effectiveness of the pretext task in self-supervised MVS. To this end, we propose to estimate epistemic uncertainty in self-supervised MVS, accounting for what the model ignores. Specially, the limitations can be resorted into two folds: ambiguous supervision in foreground and noisy disturbance in background. To address these issues, we propose a novel Uncertainty reduction Multi-view Stereo (U-MVS) framework for self-supervised learning. To alleviate ambiguous supervision in foreground, we involve extra correspondence prior with a flow-depth consistency loss. The dense 2D correspondence of optical flows is used to regularize the 3D stereo correspondence in MVS. To handle the noisy disturbance in background, we use Monte-Carlo Dropout to acquire the uncertainty map and further filter the unreliable supervision signals on invalid regions. Extensive experiments on DTU and Tank&Temples benchmark show that our U-MVS framework achieves the best performance among unsupervised MVS methods, with competitive performance with its supervised opponents.

近年来，以图像重建为借口的自监督多视点立体视觉 (MVS) 技术取得了重大进展。然而，以前的方法都是建立在直觉的基础上，缺乏对自我监督MVS中借口任务有效性的全面解释。为此，我们建议估计自我监督MVS中的认知不确定性，考虑模型忽略的内容。特别是，其局限性可以归结为两个方面：前景中的模糊监控和背景中的噪声干扰。为了解决这些问题，我们提出了一种新的用于自监督学习的不确定性减少多视点立体 (U-MVS) 框架。为了减轻前景中的模糊监控，我们在流深度一致性损失之前加入额外的通信。在MVS中，光流的稠密2D对应用于正则化3D立体对应。为了处理背景中的噪声干扰，我们使用蒙特卡罗差分法获取不确定性映射，并进一步过滤无效区域上的不可靠监控信号。在DTU和Tank&Temples基准测试上的大量实验表明，我们的U-MVS框架在无监督MVS方法中取得了最好的性能，其性能与有监督的对手相当。

DETR is a recently proposed Transformer-based method which views object detection as a set prediction problem and achieves state-of-the-art performance but demands extra-long training time to converge. In this paper, we investigate the causes of the optimization difficulty in the training of DETR. Our examinations reveal several factors contributing to the slow convergence of DETR, primarily the issues with the Hungarian loss and the Transformer cross attention mechanism. To overcome these issues we propose two solutions, namely, TSP-FCOS (Transformer-based Set Prediction with FCOS) and TSP-RCNN (Transformer-based Set Prediction with RCNN). Experimental results show that the proposed methods not only converge much faster than the original DETR, but also significantly outperform DETR and other baselines in terms of detection accuracy.

DETR是最近提出的一种基于变换器的方法，它将目标检测视为一个集合预测问题，实现了最先进的性能，但需要很长的训练时间才能收敛。本文探讨了DETR训练中优化困难的原因。我们的研究揭示了导致DETR缓慢收敛的几个因素，主要是匈牙利损失和变压器交叉注意机制的问题。为了克服这些问题，我们提出了两种解决方案，即TSP-FCOS（基于变压器的FCOS集预测）和TSP-RCNN（基于变压器的RCNN集预测）。实验结果表明，该方法不仅比原DETR收敛速度快得多，而且在检测精度方面也明显优于DETR和其他基线。

We present a simple yet effective unpaired learning based image rain removal method from an unpaired set of synthetic images and real rainy images by exploring the properties of rain maps. The proposed algorithm mainly consists of a semi-supervised learning part and a knowledge distillation part. The semi-supervised part estimates the rain map and reconstructs the derained image based on the well-established layer separation principle. To facilitate rain removal, we develop a rain direction regularizer to constrain the rain estimation network in the semi-supervised learning part. With the estimated rain maps from the semi-supervised learning part, we first synthesize a new paired set by adding to rain-free images based on the superimposition model. The real rainy images and the derained results constitute another paired set. Then we develop an effective knowledge distillation method to explore such two paired sets so that the deraining model in the semi-supervised learning part is distilled. We propose two new rainy datasets, named RainDirection and Real3000, to validate the effectiveness of the proposed method. Both quantitative and qualitative experimental results demonstrate that the proposed method achieves favorable results against state-of-the-art methods in benchmark datasets and real-world images.

通过探索雨图的性质，我们提出了一种简单而有效的基于非配对学习的图像雨去除方法，该方法从一组未配对的合成图像和真实的雨图像中去除雨。该算法主要由半监督学习部分和知识提取部分组成。半监督部分根据成熟的层分离原理估计雨图并重建降维图像。为了便于雨水去除，我们开发了一种雨水方向正则化器，用于在半监督学习部分约束雨水估计网络。利用半监督学习部分估计的雨图，我们首先在叠加模型的基础上，通过添加到无雨图像中来合成一个新的成对集。真实的雨天图像和去定义的结果构成另一对集合。然后，我们开发了一种有效的方法来提取这两个配对集，从而提取半监督学习部分的降额模型。我们提出了两个新的雨点数据集RainDirection和Real3000，以验证所提出方法的有效性。定量和定性实验结果表明，该方法在基准数据集和真实图像中取得了良好的效果。

The lack of clean images undermines the practicability of supervised image prior learning methods, of which the training schemes require a large number of clean images. To free image prior learning from the image collection burden, a novel self-supervised learning method for Gaussian Mixture Model (SS-GMM) is proposed in this paper. It can simultaneously achieve the noise level estimation and the image prior learning directly from only a single noisy image. This work is derived from our study on eigenvalues of the GMM's covariance matrix. Through statistical experiments and theoretical analysis, we conclude that (1) covariance eigenvalues for clean images hold the sparsity; and that (2) those for noisy images contain sufficient information for noise estimation. The first conclusion inspires us to impose a sparsity constraint on covariance eigenvalues during the learning process to suppress the influence of noise. The second conclusion leads to a self-contained noise estimation module of high accuracy in our proposed method. This module serves to estimate the noise level and automatically determine the specific level of the sparsity constraint. Our final derived method requires only minor modifications to the standard expectation-maximization algorithm. This makes it easy to implement. Very interestingly, the GMM learned via our proposed self-supervised learning method can even achieve better image denoising performance than its supervised counterpart, i.e., the EPLL. Also, it is on par with the state-of-the-art self-supervised deep learning method, i.e., the Self2Self. Code is available at <https://github.com/HUST-Tan/SS-GMM>.

干净图像的缺乏削弱了有监督图像先验学习方法的实用性，其中训练方案需要大量干净图像。为了减轻图像采集负担，提出了一种新的高斯混合模型自监督学习方法（SS-GMM）。它可以直接从单个噪声图像中同时实现噪声水平估计和图像先验学习。这项工作来源于我们对GMM协方差矩阵特征值的研究。通过统计实验和理论分析，我们得出以下结论：（1）干净图像的协方差特征值具有稀疏性；（2）对于噪声图像，包含足够的噪声估计信息。第一个结论启发我们在学习过程中对协方差特征值施加稀疏约束，以抑制噪声的影响。第二个结论是在我们提出的方法中引入了一个自包含的高精度噪声估计模块。该模块用于估计噪声级并自动确定稀疏性约束的特定级别。我们最终得出的方法只需要对标准的期望最大化算法进行微小的修改。这使得它易于实现。非常有趣的是，通过我们提出的自监督学习方法学习的GMM甚至可以比有监督的对应方法（即EPLL）获得更好的图像去噪性能。此外，它也与现有的自监督深度学习方法（即SELF2SUBE）相一致。代码可在<https://github.com/HUST-Tan/SS-GMM>。

Autonomous driving has attracted much attention over the years but turns out to be harder than expected, probably due to the difficulty of labeled data collection for model training. Self-supervised learning (SSL), which leverages unlabeled data only for representation learning, might be a promising way to improve model performance. Existing SSL methods, however, usually rely on the single-centric-object guarantee, which may not be applicable for multi-instance datasets such as street scenes. To alleviate this limitation, we raise two issues to solve: (1) how to define positive samples for cross-view consistency and (2) how to measure similarity in multi-instance circumstances. We first adopt an IoU threshold during random cropping to transfer global-inconsistency to local-consistency. Then, we propose two feature alignment methods to enable 2D feature maps for multi-instance similarity measurement. Additionally, we adopt intra-image clustering with self-attention for further mining intra-image similarity and translation-invariance. Experiments show that, when pre-trained on Waymo dataset, our method called Multi-instance Siamese Network (Multisiam) remarkably improves generalization ability and achieves state-of-the-art transfer performance on autonomous driving benchmarks, including Cityscapes and BDD100K, while existing SSL counterparts like MoCo, MoCo-v2, and BYOL show significant performance drop. By pre-training on SODA10M, a large-scale autonomous driving dataset, Multisiam exceeds the ImageNet pre-trained MoCo-v2, demonstrating the potential of domain-specific pre-training. Code will be available at <https://github.com/Kaichen1998/Multisiam> .

多年来，自动驾驶引起了人们的广泛关注，但事实证明它比预期的更难，这可能是因为模型训练中难以收集标记数据。自监督学习（SSL）仅利用未标记数据进行表示学习，可能是提高模型性能的一种很有前途的方法。然而，现有的SSL方法通常依赖于单中心对象保证，这可能不适用于街道场景等多实例数据集。为了缓解这一限制，我们提出了两个需要解决的问题：(1) 如何定义交叉视图一致性的正样本；(2) 如何在多实例情况下度量相似度。我们首先在随机裁剪过程中采用IoU阈值，将全局不一致性转化为局部一致性。然后，我们提出了两种特征对齐方法，使二维特征映射能够进行多实例相似性度量。此外，我们采用具有自关注的图像内聚类方法进一步挖掘图像内的相似性和翻译不变性。实验表明，当在Waymo数据集上进行预训练时，我们称之为多实例暹罗网络（MultiSiam）的方法显著提高了泛化能力，并在自动驾驶基准（包括Cityscapes和BDD100K）上实现了最先进的传输性能，而现有的SSL对应物（如MoCo、MoCo-v2、BYOL）的性能明显下降。通过对大规模自动驾驶数据集SODA1000进行预培训，MultiSiam超过了ImageNet预训练的MoCo-v2，展示了特定领域预培训的潜力。代码将在<https://github.com/KaiChen1998/MultiSiam>。

360deg videos convey holistic views for the surroundings of a scene. It provides audio-visual cues beyond predetermined normal field of views and displays distinctive spatial relations on a sphere. However, previous benchmark tasks for panoramic videos are still limited to evaluate the semantic understanding of audio-visual relationships or spherical spatial property in surroundings. We propose a novel benchmark named Pano-AVQA as a large-scale grounded audio-visual question answering dataset on panoramic videos. Using 5.4K 360deg video clips harvested online, we collect two types of novel question-answer pairs with bounding-box grounding: spherical spatial relation QAs and audio-visual relation QAs. We train several transformer-based models from Pano-AVQA, where the results suggest that our proposed spherical spatial embeddings and multimodal training objectives fairly contribute to better semantic understanding of the panoramic surroundings on the dataset.

360度视频传达场景周围的整体视图。它提供超出预定正常视野的视听线索，并在球体上显示独特的空间关系。然而，以前的全景视频基准测试任务仍然局限于评估环境中视听关系或球形空间属性的语义理解。我们提出了一个新的基准测试panoavqa，作为一个基于全景视频的大规模地面视听问答数据集。使用在线采集的5.4K 360度视频剪辑，我们收集了两种具有边界框接地的新颖问答对：球形空间关系QAs和视听关系QAs。我们从Pano AVQA训练了几个基于变换器的模型，结果表明我们提出的球形空间嵌入和多模态训练目标有助于更好地理解数据集上的全景环境。

Most successful self-supervised learning methods are trained to align the representations of two independent views from the data. State-of-the-art methods in video are inspired by image techniques, where these two views are similarly extracted by cropping and augmenting the resulting crop. However, these methods miss a crucial element in the video domain: time. We introduce BraVe, a self-supervised learning framework for video. In BraVe, one of the views has access to a narrow temporal window of the video while the other view has a broad access to the video content. Our models learn to generalise from the narrow view to the general content of the video. Furthermore, BraVe processes the views with different backbones, enabling the use of alternative augmentations or modalities into the broad view such as optical flow, randomly convolved RGB frames, audio or their combinations. We demonstrate that BraVe achieves state-of-the-art results in self-supervised representation learning on standard video and audio classification benchmarks including UCF101, HMDB51, Kinetics, ESC-50 and AudioSet.

大多数成功的自监督学习方法都经过训练，能够从数据中对齐两个独立视图的表示。视频中最先进的方法受到图像技术的启发，在图像技术中，这两个视图通过裁剪和增强生成的裁剪进行类似的提取。然而，这些方法忽略了视频领域的一个关键元素：时间。我们介绍了BraVe，一个视频自监督学习框架。在BraVe中，一个视图可以访问视频的窄时间窗口，而另一个视图可以广泛访问视频内容。我们的模型

学习从狭隘的视角概括到视频的一般内容。此外，BraVe使用不同的主干处理视图，允许在广域视图中使用替代增强或模式，如光流、随机卷积RGB帧、音频或其组合。我们证明了BraVe在标准视频和音频分类基准（包括UCF101、HMDB51、Kinetics、ESC-50和AudioSet）上实现了自我监督表征学习的最新成果。

Many video analysis tasks require temporal localization for the detection of content changes. However, most existing models developed for these tasks are pre-trained on general video action classification tasks. This is due to large scale annotation of temporal boundaries in untrimmed videos being expensive. Therefore, no suitable datasets exist that enable pre-training in a manner sensitive to temporal boundaries. In this paper for the first time, we investigate model pre-training for temporal localization by introducing a novel boundary-sensitive pretext (BSP) task. Instead of relying on costly manual annotations of temporal boundaries, we propose to synthesize temporal boundaries in existing video action classification datasets. By defining different ways of synthesizing boundaries, BSP can then be simply conducted in a self-supervised manner via the classification of the boundary types. This enables the learning of video representations that are much more transferable to downstream temporal localization tasks. Extensive experiments show that the proposed BSP is superior and complementary to the existing action classification-based pre-training counterpart, and achieves new state-of-the-art performance on several temporal localization tasks. Please visit our website for more details

<https://frostinassiky.github.io/bsp>.

许多视频分析任务需要时间定位来检测内容变化。然而，大多数为这些任务开发的现有模型都是针对一般视频动作分类任务进行预训练的。这是因为在未经剪辑的视频中对时间边界进行大规模注释成本高昂。因此，不存在能够以对时间边界敏感的方式进行预训练的合适数据集。在本文中，我们首次通过引入一种新的边界敏感借口（BSP）任务来研究时间定位的模型预训练。我们建议在现有的视频动作分类数据集中合成时间边界，而不是依赖于昂贵的时间边界手工标注。通过定义合成边界的不同方式，BSP可以通过边界类型的分类以自我监督的方式简单地进行。这使得视频表示的学习更容易转移到下游的时间定位任务中。大量实验表明，该算法优于现有的基于动作分类的预训练算法，并在多个时间定位任务上取得了新的性能。请访问我们的网站了解更多详细信息<https://frostinassiky.github.io/bsp>.

Video entailment aims at determining if a hypothesis textual statement is entailed or contradicted by a premise video. The main challenge of video entailment is that it requires fine-grained reasoning to understand the complex and long story-based videos. To this end, we propose to incorporate visual grounding to the entailment by explicitly linking the entities described in the statement to the evidence in the video. If the entities are grounded in the video, we enhance the entailment judgment by focusing on the frames where the entities occur. Besides, in entailment dataset, the real/fake statements are formed in pairs with subtle discrepancy, which allows an add-on explanation module to predict which words or phrases make the statement contradictory to the video and regularize the training of the entailment judgment. Experimental results demonstrate that our approach significantly outperforms the state-of-the-art methods.

视频蕴涵旨在确定假设文本陈述是否被前提视频蕴涵或矛盾。视频蕴涵的主要挑战是，它需要细粒度的推理来理解复杂的、基于长故事的视频。为此，我们建议通过明确地将陈述中描述的实体与视频中的证据联系起来，将视觉基础与蕴涵结合起来。如果实体在视频中扎根，我们通过聚焦实体出现的帧来增强蕴涵判断。此外，在蕴涵数据集中，真实/虚假陈述成对形成，存在细微差异，这允许附加解释模块预测哪些单词或短语使陈述与视频相矛盾，并规范蕴涵判断的训练。实验结果表明，我们的方法明显优于现有的方法。

Recent learning-based multi-view stereo (MVS) methods show excellent performance with dense cameras and small depth ranges. However, non-learning based approaches still outperform for scenes with large depth ranges and sparser wide-baseline views, in part due to their PatchMatch optimization over pixelwise estimates of depth, normals, and visibility. In this paper, we propose an end-to-end trainable PatchMatch-based MVS approach that combines advantages of trainable costs and regularizations with pixelwise estimates. To overcome the challenge of the non-differentiable PatchMatch optimization that involves iterative sampling and hard decisions, we use reinforcement learning to minimize expected photometric cost and maximize likelihood of ground truth depth and normals. We incorporate normal estimation by using dilated patch kernels, and propose a recurrent cost regularization that applies beyond frontal plane-sweep algorithms to our pixelwise depth/normal estimates. We evaluate our method on widely used MVS benchmarks, ETH3D and Tanks and Temples (TnT), and compare to other state of the art learning based MVS models. On ETH3D, our method outperforms other recent learning-based approaches and performs comparably on advanced TnT.

最近的基于学习的多视点立体 (MVS) 方法在摄像机密集、深度范围小的情况下表现出优异的性能。然而，对于深度范围较大且基线视图较稀疏的场景，非学习方法仍优于其他方法，部分原因是它们对深度、法线和可见性的像素级估计进行了补丁匹配优化。在本文中，我们提出了一种端到端可训练的基于补丁匹配的MVS方法，该方法将可训练成本和正则化的优点与像素估计相结合。为了克服涉及迭代采样和硬决策的不可微修补匹配优化的挑战，我们使用强化学习来最小化预期光度成本，并最大化地真实深度和法线的可能性。我们通过使用扩展的面片核来合并法线估计，并提出了一种循环成本正则化，该正则化应用于我们的像素深度/法线估计中的超前沿平面扫描算法。我们在广泛使用的MVS基准、ETH3D 和坦克庙宇 (TnT) 上评估了我们的方法，并将其与其他基于最先进学习的MVS模型进行了比较。在ETH3D上，我们的方法优于其他最近基于学习的方法，并且在高级TnT上的性能相当。

Recent state-of-the-art learning-based approaches to point cloud registration have largely been based on graph neural networks (GNN). However, these prominent GNN backbones suffer from the indistinguishable features problem associated with over-smoothing and structural ambiguity of the high-level features, a crucial bottleneck to point cloud registration that has evaded scrutiny in the recent relevant literature. To address this issue, we propose the Distinctiveness oriented Positional Equilibrium (DoPE) module, a novel positional embedding scheme that significantly improves the distinctiveness of the high-level features within both the source and target point clouds, resulting in superior point matching and hence registration accuracy. Specifically, we use the DoPE module in an iterative registration framework, whereby the two point clouds are gradually registered via rigid transformations that are computed from DoPE's position-aware features. With every successive iteration, the DoPE module feeds increasingly consistent positional information to would-be corresponding pairs, which in turn enhances the resulting point-to-point correspondence predictions used to estimate the rigid transformation. Within only a few iterations, the network converges to a desired equilibrium, where the positional embeddings given to matching pairs become essentially identical. We validate the effectiveness of DoPE through comprehensive experiments on various registration benchmarks, registration task settings, and prominent backbones, yielding unprecedented performance improvement across all combinations.

最近基于学习的最新点云配准方法主要基于图形神经网络 (GNN)。然而，这些突出的GNN主干受到与高级特征的过度平滑和结构模糊相关的难以区分的特征问题的困扰，这是点云注册的一个关键瓶颈，在最近的相关文献中，这一瓶颈回避了详细审查。为了解决这个问题，我们提出了面向显著性位置平衡 (DoPE) 模块，这是一种新的位置嵌入方案，它显著提高了源点云和目标点云中高级特征的显著性，从而实现了更高的点匹配和配准精度。具体地说，我们在迭代注册框架中使用DoPE模块，通过从DoPE的位置感知特征计算的刚性变换逐渐注册两点云。在每次连续迭代中，“摄影”模块向可能的对应提供越

来越一致的位置信息，这反过来增强了用于估计刚性变换的结果点到点对应预测。只需几次迭代，网络就会收敛到所需的平衡点，在这个平衡点上，匹配对的位置嵌入变得基本相同。我们通过在各种注册基准、注册任务设置和突出主干上的综合实验来验证DoPE的有效性，从而在所有组合中产生前所未有的性能改进。

We present the novel Efficient Line Segment Detector and Descriptor (ELSD) to simultaneously detect line segments and extract their descriptors in an image. Unlike the traditional pipelines that conduct detection and description separately, ELSD utilizes a shared feature extractor for both detection and description, to provide the essential line features to the higher-level tasks like SLAM and image matching in real time. First, we design a one-stage compact model, and propose to use the mid-point, angle and length as the minimal representation of line segment, which also guarantees the center-symmetry. The non-centerness suppression is proposed to filter out the fragmented line segments caused by lines' intersections. The fine offset prediction is designed to refine the mid-point localization. Second, the line descriptor branch is integrated with the detector branch, and the two branches are jointly trained in an end-to-end manner. In the experiments, the proposed ELSD achieves the state-of-the-art performance on the Wireframe dataset and YorkUrban dataset, in both accuracy and efficiency. The line description ability of ELSD also outperforms the previous works on the line matching task.

我们提出了一种新的高效线段检测器和描述符 (ELSD)，用于同时检测图像中的线段并提取其描述符。与传统的分别进行检测和描述的管道不同，ELSD利用共享特征提取器进行检测和描述，为SLAM和图像匹配等更高级别的任务实时提供基本的线条特征。首先，我们设计了一个单阶段的紧凑模型，并建议使用中点、角度和长度作为线段的最小表示，这也保证了中心对称性。提出了一种非中心抑制方法来滤除由直线相交引起的线段碎片。精细偏移量预测用于细化中点定位。其次，行描述符分支与检测器分支集成，并以端到端的方式联合训练这两个分支。在实验中，提出的ELSD在线框数据集和YorkUrban数据集上实现了最先进的性能，在准确性和效率上都达到了最高水平。ELSD的行描述能力也优于以往的行匹配任务。

Real world images often gets corrupted due to unwanted reflections and their removal is highly desirable. A major share of such images originate from smart phone cameras capable of very high resolution captures. Most of the existing methods either focus on restoration quality by compromising on processing speed and memory requirements or, focus on removing reflections at very low resolutions, thereby limiting their practical deploy-ability. We propose a light weight deep learning model for reflection removal using a novel scale space architecture. Our method processes the corrupted image in two stages, a Low Scale Sub-network (LSSNet) to process the lowest scale and a Progressive Inference (PI) stage to process all the higher scales. In order to reduce the computational complexity, the sub-networks in PI stage are designed to be much shallower than LSSNet. Moreover, we employ weight sharing between various scales within the PI stage to limit the model size. This also allows our method to generalize to very high resolutions without explicit retraining. Our method is superior both qualitatively and quantitatively compared to the state of the art methods and at the same time 20x faster with 50x less number of parameters compared to the most recent state-of-the-art algorithm RAGNet. We implemented our method on an android smart phone, where a high resolution 12 MP image is restored in under 5 seconds.

现实世界中的图像通常会由于不必要的反射而损坏，因此非常需要将其删除。这些图像中的大部分来自能够进行高分辨率拍摄的智能手机摄像头。现有的大多数方法要么通过降低处理速度和内存需求来关注恢复质量，要么通过限制实际部署能力来关注在极低分辨率下消除反射。我们提出了一个轻量级的深度学习模型，使用一种新的尺度空间结构去除反射。我们的方法分两个阶段处理损坏的图像，一个低尺度子网络 (LSSNet) 处理最低尺度，一个渐进推理 (PI) 阶段处理所有较高尺度。为了降低计算复杂度，

PI阶段的子网络被设计为比LSSNet浅得多。此外，我们在PI阶段使用不同尺度之间的权重共享来限制模型大小。这也使得我们的方法可以推广到非常高的分辨率，而无需显式的重新训练。与最先进的方法相比，我们的方法在定性和定量上都优于最先进的方法，同时与最新最先进的算法RAGNet相比，我们的方法速度快20倍，参数数量少50倍。我们在安卓智能手机上实现了我们的方法，在5秒钟内恢复高分辨率12MP图像。

Cracks are irregular line structures that are of interest in many computer vision applications. Crack detection (e.g., from pavement images) is a challenging task due to intensity in-homogeneity, topology complexity, low contrast and noisy background. The overall crack detection accuracy can be significantly affected by the detection performance on fine-grained cracks. In this work, we propose a Crack Transformer network (CrackFormer) for fine-grained crack detection. The CrackFormer is composed of novel attention modules in a SegNet-like encoder-decoder architecture. Specifically, it consists of novel self-attention modules with 1x1 convolutional kernels for efficient contextual information extraction across feature-channels, and efficient positional embedding to capture large receptive field contextual information for long range interactions. It also introduces new scaling-attention modules to combine outputs from the corresponding encoder and decoder blocks to suppress non-semantic features and sharpen semantic cracks. The CrackFormer is trained and evaluated on three classical crack datasets. The experimental results show that CrackFormer achieves ODS values of 0.871, 0.877 and 0.881, respectively, on the three datasets and outperforms the state-of-the-art methods.

裂纹是许多计算机视觉应用中感兴趣的不规则线结构。裂缝检测（例如，从路面图像）是一项具有挑战性的任务，因为强度均匀，拓扑结构复杂，对比度低，背景噪声大。细粒度裂纹的检测性能会显著影响整体裂纹检测精度。在这项工作中，我们提出了一种用于细粒度裂纹检测的裂纹变压器网络（CrackFormer）。CrackFormer由新型注意力模块组成，采用类似SegNet的编解码器结构。具体而言，它包括具有1x1卷积核的新型自我注意模块，用于跨特征通道有效提取上下文信息，以及用于捕获大的感受野上下文信息以进行远程交互的有效位置嵌入。它还引入了新的缩放注意模块，以组合来自相应编码器和解码器块的输出，以抑制非语义特征并锐化语义裂缝。CrackFormer在三个经典裂纹数据集上进行训练和评估。实验结果表明，CrackFormer在三个数据集上的ODS值分别为0.871、0.877和0.881，优于最新的方法。

Performing simple household tasks based on language directives is very natural to humans, yet it remains an open challenge for an AI agent. The 'interactive instruction following' task attempts to make progress towards building an agent that can jointly navigate, interact, and reason in the environment at every step. To address the multifaceted problem, we propose a model that factorizes the task into interactive perception and action policy streams with enhanced components. We empirically validate that our model outperforms prior arts by significant margins on the ALFRED benchmark in all metrics with improved generalization.

基于语言指令执行简单的家庭任务对人类来说是非常自然的，但对于人工智能代理来说，这仍然是一个公开的挑战。“交互式指令跟踪”任务试图在构建一个代理方面取得进展，该代理可以在环境中的每一步进行联合导航、交互和推理。为了解决这个多方面的问题，我们提出了一个模型，该模型将任务分解为具有增强组件的交互式感知和行动策略流。我们通过经验验证，我们的模型在所有指标上都优于现有技术，在阿尔弗雷德基准上有显著的优势，并且具有更好的泛化能力。

Monocular object detection and tracking have improved drastically in recent years, but rely on a key assumption: that objects are visible to the camera. Many offline tracking approaches reason about occluded objects post-hoc, by linking together tracklets after the object re-appears, making use of reidentification (ReID). However, online tracking in embodied robotic agents (such as a self-driving vehicle) fundamentally requires object permanence, which is the ability to reason about occluded objects before they re-appear. In this work, we re-purpose tracking benchmarks and propose new metrics for the task of detecting invisible objects, focusing on the illustrative case of people. We demonstrate that current detection and tracking systems perform dramatically worse on this task. We introduce two key innovations to recover much of this performance drop. We treat occluded object detection in temporal sequences as a short-term forecasting challenge, bringing to bear tools from dynamic sequence prediction. Second, we build dynamic models that explicitly reason in 3D from monocular videos without calibration, using observations produced by monocular depth estimators. To our knowledge, ours is the first work to demonstrate the effectiveness of monocular depth estimation for the task of tracking and detecting occluded objects. Our approach strongly improves by 11.4% over the baseline in ablations and by 5.0% over the state-of-the-art in F1 score.

近几年来，单目目标检测和跟踪技术有了很大的进步，但这取决于一个关键假设：目标对相机是可见的。许多离线跟踪方法通过在对象重新出现后将tracklet链接在一起，利用重新识别（ReID）来事后推断被遮挡的对象。然而，嵌入式机器人代理（如自动驾驶车辆）中的在线跟踪从根本上要求对象持久性，即在被遮挡对象再次出现之前对其进行推理的能力。在这项工作中，我们重新设计了目标跟踪基准，并针对检测不可见对象的任务提出了新的度量标准，重点放在人的示例上。我们证明了当前的检测和跟踪系统在这项任务上的表现非常糟糕。我们引入了两项关键创新来弥补大部分性能下降。我们将时间序列中的遮挡目标检测视为短期预测挑战，从而带来了来自动态序列预测的工具。第二，我们使用单目深度估计器产生的观测值，在无需校准的情况下，从单目视频中建立动态模型，在3D中明确推理。据我们所知，我们的工作是第一次证明单目深度估计在跟踪和检测遮挡目标任务中的有效性。我们的方法在消融方面比基线水平提高了11.4%，在F1成绩方面比最先进水平提高了5.0%。

weakly-supervised object detection (wsod) has emerged as an inspiring recent topic to avoid expensive instance-level object annotations. However, the bounding boxes of most existing wsod methods are mainly determined by precomputed proposals, thereby being limited in precise object localization. In this paper, we defend the problem setting for improving localization performance by leveraging the bounding box regression knowledge from a well-annotated auxiliary dataset. First, we use the well-annotated auxiliary dataset to explore a series of learnable bounding box adjusters (LBBAs) in a multi-stage training manner, which is class-agnostic. Then, only LBBAs and a weakly-annotated dataset with non-overlapped classes are used for training LBBA-boosted wsod. As such, our LBBAs are practically more convenient and economical to implement while avoiding the leakage of the auxiliary well-annotated dataset. In particular, we formulate learning bounding box adjusters as a bi-level optimization problem and suggest an EM-like multi-stage training algorithm. Then, a multi-stage scheme is further presented for LBBA-boosted wsod. Additionally, a masking strategy is adopted to improve proposal classification. Experimental results verify the effectiveness of our method. Our method performs favorably against state-of-the-art wsod methods and knowledge transfer model with similar problem setting. Code is publicly available at [https://github.com/DongSky/lbba\\_boosted\\_wsod](https://github.com/DongSky/lbba_boosted_wsod).

弱监督对象检测（WSOD）已成为避免昂贵的实例级对象注释的最新主题。然而，大多数现有WSOD方法的边界框主要由预先计算的方案确定，因此在精确的目标定位方面受到限制。在本文中，我们通过利用来自注释良好的辅助数据集的边界框回归知识，为提高本地化性能的问题设置辩护。首先，我们使用注释良好的辅助数据集，以多阶段训练的方式探索一系列可学习的边界框调整器（LBBA），这是类不可

知的。然后，只使用LBBA和具有非重叠类的弱注释数据集来训练LBBA增强的WSOD。因此，我们的LBBA实际上更方便、更经济地实现，同时避免了辅助并注数据集的泄漏。特别地，我们将学习包围盒调整器描述为一个双层优化问题，并提出了一种类似EM的多级训练算法。然后，针对LBBA增强的WSOD，进一步提出了一种多级方案。此外，采用掩蔽策略来改进提案分类。实验结果验证了该方法的有效性。我们的方法与最先进的WSOD方法和具有类似问题设置的知识转移模型相比，具有良好的性能。该守则可于<https://github.com/DongSky/lbba boosted wsod>。

The convolutional neural network (CNN) is vulnerable to degraded images with even very small variations (e.g. corrupted and adversarial samples). One of the possible reasons is that CNN pays more attention to the most discriminative regions, but ignores the auxiliary features when learning, leading to the lack of feature diversity for final judgment. In our method, we propose to dynamically suppress significant activation values of CNN by group-wise inhibition, but not fixedly or randomly handle them when training. The feature maps with different activation distribution are then processed separately to take the feature independence into account. CNN is finally guided to learn richer discriminative features hierarchically for robust classification according to the proposed regularization. Our method is comprehensively evaluated under multiple settings, including classification against corruptions, adversarial attacks and low data regime. Extensive experimental results show that the proposed method can achieve significant improvements in terms of both robustness and generalization performances, when compared with the state-of-the-art methods. Code is available at [https://github.com/LinusWu/TENET\\_Training](https://github.com/LinusWu/TENET_Training).

卷积神经网络 (CNN) 易受退化图像的影响，即使变化很小（例如损坏和敌对样本）。其中一个可能的原因是，CNN更多地关注最具辨别力的区域，而在学习时忽略了辅助特征，导致最终判断缺乏特征多样性。在我们的方法中，我们建议通过群体抑制来动态抑制CNN的显著激活值，而不是在训练时固定或随机处理它们。然后分别处理具有不同激活分布的特征映射，以考虑特征独立性。最后，根据所提出的正则化方法，引导CNN分层学习更丰富的鉴别特征，实现鲁棒分类。我们的方法在多种环境下进行了综合评估，包括针对腐蚀、对抗性攻击和低数据模式的分类。大量的实验结果表明，与现有的方法相比，该方法在鲁棒性和泛化性能方面都有显著的提高。代码可在[https://github.com/LinusWu/TENET\\_Training](https://github.com/LinusWu/TENET_Training) 培训。

Motivated by the success of Transformers in natural language processing (NLP) tasks, there exist some attempts (e.g., ViT and DeiT) to apply Transformers to the vision domain. However, pure Transformer architectures often require a large amount of training data or extra supervision to obtain comparable performance with convolutional neural networks (CNNs). To overcome these limitations, we analyze the potential drawbacks when directly borrowing Transformer architectures from NLP. Then we propose a new Convolution-enhanced image Transformer (CeIT) which combines the advantages of CNNs in extracting low-level features, strengthening locality, and the advantages of Transformers in establishing long-range dependencies. Three modifications are made to the original Transformer: 1) instead of the straightforward tokenization from raw input images, we design an Image-to-Tokens (I2T) module that extracts patches from generated low-level features; 2) the feed-forward network in each encoder block is replaced with a Locally-enhanced Feed-Forward (LeFF) layer that promotes the correlation among neighboring tokens in the spatial dimension; 3) a Layer-wise Class token Attention (LCA) is attached at the top of the Transformer that utilizes the multi-level representations. Experimental results on ImageNet and seven downstream tasks show the effectiveness and generalization ability compared with previous Transformers and state-of-the-art CNNs, without requiring a large amount of training data and extra CNN teachers. Besides, CeIT models also demonstrate better convergence with 3x fewer training iterations, which can reduce the training cost significantly.

由于自然语言处理 (NLP) 任务中变形金刚的成功，有一些尝试（如ViT和DeiT）将变形金刚应用于视觉领域。然而，纯变压器结构通常需要大量的训练数据或额外的监督才能获得与卷积神经网络 (CNN) 相当的性能。为了克服这些限制，我们分析了直接从NLP借用Transformer架构时可能存在的缺点。然后，我们提出了一种新的卷积增强图像变换器 (CeIT)，它结合了CNN在提取低级特征、增强局部性方面的优势，以及变换器在建立远程依赖方面的优势。对原始转换器进行了三个修改：1) 我们设计了一个图像到标记 (I2T) 模块，从生成的低级特征中提取补丁，而不是直接从原始输入图像进行标记化；2) 将每个编码器块中的前馈网络替换为局部增强前馈 (LeFF) 层，该层促进空间维度中相邻令牌之间的相关性；3) 一个分层类标记注意 (LCA) 附加在转换器的顶部，该转换器使用多级表示。在ImageNet和七个下游任务上的实验结果表明，与以前的Transformer和最先进的CNN相比，它具有有效性和泛化能力，而不需要大量的培训数据和额外的CNN教师。此外，CeIT模型还具有更好的收敛性和更少的训练迭代次数，这可以显著降低训练成本。

A large gap exists between fully-supervised object detection and weakly-supervised object detection. To narrow this gap, some methods consider knowledge transfer from additional fully-supervised dataset. But these methods do not fully exploit discriminative category information in the fully-supervised dataset, thus causing low mAP. To solve this issue, we propose a novel category transfer framework for weakly supervised object detection. The intuition is to fully leverage both visually-discriminative and semantically-correlated category information in the fully-supervised dataset to enhance the object-classification ability of a weakly-supervised detector. To handle overlapping category transfer, we propose a double-supervision mean teacher to gather common category information and bridge the domain gap between two datasets. To handle non-overlapping category transfer, we propose a semantic graph convolutional network to promote the aggregation of semantic features between correlated categories. Experiments are conducted with Pascal VOC 2007 as the target weakly-supervised dataset and COCO as the source fully-supervised dataset. Our category transfer framework achieves 63.5% mAP and 80.3% CorLoc with 5 overlapping categories between two datasets, which outperforms the state-of-the-art methods. Codes are available at <https://github.com/MediaBrain-SJTU/CaT>.

全监督目标检测与弱监督目标检测之间存在较大差距。为了缩小这一差距，一些方法考虑从额外的全监督数据集的知识转移。但这些方法并没有充分利用全监督数据集中的区分性类别信息，从而导致低mAP。为了解决这个问题，我们提出了一种新的类别转移框架，用于弱监督目标检测。直觉是充分利用全监督数据集中的视觉辨别性和语义相关的类别信息来增强弱监督检测器的对象分类能力。为了处理重叠的类别转移，我们提出了一种双重监督的方法来收集共同的类别信息并弥合两个数据集之间的领域鸿沟。为了处理不重叠的类别转移，我们提出了一种语义图卷积网络来促进相关类别之间语义特征的聚合。实验以Pascal VOC 2007为目标弱监督数据集，COCO为源全监督数据集。我们的类别转移框架实现了63.5%的mAP和80.3%的CorLoc，两个数据集之间有5个重叠类别，这优于最先进的方法。代码可在<https://github.com/MediaBrain-SJTU/CaT>.

The state-of-the-art object detection and image classification methods can perform impressively on more than 9k and 10k classes respectively. In contrast, the number of classes in semantic segmentation datasets is relatively limited. This is not surprising when the restrictions caused by the lack of labelled data and high computation demand for segmentation are considered. In this paper, we propose a novel training methodology to train and scale the existing semantic segmentation models for a large number of semantic classes without increasing the memory overhead. In our approach, we reduce the space complexity of the segmentation model's output from  $O(C)$  to  $O(1)$ , propose an approximation method for ground-truth class probability, and use it to compute cross-entropy loss. The proposed approach is general and can be adopted by any state-of-the-art segmentation model to gracefully scale it for any number of semantic classes with only one GPU. Our approach achieves similar, and in some cases even better mIoU for Cityscapes, Pascal VOC and ADE20k dataset when adopted to DeeplabV3+ model with different backbones. We demonstrate a clear benefit of our approach on a dataset with 1284 classes, bootstrapped from LVIS and COCO annotations, with almost three times better mIoU when compared to DeeplabV3+. Code is available at: <https://github.com/shipra25jain/ESSNet>.

最先进的目标检测和图像分类方法可以分别在9k和10k以上的类别上表现出色。相比之下，语义分割数据集中的类数量相对有限。当考虑到缺乏标记数据和分割的高计算需求所造成的限制时，这并不奇怪。在本文中，我们提出了一种新的训练方法来训练和扩展大量语义类的现有语义分割模型，而不增加内存开销。在我们的方法中，我们将分割模型输出的空间复杂度从 $O(C)$ 降低到 $O(1)$ ，提出了一种基本真值类概率的近似方法，并将其用于计算交叉熵损失。所提出的方法是通用的，任何最先进的分割模型都可以采用这种方法，只需一个GPU就可以对任意数量的语义类进行优雅的缩放。当采用不同主干的DeeplabV3+模型时，我们的方法在城市景观、Pascal VOC和ADE20k数据集上实现了类似的，在某些情况下甚至更好的mIoU。我们在一个包含1284个类的数据集上展示了我们的方法的明显优势，该数据集由LVIS和COCO注释引导，mIoU几乎是DeeplabV3+的三倍。代码可从以下网址获取：<https://github.com/shipra25jain/ESSNet>。

We present a "learning to learn" approach for discovering white-box classification loss functions that are robust to label noise in the training data. We parameterise a flexible family of loss functions using Taylor polynomials, and apply evolutionary strategies to search for noise-robust losses in this space. To learn re-usable loss functions that can apply to new tasks, our fitness function scores their performance in aggregate across a range of training datasets and architectures. The resulting white-box loss provides a simple and fast "plug-and-play" module that enables effective label-noise-robust learning in diverse downstream tasks, without requiring a special training procedure or network architecture. The efficacy of our loss is demonstrated on a variety of datasets with both synthetic and real label noise, where we compare favourably to prior work.

我们提出了一种“从学习到学习”的方法来发现白盒分类损失函数，它对训练数据中的噪声具有鲁棒性。我们使用泰勒多项式参数化了一个灵活的损失函数族，并应用进化策略在该空间中搜索噪声鲁棒损失。为了学习可应用于新任务的可重用损失函数，我们的适应度函数在一系列训练数据集和体系结构中对其性能进行综合评分。由此产生的白盒损失提供了一个简单而快速的“即插即用”模块，该模块能够在各种下游任务中实现有效的标签噪声鲁棒性学习，而无需特殊的培训程序或网络体系结构。我们的损失的有效性在合成和真实标签噪声的各种数据集上得到了证明，与之前的工作相比，我们的损失是有利的。

We propose a novel transformer-based styled handwritten text image generation approach, HWT, that strives to learn both style-content entanglement as well as global and local style patterns. The proposed HWT captures the long and short range relationships within the style examples through a self-attention mechanism, thereby encoding both global and local style patterns. Further, the proposed transformer-based HWT comprises an encoder-decoder attention that enables style-content entanglement by gathering the style features of each query character. To the best of our knowledge, we are the first to introduce a transformer-based network for styled handwritten text generation. Our proposed HWT generates realistic styled handwritten text images and outperforms the state-of-the-art demonstrated through extensive qualitative, quantitative and human-based evaluations. The proposed HWT can handle arbitrary length of text and any desired writing style in a few-shot setting. Further, our HWT generalizes well to the challenging scenario where both words and writing style are unseen during training, generating realistic styled handwritten text images.

我们提出了一种新的基于转换器的手写文本图像生成方法HWT，该方法致力于学习样式内容纠缠以及全局和局部样式模式。建议的HWT通过自我注意机制捕获样式示例中的长程和短程关系，从而编码全局和局部样式模式。此外，所提出的基于转换器的HWT包括编码器-解码器，其通过收集每个查询字符的样式特征来实现样式内容纠缠。据我们所知，我们是第一个引入基于转换器的网络来生成样式化手写文本的人。我们提出的HWT生成逼真风格的手写文本图像，并通过广泛的定性、定量和基于人的评估，优于最先进的技术。建议的HWT可以处理任意长度的文本和任意所需的书写风格，只需几个镜头设置。此外，我们的HWT很好地概括了在训练过程中看不到单词和书写风格的挑战场景，生成了逼真的手写文本图像。

In this paper, we present a novel Dynamic DETR (Detection with Transformers) approach by introducing dynamic attentions into both the encoder and decoder stages of DETR to break its two limitations on small feature resolution and slow training convergence. To address the first limitation, which is due to the quadratic computational complexity of the self-attention module in Transformer encoders, we propose a dynamic encoder to approximate the Transformer encoder's attention mechanism using a convolution-based dynamic encoder with various attention types. Such an encoder can dynamically adjust attentions based on multiple factors such as scale importance, spatial importance, and representation (i.e., feature dimension) importance. To mitigate the second limitation of learning difficulty, we introduce a dynamic decoder by replacing the cross-attention module with a ROI-based dynamic attention in the Transformer decoder. Such a decoder effectively assists Transformers to focus on region of interests from a coarse-to-fine manner and dramatically lowers the learning difficulty, leading to a much faster convergence with fewer training epochs. We conduct a series of experiments to demonstrate our advantages. Our Dynamic DETR significantly reduces the training epochs (by  $\backslash bf 14x$ ), yet results in a much better performance (by  $\backslash bf 3.6$  on mAP). Meanwhile, in the standard 1x setup with ResNet-50 backbone, we archive a new state-of-the-art performance that further proves the learning effectiveness of the proposed approach. Code will be released soon.

在本文中，我们提出了一种新的动态DETR（变压器检测）方法，通过在DETR的编码器和解码器阶段引入动态注意来打破其特征分辨率小和训练收敛慢的两个限制。为了解决第一个限制，即变压器编码器中自注意模块的二次计算复杂性，我们提出了一种动态编码器，使用基于卷积的具有各种注意类型的动态编码器来近似变压器编码器的注意机制。这种编码器可以基于多个因素动态调整注意事项，例如尺度重要性、空间重要性和表示（即特征维度）重要性。为了缓解学习困难的第二个限制，我们在Transformer解码器中引入了一个动态解码器，将交叉注意模块替换为基于ROI的动态注意。这样的解码器有效地帮助变形金刚从粗到细地关注感兴趣的区域，并显著降低学习难度，从而以更少的训练周期实现更快的收敛。我们进行了一系列实验来证明我们的优势。我们的动态DETR显著缩短了训练时间（缩短

了 $\text{bf } 14x$ ），但却带来了更好的性能（缩短了 $\text{bf } 3.6$ ）。同时，在带有ResNet-50主干网的标准 $1x$ 设置中，我们存档了新的最先进性能，进一步证明了所提出方法的学习效率。代码将很快发布。

We propose an end-to-end pipeline, named watch once only (woo), for video action detection. Current methods either decouple video action detection task into separated stages of actor localization and action classification or train two separated models within one stage. In contrast, our approach solves the actor localization and action classification simultaneously in a unified network. The whole pipeline is significantly simplified by unifying the backbone network and eliminating many hand-crafted components. WOO takes a unified video backbone to simultaneously extract features for actor location and action classification. In addition, we introduce spatial-temporal action embeddings into our framework and design a spatial-temporal fusion module to obtain more discriminative features with richer information, which further boosts the action classification performance. Extensive experiments on AVA and JHMDB datasets show that WOO achieves state-of-the-art performance, while still reduces up to 16.7% GFLOPs compared with existing methods. We hope our work can inspire rethinking the convention of action detection and serve as a solid baseline for end-to-end action detection. Code is available.

我们提出了一种端到端的视频动作检测管道，名为watchonce Only (WOO)。现有的方法要么将视频动作检测任务分解为演员定位和动作分类的分离阶段，要么在一个阶段内训练两个分离的模型。相比之下，我们的方法在一个统一的网络中同时解决了角色定位和动作分类问题。通过统一主干网并消除许多手工制作的组件，整个管道大大简化。WOO采用统一的视频主干，同时提取演员位置和动作分类的特征。此外，我们将时空动作嵌入到我们的框架中，并设计了时空融合模块，以获得更具辨别力的特征和更丰富的信息，这进一步提高了动作分类的性能。在AVA和JHMDB数据集上进行的大量实验表明，与现有方法相比，WOO实现了最先进的性能，同时仍减少了高达16.7%的GFLOPs。我们希望我们的工作能够启发人们重新思考动作检测的惯例，并作为端到端动作检测的坚实基线。代码是可用的。

Trajectory forecasting is a crucial step for autonomous vehicles and mobile robots in order to navigate and interact safely. In order to handle the spatial interactions between objects, graph-based approaches have been proposed. These methods, however, model motion on a frame-to-frame basis and do not provide a strong temporal model. To overcome this limitation, we propose a compact model called Spatial-Temporal Consistency Network (STC-Net). In STC-Net, dilated temporal convolutions are introduced to model long-range dependencies along each trajectory for better temporal modeling while graph convolutions are employed to model the spatial interaction among different trajectories. Furthermore, we propose a feature-wise convolution to generate the predicted trajectories in one pass and refine the forecast trajectories together with the reconstructed observed trajectories. We demonstrate that STC-Net generates spatially and temporally consistent trajectories and outperforms other graph-based methods. Since STC-Net requires only 0.7k parameters and forecasts the future with a latency of only 1.3ms, it advances the state-of-the-art and satisfies the requirements for realistic applications.

轨迹预测是自主车辆和移动机器人安全导航和交互的关键步骤。为了处理对象之间的空间交互，提出了基于图的方法。然而，这些方法在帧到帧的基础上对运动进行建模，并且不提供强的时间模型。为了克服这一局限性，我们提出了一种称为时空一致性网络 (STC-Net) 的紧凑模型。在STC网络中，为了更好地进行时间建模，引入了扩展的时间卷积来建模每条轨迹上的长期依赖关系，而图卷积用于建模不同轨迹之间的空间交互。此外，我们还提出了一种基于特征的卷积方法，一次生成预测轨迹，并将预测轨迹与重构的观测轨迹一起细化。我们证明了STC网络能够生成空间和时间一致的轨迹，并且优于其他基于图的方法。由于STC网络只需要0.7k参数，预测未来的延迟仅为1.3ms，因此它提高了最先进的技术水平，满足了实际应用的要求。

Despite the extensive usage of point clouds in 3D vision, relatively limited data are available for training deep neural networks. Although data augmentation is a standard approach to compensate for the scarcity of data, it has been less explored in the point cloud literature. In this paper, we propose a simple and effective augmentation method called PointWOLF for point cloud augmentation. The proposed method produces smoothly varying non-rigid deformations by locally weighted transformations centered at multiple anchor points. The smooth deformations allow diverse and realistic augmentations. Furthermore, in order to minimize the manual efforts to search the optimal hyperparameters for augmentation, we present AugTune, which generates augmented samples of desired difficulties producing targeted confidence scores. Our experiments show that our framework consistently improves the performance for both shape classification and part segmentation tasks. In particular, with PointNet++, PointWOLF achieves the state-of-the-art 89.7 accuracy on shape classification with the real-world ScanObjectNN dataset. The code is available at <https://github.com/mlvlab/PointWOLF>.

尽管在3D视觉中广泛使用点云，但用于训练深层神经网络的数据相对有限。尽管数据扩充是弥补数据稀缺性的标准方法，但在点云文献中对其进行的探讨较少。本文提出了一种简单有效的点云增强方法PointWOLF。该方法通过以多个锚定点为中心的局部加权变换产生平滑变化的非刚性变形。平滑变形允许进行各种逼真的增强。此外，为了尽可能减少人工搜索最优超参数以进行增强，我们提出了AugTune，它生成期望困难的增强样本以产生目标置信度分数。我们的实验表明，我们的框架一致地提高了形状分类和零件分割任务的性能。特别是，通过PointNet++，PointWOLF在使用真实世界的ScanObjectNN数据集进行形状分类时达到了最先进的89.7精度。该守则可于<https://github.com/mlvlab/PointWOLF>。

Image classification models can depend on multiple different semantic attributes of the image. An explanation of the decision of the classifier needs to both discover and visualize these properties. Here we present StylEx, a method for doing this, by training a generative model to specifically explain multiple attributes that underlie classifier decisions. A natural source for such attributes is the StyleSpace of StyleGAN, which is known to generate semantically meaningful dimensions in the image. However, because standard GAN training is not dependent on the classifier, it may not represent those attributes which are important for the classifier decision, and the dimensions of StyleSpace may represent irrelevant attributes. To overcome this, we propose a training procedure for a StyleGAN, which incorporates the classifier model, in order to learn a classifier-specific StyleSpace. Explanatory attributes are then selected from this space. These can be used to visualize the effect of changing multiple attributes per image, thus providing image-specific explanations. We apply StylEx to multiple domains, including animals, leaves, faces and retinal images. For these, we show how an image can be modified in different ways to change its classifier output. Our results show that the method finds attributes that align well with semantic ones, generate meaningful image-specific explanations, and are human-interpretable as measured in user-studies.

图像分类模型可以依赖于图像的多个不同语义属性。对分类器决策的解释需要发现并可视化这些属性。在这里，我们介绍了StylEx，一种实现这一点的方法，它通过训练一个生成模型来具体解释分类器决策所依据的多个属性。此类属性的一个自然来源是StyleGAN的StyleSpace，已知它在图像中生成语义上有意义的维度。然而，由于标准GAN训练不依赖于分类器，因此它可能不表示对分类器决策重要的属性，并且样式空间的维度可能表示不相关的属性。为了克服这一问题，我们提出了一个样式表的训练过程，其中包含了分类器模型，以便学习特定于分类器的样式空间。然后从该空间中选择解释性属性。这些可用于可视化更改每个图像的多个属性的效果，从而提供特定于图像的解释。我们将StylEx应用于多个领域，包括动物、树叶、人脸和视网膜图像。对于这些，我们展示了如何以不同的方式修改图像以改变其

分类器输出。我们的结果表明，该方法发现的属性与语义属性很好地一致，生成有意义的图像特定解释，并且在用户研究中可以被人类理解。

Visual engagement in social media platforms comprises interactions with photo posts including comments, shares, and likes. In this paper, we leverage such visual engagement clues as supervisory signals for representation learning. However, learning from engagement signals is non-trivial as it is not clear how to bridge the gap between low-level visual information and high-level social interaction. We present VisE, a weakly supervised learning approach, which maps social images to pseudo labels derived by clustered engagement signals. We then study how models trained in this way benefit subjective downstream computer vision tasks such as emotion recognition or political bias detection. Through extensive studies, we empirically demonstrate the effectiveness of VisE across a diverse set of classification tasks beyond the scope of conventional recognition.

社交媒体平台的视觉参与包括与照片帖子的互动，包括评论、分享和喜欢。在本文中，我们利用视觉参与线索作为监督信号进行表征学习。然而，从参与信号中学习并非易事，因为不清楚如何弥合低层次视觉信息和高层次社会互动之间的差距。我们提出了一种弱监督学习方法VisE，它将社会图像映射到由聚集的参与信号导出的伪标签。然后，我们研究以这种方式训练的模型如何有利于主观的下游计算机视觉任务，如情感识别或政治偏见检测。通过广泛的研究，我们实验证明了VisE在传统识别范围之外的一系列分类任务中的有效性。

Image quality assessment (IQA) is an important research topic for understanding and improving visual experience. The current state-of-the-art IQA methods are based on convolutional neural networks (CNNs). The performance of CNN-based models is often compromised by the fixed shape constraint in batch training. To accommodate this, the input images are usually resized and cropped to a fixed shape, causing image quality degradation. To address this, we design a multi-scale image quality Transformer (MUSIQ) to process native resolution images with varying sizes and aspect ratios. With a multi-scale image representation, our proposed method can capture image quality at different granularities. Furthermore, a novel hash-based 2D spatial embedding and a scale embedding is proposed to support the positional embedding in the multi-scale representation. Experimental results verify that our method can achieve state-of-the-art performance on multiple large scale IQA datasets such as PaQ-2-PiQ, SPAQ and KonIQ-10k.

图像质量评估 (IQA) 是理解和改善视觉体验的重要研究课题。目前最先进的IQA方法基于卷积神经网络 (CNN)。在批量训练中，基于CNN的模型的性能经常受到固定形状约束的影响。为了适应这种情况，输入图像通常会调整大小并裁剪为固定形状，从而导致图像质量下降。为了解决这个问题，我们设计了一个多尺度图像质量转换器 (MUSIQ) 来处理具有不同大小和纵横比的本机分辨率图像。通过多尺度图像表示，我们提出的方法可以捕获不同粒度的图像质量。此外，本文还提出了一种新的基于哈希的二维空间嵌入和尺度嵌入方法来支持多尺度表示中的位置嵌入。实验结果证明，我们的方法可以在多个大规模IQA数据集（如PaQ-2-PiQ、SPAQ和KonIQ-10k）上实现最先进的性能。

Attention mechanism, especially channel attention, has gained great success in the computer vision field. Many works focus on how to design efficient channel attention mechanisms while ignoring a fundamental problem, i.e., channel attention mechanism uses scalar to represent channel, which is difficult due to massive information loss. In this work, we start from a different view and regard the channel representation problem as a compression process using frequency analysis. Based on the frequency analysis, we mathematically prove that the conventional global average pooling is a special case of the feature decomposition in the frequency domain. With the proof, we naturally generalize the compression of the channel attention mechanism in the frequency domain and propose our method with multi-spectral channel attention, termed as FcaNet. FcaNet is simple but effective. We can change a few lines of code in the calculation to implement our method within existing channel attention methods. Moreover, the proposed method achieves state-of-the-art results compared with other channel attention methods on image classification, object detection, and instance segmentation tasks. Our method could consistently outperform the baseline SENet, with the same number of parameters and the same computational cost. Our code and models are publicly available at <https://github.com/cfzd/FcaNet>.

注意机制，特别是通道注意，在计算机视觉领域取得了巨大的成功。许多研究集中在如何设计有效的通道注意机制上，而忽略了一个基本问题，即通道注意机制使用标量来表示通道，这是由于大量信息丢失而造成的困难。在这项工作中，我们从不同的角度出发，将信道表示问题视为使用频率分析的压缩过程。基于频率分析，我们从数学上证明了传统的全局平均池是频域特征分解的一种特例。通过证明，我们自然地推广了频域中的信道注意机制的压缩，并提出了我们的多光谱信道注意方法，称为FcaNet。FcaNet简单但有效。我们可以在计算中更改几行代码，以便在现有的通道注意方法中实现我们的方法。此外，与其他通道注意方法相比，该方法在图像分类、目标检测和实例分割任务方面取得了最新的结果。在相同的参数数量和相同的计算成本下，我们的方法始终优于基线SENet。我们的代码和模型在<https://github.com/cfzd/FcaNet>。

This paper considers matching images of low-light scenes, aiming to widen the frontier of SfM and visual SLAM applications. Recent image sensors can record the brightness of scenes with more than eight-bit precision, available in their RAW-format image. We are interested in making full use of such high-precision information to match extremely low-light scene images that conventional methods cannot handle. For extreme low-light scenes, even if some of their brightness information exists in the RAW format images' low bits, the standard raw image processing fails to utilize them properly. As was recently shown by Chen et al., CNNs can learn to produce images with a natural appearance from such RAW-format images. To consider if and how well we can utilize such information stored in RAW-format images for image matching, we have created a new dataset named MID (matching in the dark). Using it, we experimentally evaluated combinations of eight image-enhancing methods and eleven image matching methods consisting of classical/neural local descriptors and classical/neural initial point-matching methods. The results show the advantage of using the RAW-format images and the strengths and weaknesses of the above component methods. They also imply there is room for further research.

本文考虑弱光场景的匹配图像，旨在拓宽SfM和视觉SLAM应用的前沿。最新的图像传感器可以以超过8位的精度记录场景的亮度，并以原始图像格式提供。我们感兴趣的是充分利用这些高精度信息来匹配传统方法无法处理的极弱光场景图像。对于极弱光场景，即使某些亮度信息存在于原始格式图像的低位中，标准原始图像处理也无法正确利用它们。正如Chen等人最近所展示的，CNN可以学习从这些原始格式的图像生成具有自然外观的图像。为了考虑是否以及如何利用这种存储在原始格式图像中的信息进行图像匹配，我们已经创建了一个名为MID的新数据集（在黑暗中匹配）。使用它，我们实验评估了八种

图像增强方法和十一种图像匹配方法的组合，包括经典/神经局部描述符和经典/神经初始点匹配方法。结果显示了使用原始格式图像的优势以及上述组件方法的优缺点。他们还暗示有进一步研究的空间。

Vision-Dialog Navigation (VDN) requires an agent to ask questions and navigate following the human responses to find target objects. Conventional approaches are only allowed to ask questions at predefined locations, which are built upon expensive dialogue annotations, and inconvenience the real-word human-robot communication and cooperation. In this paper, we propose a Self-Motivated Communication Agent (SCoA) that learns whether and what to communicate with human adaptively to acquire instructive information for realizing dialogue annotation-free navigation and enhancing the transferability in real-world unseen environment. Specifically, we introduce a whether-to-ask (WeTA) policy, together with uncertainty of which action to choose, to indicate whether the agent should ask a question. Then, a what-to-ask (WaTA) policy is proposed, in which, along with the oracle's answers, the agent learns to score question candidates so as to pick up the most informative one for navigation, and meanwhile mimic oracle's answering. Thus, the agent can navigate in a self-Q&A manner even in real-world environment where the human assistance is often unavailable. Through joint optimization of communication and navigation in a unified imitation learning and reinforcement learning framework, SCoA asks a question if necessary and obtains a hint for guiding the agent to move towards the target with less communication cost. Experiments on seen and unseen environments demonstrate that SCoA shows not only superior performance over existing baselines without dialog annotations, but also competing results compared with rich dialog annotations based counterparts.

视觉对话导航 (VDN) 要求代理提出问题，并按照人类的反应进行导航，以找到目标对象。传统的方法只允许在预定义的位置提问，这是建立在昂贵的对话注释基础上的，给真正的人机交流与合作带来不便。在本文中，我们提出了一种自我激励的通信代理 (SCoA)，它自适应地学习是否与人通信以及与人通信的内容，以获取有指导意义的信息，从而实现无对话注释的导航，增强现实世界中不可见环境中的可转移性。具体而言，我们引入了“是否提问” (WeTA) 政策，以及选择哪种行动的不确定性，以表明代理人是否应该提问。然后，提出了一种询问什么 (WaTA) 策略，在该策略中，代理与oracle的答案一起学习对候选问题打分，以便选择信息量最大的问题进行导航，同时模仿oracle的回答。因此，即使在人的帮助通常不可用的真实环境中，代理也可以以自我问答的方式进行导航。通过在统一的模仿学习和强化学习框架中对通信和导航进行联合优化，SCoA在必要时提出问题，并获得提示，指导agent以较少的通信成本向目标移动。在可见和不可见环境上的实验表明，SCoA不仅比没有对话框注释的现有基线表现出更高的性能，而且与基于丰富对话框注释的同行相比，还显示出竞争性的结果。

How to make the appearance and motion information interact effectively to accommodate complex scenarios is a fundamental issue in flow-based zero-shot video object segmentation. In this paper, we propose an Attentive Multi-Modality Collaboration Network (AMC-Net) to utilize appearance and motion information uniformly. Specifically, AMC-Net fuses robust information from multi-modality features and promotes their collaboration in two stages. First, we propose a Multi-Modality Co-Attention Gate (MCG) on the bilateral encoder branches, in which a gate function is used to formulate co-attention scores for balancing the contributions of multi-modality features and suppressing the redundant and misleading information. Then, we propose a Motion Correction Module (MCM) with a visual-motion attention mechanism, which is constructed to emphasize the features of foreground objects by incorporating the spatio-temporal correspondence between appearance and motion cues. Extensive experiments on three public challenging benchmark datasets verify that our proposed network performs favorably against existing state-of-the-art methods via training with fewer data.

在基于流的零镜头视频对象分割中，如何使外观和运动信息有效地交互以适应复杂场景是一个基本问题。在本文中，我们提出了一种注意力集中的多模态协作网络（AMC-Net），以统一利用外观和运动信息。具体而言，AMC网络融合了来自多模态功能的可靠信息，并分两个阶段促进它们的协作。首先，我们在双边编码器分支上提出了一种多模态共同注意门（MCG），该门函数用于制定共同注意分数，以平衡多模态特征的贡献并抑制冗余和误导信息。然后，我们提出了一个带有视觉运动注意机制的运动校正模块（MCM），该模块通过结合外观和运动线索之间的时空对应关系来强调前景对象的特征。在三个公开的具有挑战性的基准数据集上进行的大量实验证明了我们提出的网络通过使用较少的数据进行训练，相对于现有的最先进的方法具有良好的性能。

We present DietNeRF, a 3D neural scene representation estimated from a few images. Neural Radiance Fields (NeRF) learn a continuous volumetric representation of a scene through multi-view consistency, and can be rendered from novel viewpoints by ray casting. While NeRF has an impressive ability to reconstruct geometry and fine details given many images, up to 100 for challenging 360 degree scenes, it often finds a degenerate solution to its image reconstruction objective when only a few input views are available. To improve few-shot quality, we propose DietNeRF. We introduce an auxiliary semantic consistency loss that encourages realistic renderings at novel poses. DietNeRF is trained on individual scenes to (1) correctly render given input views from the same pose, and (2) match high-level semantic attributes across different, random poses. Our semantic loss allows us to supervise DietNeRF from arbitrary poses. We extract these semantics using a pre-trained visual encoder such as CLIP, a Vision Transformer trained on hundreds of millions of diverse single-view, 2D photographs mined from the web with natural language supervision. In experiments, DietNeRF improves the perceptual quality of few-shot view synthesis when learned from scratch, can render novel views with as few as one observed image when pre-trained on a multi-view dataset, and produces plausible completions of completely unobserved regions. Our project website is available at <https://www.ajayj.com/dietnerf>.

我们介绍了DietNeRF，一种从一些图像估计的3D神经场景表示。神经辐射场（NeRF）通过多视图一致性学习场景的连续体积表示，并且可以通过光线投射从新的视点进行渲染。尽管NeRF在重建几何体和精细细节方面有着令人印象深刻的能力，许多图像（挑战360度场景时高达100），但当只有少数输入视图可用时，它通常会找到退化的图像重建目标解决方案。为了提高少数镜头的质量，我们建议DietNeRF。我们引入了一个辅助的语义一致性损失，它鼓励以新颖的姿势进行真实的渲染。DietNeRF在各个场景上进行训练，以（1）正确渲染来自同一姿势的给定输入视图，以及（2）匹配不同随机姿势的高级语义属性。我们的语义缺失使我们能够监督DietNeRF的任意姿势。我们使用预先训练过的视觉编码器（如CLIP）提取这些语义，CLIP是一种视觉转换器，通过自然语言监控从网络上挖掘出数亿张不同的单视图、二维照片。在实验中，DietNeRF改进了从零开始学习的少镜头视图合成的感知质量，在多视图数据集上进行预训练时，可以用一张观察到的图像渲染新视图，并生成完全未观察到的区域的合理完整。我们的项目网站位于<https://www.ajayj.com/dietnerf>。

Object detection aims to accurately locate and classify objects in an image, which requires precise object representations. Existing methods usually use rectangular anchor boxes or a set of points to represent objects. However, these methods either introduce background noise or miss the continuous appearance information inside the object, and thus cause incorrect detection results. In this paper, we propose a novel anchor-free object detection network, called CrossDet, which uses a set of growing cross lines along horizontal and vertical axes as object representations. An object can be flexibly represented as cross lines in different combinations. It not only can effectively reduce the interference of noise, but also takes into account the continuous object information, which is useful to enhance the discriminability of object features and find the object boundaries. Based on the learned cross lines, we propose a crossline extraction module to adaptively capture features of cross lines. Furthermore, we design a decoupled regression mechanism to regress the localization along the horizontal and vertical directions respectively, which helps to decrease the optimization difficulty because the optimization space is limited to a specific direction. Our method achieves consistently improvement on the PASCAL VOC and MS-COCO datasets. The experiment results demonstrate the effectiveness of our proposed method.

目标检测的目的是准确定位和分类图像中的目标，这需要精确的对象表示。现有方法通常使用矩形定位框或一组点来表示对象。然而，这些方法要么引入背景噪声，要么漏掉目标内部的连续外观信息，从而导致不正确的检测结果。在本文中，我们提出了一种新的无锚目标检测网络，称为CrossDet，它使用一组沿水平和垂直轴生长的交叉线作为目标表示。对象可以灵活地表示为不同组合中的交叉线。该方法不仅能有效地降低噪声的干扰，而且考虑了连续的目标信息，有利于增强目标特征的可分辨性和发现目标边界。基于学习到的交叉线，我们提出了一个交叉线提取模块来自适应地捕获交叉线的特征。此外，我们还设计了一种解耦回归机制，分别沿水平方向和垂直方向对定位进行回归，这有助于降低优化难度，因为优化空间仅限于特定方向。我们的方法在PASCAL VOC和MS-COCO数据集上实现了一致的改进。实验结果证明了该方法的有效性。

High quality imaging usually requires bulky and expensive lenses to compensate geometric and chromatic aberrations. This poses high constraints on the optical hash or low cost applications. Although one can utilize algorithmic reconstruction to remove the artifacts of low-end lenses, the degeneration from optical aberrations is spatially varying and the computation has to trade off efficiency for performance. For example, we need to conduct patch-wise optimization or train a large set of local deep neural networks to achieve high reconstruction performance across the whole image. In this paper, we propose a PSF aware plug-and-play deep network, which takes the aberrant image and PSF map as input and produces the latent high quality version via incorporating lens-specific deep priors, thus leading to a universal and flexible optical aberration correction method. Specifically, we pre-train a base model from a set of diverse lenses and then adapt it to a given lens by quickly refining the parameters, which largely alleviates the time and memory consumption of model learning. The approach is of high efficiency in both training and testing stages. Extensive results verify the promising applications of our proposed approach for compact low-end cameras.

高质量成像通常需要体积庞大且价格昂贵的透镜来补偿几何和色差。这对光学散列或低成本应用程序造成了很大的限制。虽然可以利用算法重建来消除低端透镜的伪影，但光学像差的退化在空间上是变化的，计算必须在效率和性能之间进行权衡。例如，我们需要进行面片优化或训练大量局部深层神经网络，以在整个图像上实现高重建性能。在本文中，我们提出了一种支持PSF的即插即用深度网络，该网络以畸变图像和PSF图为输入，通过结合镜头特定的深度先验产生潜在的高质量版本，从而实现了一种通用灵活的光学像差校正方法。具体来说，我们从一组不同的镜头中预先训练一个基本模型，然后通过

快速细化参数使其适应给定镜头，这在很大程度上减少了模型学习的时间和内存消耗。该方法在训练和测试阶段都具有较高的效率。大量结果验证了我们提出的方法在紧凑型低端相机上的应用前景。

Recently, there has been an increasing number of efforts to introduce models capable of generating natural language explanations (NLEs) for their predictions on vision-language (VL) tasks. Such models are appealing, because they can provide human-friendly and comprehensive explanations. However, there is a lack of comparison between existing methods, which is due to a lack of re-usable evaluation frameworks and a scarcity of datasets. In this work, we introduce e-ViL and e-SNLI-VE. e-ViL is a benchmark for explainable vision-language tasks that establishes a unified evaluation framework and provides the first comprehensive comparison of existing approaches that generate NLEs for VL tasks. It spans four models and three datasets and both automatic metrics and human evaluation are used to assess model-generated explanations. e-SNLI-VE is currently the largest existing VL dataset with NLEs (over 430k instances). We also propose a new model that combines UNITER, which learns joint embeddings of images and text, and GPT-2, a pre-trained language model that is well-suited for text generation. It surpasses the previous state of the art by a large margin across all datasets. Code and data are available here:

<https://github.com/maximek3/e-ViL>.

最近，有越来越多的努力引入能够生成自然语言解释（NLE）的模型来预测视觉语言（VL）任务。这些模型很有吸引力，因为它们可以提供人性化和全面的解释。然而，由于缺乏可重用的评估框架和数据集，现有方法之间缺乏比较。在这项工作中，我们介绍了e-ViL和e-SNLI-VE。e-ViL是可解释视觉语言任务的基准，它建立了一个统一的评估框架，并首次全面比较了为VL任务生成NLE的现有方法。它跨越四个模型和三个数据集，自动度量和人工评估都用于评估模型生成的解释。e-SNLI-VE是目前最大的现有VL数据集，具有NLE（超过430k个实例）。我们还提出了一个新的模型，该模型结合了UNITER（学习图像和文本的联合嵌入）和GPT-2（一种非常适合文本生成的预训练语言模型）。它在所有数据集上都大大超过了以前的技术水平。代码和数据可在此处获得：<https://github.com/maximek3/e-ViL>.

Anomaly detection (AD) aims to address the task of classification or localization of image anomalies. This paper addresses two pivotal issues of reconstruction-based approaches to AD in images, namely, model adaptation and reconstruction gap. The former generalizes an AD model to tackling a broad range of object categories, while the latter provides useful clues for localizing abnormal regions. At the core of our method is an unsupervised universal model, termed as Metaformer, which leverages both meta-learned model parameters to achieve high model adaptation capability and instance-aware attention to emphasize the focal regions for localizing abnormal regions, i.e., to explore the reconstruction gap at those regions of interest. We justify the effectiveness of our method with SOTA results on the MVTec AD dataset of industrial images and highlight the adaptation flexibility of the universal Metaformer with multi-class and few-shot scenarios.

异常检测（AD）旨在解决图像异常的分类或定位任务。本文讨论了基于重建的图像AD方法的两个关键问题，即模型自适应和重建间隙。前者将AD模型推广到处理范围广泛的对象类别，而后者为定位异常区域提供了有用的线索。我们方法的核心是一个无监督的通用模型，称为Metaformer，它利用元学习模型参数来实现高模型适应能力和实例感知注意来强调用于定位异常区域的焦点区域，即探索这些感兴趣区域的重建差距。我们用工业图像MVTec AD数据集上的SOTA结果证明了我们方法的有效性，并强调了通用元模型在多类和少镜头场景下的适应性。

We present a domain- and user-preference-agnostic approach to detect highlightable excerpts from human-centric videos. Our method works on the graph-based representation of multiple observable human-centric modalities in the videos, such as poses and faces. We use an autoencoder network equipped with spatial-temporal graph convolutions to detect human activities and interactions based on these modalities. We train our network to map the activity- and interaction-based latent structural representations of the different modalities to per-frame highlight scores based on the representativeness of the frames. We use these scores to compute which frames to highlight and stitch contiguous frames to produce the excerpts. We train our network on the large-scale AVA-Kinetics action dataset and evaluate it on four benchmark video highlight datasets: DSH, TVSum, PHD<sup>2</sup>, and SumMe. We observe a 4-12% improvement in the mean average precision of matching the human-annotated highlights over state-of-the-art methods in these datasets, without requiring any user-provided preferences or dataset-specific fine-tuning.

我们提出了一种领域和用户偏好无关的方法来检测以人类为中心的视频中的可突出显示的摘录。我们的方法工作于视频中多个可观察到的以人为中心的模式的基于图形的表示，例如姿势和面部。我们使用配备时空图卷积的自动编码器网络来检测基于这些模式的人类活动和交互。我们训练我们的网络将不同模式的基于活动和交互的潜在结构表示映射到基于帧代表性的每帧突出显示分数。我们使用这些分数来计算要高亮显示的帧，并缝合相邻帧以生成摘录。我们在大规模AVA动力学动作数据集上训练我们的网络，并在四个基准视频突出显示数据集上对其进行评估：DSH、TVSum、PHD<sup>2</sup>和SumMe。我们观察到，在这些数据集中，与最先进的方法相比，匹配人类注释高光的平均精度提高了4-12%，而无需任何用户提供的偏好或数据集特定的微调。

We propose a versatile deep image compression network based on Spatial Feature Transform (SFT), which takes a source image and a corresponding quality map as inputs and produce a compressed image with variable rates. Our model covers a wide range of compression rates using a single model, which is controlled by arbitrary pixel-wise quality maps. In addition, the proposed framework allows us to perform task-aware image compressions for various tasks, e.g., classification, by efficiently estimating optimized quality maps specific to target tasks for our encoding network. This is even possible with a pretrained network without learning separate models for individual tasks. Our algorithm achieves outstanding rate-distortion trade-off compared to the approaches based on multiple models that are optimized separately for several different target rates. At the same level of compression, the proposed approach successfully improves performance on image classification and text region quality preservation via task-aware quality map estimation without additional model training. The code is available at the project website <https://github.com/micmic123/QmapCompression>.

我们提出了一种基于空间特征变换（SFT）的多功能深度图像压缩网络，该网络以源图像和相应的质量图作为输入，以可变速率生成压缩图像。我们的模型使用单个模型覆盖了广泛的压缩率，该模型由任意像素质量贴图控制。此外，所提出的框架允许我们通过有效地估计特定于编码网络目标任务的优化质量映射，对各种任务（例如分类）执行任务感知图像压缩。这甚至可以通过预先训练的网络实现，而无需学习单独任务的单独模型。与基于多个模型的方法相比，我们的算法实现了出色的率失真折衷，这些模型分别针对多个不同的目标速率进行了优化。在相同的压缩水平下，该方法通过任务感知质量图估计成功地提高了图像分类和文本区域质量保持的性能，而无需额外的模型训练。该代码可在项目网站上获得<https://github.com/micmic123/QmapCompression>。

We present a novel solution to the garment animation problem through deep learning. Our contribution allows animating any template outfit with arbitrary topology and geometric complexity. Recent works develop models for garment edition, resizing and animation at the same time by leveraging the support body model (encoding garments as body homotopies). This leads to complex engineering solutions that suffer from scalability, applicability and compatibility. By limiting our scope to garment animation only, we are able to propose a simple model that can animate any outfit, independently of its topology, vertex order or connectivity. Our proposed architecture maps outfits to animated 3D models into the standard format for 3D animation (blend weights and blend shapes matrices), automatically providing of compatibility with any graphics engine. We also propose a methodology to complement supervised learning with an unsupervised physically based learning that implicitly solves collisions and enhances cloth quality.

我们提出了一个新的解决方案，服装动画问题，通过深入学习。我们的贡献允许为任何具有任意拓扑和几何复杂性的模板装备制作动画。最近的工作通过利用支持身体模型（将衣服编码为身体同伦）同时开发服装编辑、尺寸调整和动画模型。这导致复杂的工程解决方案受到可伸缩性、适用性和兼容性的影响。通过将我们的范围仅限于服装动画，我们能够提出一个简单的模型，该模型可以为任何服装制作动画，而不依赖于其拓扑、顶点顺序或连接性。我们提出的体系结构将装备映射到三维动画的标准格式（混合权重和混合形状矩阵），自动提供与任何图形引擎的兼容性。我们还提出了一种方法来补充监督学习与非监督的基于物理的学习，隐式解决冲突，提高布料质量。

Coarse-to-fine strategies have been extensively used for the architecture design of single image deblurring networks. Conventional methods typically stack sub-networks with multi-scale input images and gradually improve sharpness of images from the bottom sub-network to the top sub-network, yielding inevitably high computational costs. Toward a fast and accurate deblurring network design, we revisit the coarse-to-fine strategy and present a multi-input multi-output U-net (MIMO-UNet). The MIMO-UNet has three distinct features. First, the single encoder of the MIMO-UNet takes multi-scale input images to ease the difficulty of training. Second, the single decoder of the MIMO-UNet outputs multiple deblurred images with different scales to mimic multi-cascaded U-nets using a single U-shaped network. Last, asymmetric feature fusion is introduced to merge multi-scale features in an efficient manner. Extensive experiments on the GoPro and RealBlur datasets demonstrate that the proposed network outperforms the state-of-the-art methods in terms of both accuracy and computational complexity. Source code is available for research purposes at <https://github.com/chosj95/MIMO-UNet>.

从粗到精的策略已被广泛用于单图像去模糊网络的结构设计。传统方法通常使用多尺度输入图像堆叠子网络，并逐渐提高从底部子网络到顶部子网络的图像清晰度，不可避免地产生较高的计算成本。为了快速准确的去模糊网络设计，我们重新研究了从粗到精的策略，并提出了一种多输入多输出U网络（MIMO UNet）。MIMO-UNet有三个不同的特性。首先，MIMO-UNet的单个编码器获取多尺度输入图像以减轻训练的难度。其次，MIMO-UNet的单个解码器输出具有不同尺度的多个去模糊图像，以使用单个U形网络模拟多个级联U形网络。最后，引入非对称特征融合，有效地融合多尺度特征。在GoPro和RealBlur数据集上的大量实验表明，所提出的网络在精度和计算复杂度方面都优于最先进的方法。源代码可用于研究目的，网址为<https://github.com/chosj95/MIMO-UNet>。

A key assumption of top-down human pose estimation approaches is their expectation of having a single person/instance present in the input bounding box. This often leads to failures in crowded scenes with occlusions. We propose a novel solution to overcome the limitations of this fundamental assumption. Our Multi-Instance Pose Network (MIPNet) allows for predicting multiple 2D pose instances within a given bounding box. We introduce a Multi-Instance Modulation Block (MIMB) that can adaptively modulate channel-wise feature responses for each instance and is parameter efficient. We demonstrate the efficacy of our approach by evaluating on COCO, CrowdPose, and OCHuman datasets. Specifically, we achieve 70.0 AP on CrowdPose and 42.5 AP on OCHuman test sets, a significant improvement of 2.4 AP and 6.5 AP over the prior art, respectively. When using ground truth bounding boxes for inference, MIPNet achieves an improvement of 0.7 AP on COCO, 0.9 AP on CrowdPose, and 9.1 AP on OCHuman validation sets compared to HRNet. Interestingly, when fewer, high confidence bounding boxes are used, HRNet's performance degrades (by 5 AP) on OCHuman, whereas MIPNet maintains a relatively stable performance (drop of 1 AP) for the same inputs.

自上而下的人体姿势估计方法的一个关键假设是，它们期望在输入边界框中有一个人/实例。这通常会导致在有遮挡的拥挤场景中失败。我们提出了一种新的解决方案来克服这一基本假设的局限性。我们的多实例姿势网络 (MIPNet) 允许预测给定边界框内的多个2D姿势实例。我们介绍了一种多实例调制块 (MIMB)，它可以自适应地调制每个实例的通道特性响应，并且参数效率高。通过对COCO、 CrowdPose和Ohuman数据集的评估，我们证明了我们方法的有效性。具体而言，我们在CrowdPose上实现了70.0 AP，在Ohuman测试集上实现了42.5 AP，分别比现有技术显著提高了2.4 AP和6.5 AP。当使用地面真值边界框进行推理时，与HRNet相比，MIPNet在COCO上实现了0.7 AP，在CrowdPose上实现了0.9 AP，在Ohuman验证集上实现了9.1 AP。有趣的是，当使用较少的高置信度边界框时， HRNet在Ohuman上的性能会下降（下降5 AP），而MIPNet在相同的输入下保持相对稳定的性能（下降1 AP）。

Satellite multi-view stereo (MVS) imagery is particularly suited for large-scale Earth surface reconstruction. Differing from the perspective camera model (pin-hole model) that is commonly used for close-range and aerial cameras, the cubic rational polynomial camera (RPC) model is the mainstream model for push-broom linear-array satellite cameras. However, the homography warping used in the prevailing learning based MVS methods is only applicable to pin-hole cameras. In order to apply the SOTA learning based MVS technology to the satellite MVS task for large-scale Earth surface reconstruction, RPC warping should be considered. In this work, we propose, for the first time, a rigorous RPC warping module. The rational polynomial coefficients are recorded as a tensor, and the RPC warping is formulated as a series of tensor transformations. Based on the RPC warping, we propose the deep learning based satellite MVS (SatMVS) framework for large-scale and wide depth range Earth surface reconstruction. We also introduce a large-scale satellite image dataset consisting of 519 5120x5120 images, which we call the TLC SatMVS dataset. The satellite images were acquired from a three-line camera (TLC) that catches triple-view images simultaneously, forming a valuable supplement to the existing open-source WorldView-3 datasets with single-scanline images. Experiments show that the proposed RPC warping module and the SatMVS framework can achieve a superior reconstruction accuracy compared to the pin-hole fitting method and conventional MVS methods. Code and data are available at <https://github.com/WHU-GPCV/SatMVS>.

卫星多视图立体 (MVS) 图像特别适合于大规模地表重建。与近景和航空相机常用的透视相机模型（针孔模型）不同，三次有理多项式相机 (RPC) 模型是推扫式线阵卫星相机的主流模型。然而，主流的基于学习的MVS方法中使用的单应变形仅适用于针孔相机。为了将基于SOTA学习的MVS技术应用于卫星 MVS任务中进行大规模地表重建，应考虑RPC翘曲。在这项工作中，我们首次提出了一个严格的RPC扭曲模块。有理多项式系数记录为张量，RPC翘曲表示为一系列张量变换。基于RPC翘曲，我们提出了基

于深度学习的卫星MVS (SatMVS) 框架，用于大规模、宽深度范围的地表重建。我们还介绍了由519 5120x5120图像组成的大规模卫星图像数据集，我们称之为TLC SatMVS数据集。卫星图像是从三线摄像机 (TLC) 获取的，该摄像机可同时捕获三视图图像，对现有开源WorldView-3数据集的单扫描线图像形成了有价值的补充。实验表明，与针孔拟合法和传统的MVS方法相比，本文提出的RPC翘曲模块和SatMVS框架能够获得更高的重建精度。有关代码和数据，请访问<https://github.com/WHU-GPCV/SatMVS>。

Nowadays modern displays are capable to render video content with high dynamic range (HDR) and wide color gamut (WCG). However, most available resources are still in standard dynamic range (SDR). Therefore, there is an urgent demand to transform existing SDR-TV contents into their HDR-TV versions. In this paper, we conduct an analysis of SDRTV-to-HDRTV task by modeling the formation of SDRTV/HDRTV content. Base on the analysis, we propose a three-step solution pipeline including adaptive global color mapping, local enhancement and highlight generation. Moreover, the above analysis inspires us to present a lightweight network that utilizes global statistics as guidance to conduct image-adaptive color mapping. In addition, we construct a dataset using HDR videos in HDR10 standard, named HDRTV1K, and select five metrics to evaluate the results of SDRTV-to-HDRTV algorithms. Furthermore, our final results achieve state-of-the-art performance in quantitative comparisons and visual quality. The code and dataset are available at <https://github.com/chxy95/HDRTVNet>.

如今，现代显示器能够以高动态范围 (HDR) 和宽色域 (WCG) 呈现视频内容。然而，大多数可用资源仍在标准动态范围 (SDR) 内。因此，迫切需要将现有SDR-TV内容转换为HDR-TV版本。在本文中，我们通过建模SDRTV/HDRTV内容的形成来分析SDRTV到HDRTV任务。在此基础上，我们提出了一个包括自适应全局颜色映射、局部增强和高光生成三步的解决方案。此外，上述分析启发我们提出了一个轻量级网络，该网络利用全球统计数据作为指导来进行图像自适应颜色映射。此外，我们使用HDR10标准中的HDR视频构建了一个数据集，命名为HDRTV1K，并选择五个指标来评估SDRTV到HDRTV算法的结果。此外，我们的最终结果在定量比较和视觉质量方面达到了最先进的水平。代码和数据集可在<https://github.com/chxy95/HDRTVNet>。

Autonomous systems need to understand the semantics and geometry of their surroundings in order to comprehend and safely execute object-level task specifications. This paper proposes an expressive yet compact model for joint object pose and shape optimization, and an associated optimization algorithm to infer an object-level map from multi-view RGB-D camera observations. The model is expressive because it captures the identities, positions, orientations, and shapes of objects in the environment. It is compact because it relies on a low-dimensional latent representation of implicit object shape, allowing onboard storage of large multi-category object maps. Different from other works that rely on a single object representation format, our approach has a bi-level object model that captures both the coarse level scale as well as the fine level shape details. Our approach is evaluated on the large-scale real-world ScanNet dataset and compared against state-of-the-art methods.

自治系统需要理解其周围环境的语义和几何结构，以便理解和安全地执行对象级任务规范。本文提出了一种用于关节对象姿态和形状优化的表达性但紧凑的模型，以及一种从多视点RGB-D相机观测推断对象级地图的相关优化算法。该模型具有表现力，因为它捕获了环境中对象的身份、位置、方向和形状。它是紧凑的，因为它依赖于隐式对象形状的低维潜在表示，允许机载存储大型多类别对象贴图。与其他依赖于单个对象表示格式的工作不同，我们的方法有一个双层对象模型，该模型既捕获了粗略级别的比例，也捕获了精细级别的形状细节。我们的方法在大规模现实世界的ScanNet数据集上进行了评估，并与最先进的方法进行了比较。

Explainable artificial intelligence has been gaining attention in the past few years. However, most existing methods are based on gradients or intermediate features, which are not directly involved in the decision-making process of the classifier. In this paper, we propose a slot attention-based classifier called SCOUTER for transparent yet accurate classification. Two major differences from other attention-based methods include: (a) SCOUTER's explanation is involved in the final confidence for each category, offering more intuitive interpretation, and (b) all the categories have their corresponding positive or negative explanation, which tells "why the image is of a certain category" or "why the image is not of a certain category." We design a new loss tailored for SCOUTER that controls the model's behavior to switch between positive and negative explanations, as well as the size of explanatory regions. Experimental results show that SCOUTER can give better visual explanations in terms of various metrics while keeping good accuracy on small and medium-sized datasets.

可解释人工智能在过去的几年里得到了广泛的关注。然而，现有的大多数方法都是基于梯度或中间特征的，它们并不直接参与分类器的决策过程。在本文中，我们提出了一种基于时隙注意的分类器，称为SCOUTER，用于透明而准确的分类。与其他基于注意力的方法的两个主要区别包括：(a) SCOUTER的解释涉及到每个类别的最终置信度，提供更直观的解释；(b) 所有类别都有相应的正面或负面解释，说明“为什么图像属于某一类别”或“为什么图像不属于某一类别”我们为SCOUTER设计了一个新的损失控制模型的行为，以在积极和消极解释之间切换，以及解释区域的大小。实验结果表明，在中小型数据集上，SCOUTER能够提供更好的视觉解释，同时保持良好的准确性。

To date, various 3D scene understanding tasks still lack practical and generalizable pre-trained models, primarily due to the intricate nature of 3D scene understanding tasks and their immense variations due to camera views, lighting, occlusions, etc. In this paper, we tackle this imminent challenge by introducing a spatio-temporal representation learning (STRL) framework, capable of learning from unlabeled 3D point clouds in a self-supervised fashion. Inspired by how infants learn from visual data in-the-wild, we explore the rich spatio-temporal cues derived from the 3D data. Specifically, STRL takes two temporally-correlated frames from a 3D point cloud sequence as the input, transforms it with spatial data augmentation, and learns the invariant representation self-supervisedly. To corroborate the efficacy of STRL, we conduct extensive experiments on synthetic, indoor, and outdoor datasets. Experimental results demonstrate that, compared with supervised learning methods, the learned self-supervised representation facilitates various models to attain comparable or even better performances while capable of generalizing pre-trained models to downstream tasks, including 3D shape classification, 3D object detection, and 3D semantic segmentation. Moreover, spatio-temporal contextual cues embedded in 3D point clouds significantly improve the learned representations.

到目前为止，各种3D场景理解任务仍然缺乏实用且可概括的预训练模型，这主要是由于3D场景理解任务的复杂性及其因摄像机视图、照明、遮挡等而产生的沉浸变化，我们通过引入时空表示学习（STRL）框架来应对这一内在挑战，该框架能够以自我监督的方式从未标记的3D点云中学习。受婴儿如何在野外从视觉数据中学习的启发，我们探索了从3D数据中获得的丰富的时空线索。具体地说，STRL从一个3D点云序列中获取两个时间相关帧作为输入，通过空间数据扩充对其进行变换，并自我监督地学习不变表示。为了证实STRL的有效性，我们在合成、室内和室外数据集上进行了广泛的实验。实验结果表明，与有监督学习方法相比，学习的自监督表示方法有助于各种模型获得可比甚至更好的性能，同时能够将预先训练的模型推广到下游任务，包括三维形状分类、三维目标检测、，三维语义分割。此外，嵌入在三维点云中的时空上下文线索显著改善了学习的表示。

Vision-language Navigation (VLN) task requires an agent to perceive both the visual scene and natural language and navigate step-by-step. Large data bias makes the VLN task challenging, which is caused by the disparity ratio between small data scale and large navigation space. Previous works have proposed many data augmentation methods to reduce data bias. However, these works do not explicitly reduce the data bias across different house scenes. Therefore, the agent would be overfitting to the seen scenes and perform navigation poorly in the unseen scenes. To tackle this problem, we propose the random environmental mixup (REM) method, which generates augmentation data in cross-connected house scenes. This method consists of three steps: 1) we select the key viewpoints according to the room connection graph for each scene in the training split; 2) we cross-connect the key views of different scenes to construct augmented scenes; 3) we generate augmentation data triplets (environment, path, instruction) in the cross-connected scenes. Our experiments prove that the augmentation data helps the agent reduce its performance gap between the seen and unseen environment and improve its performance, making our model be the best existing approach on the standard benchmark.

视觉语言导航 (VLN) 任务要求agent感知视觉场景和自然语言，并逐步导航。大数据偏差使得VLN任务具有挑战性，这是由小数据规模和大导航空间之间的视差比造成的。以前的工作已经提出了许多数据增强方法来减少数据偏差。然而，这些工作并没有明确减少不同房屋场景的数据偏差。因此，代理将过度适应已看到的场景，并且在未看到的场景中执行导航效果不佳。为了解决这个问题，我们提出了随机环境混合 (REM) 方法，该方法在交叉连接的房屋场景中生成增强数据。该方法包括三个步骤：1) 根据训练分割中每个场景的房间连接图选择关键视点；2) 我们交叉连接不同场景的关键视图来构建增强场景；3) 我们在交叉连接的场景中生成增强数据三元组（环境、路径、指令）。我们的实验证明，增强数据有助于agent缩小其在可见和不可见环境中的性能差距，并提高其性能，使我们的模型成为标准基准上现有的最佳方法。

Learning from image-text data has demonstrated recent success for many recognition tasks, yet is currently limited to visual features or individual visual concepts such as objects. In this paper, we propose one of the first methods that learn from image-sentence pairs to extract a graphical representation of localized objects and their relationships within an image, known as scene graph. To bridge the gap between images and texts, we leverage an off-the-shelf object detector to identify and localize object instances, match labels of detected regions to concepts parsed from captions, and thus create "pseudo" labels for learning scene graph. Further, we design a Transformer-based model to predict these "pseudo" labels via a masked token prediction task. Learning from only image-sentence pairs, our model achieves 30% relative gain over a latest method trained with human-annotated unlocalized scene graphs. Our model also shows strong results for weakly and fully supervised scene graph generation. In addition, we explore an open-vocabulary setting for detecting scene graphs, and present the first result for open-set scene graph generation.

从图像文本数据中学习已证明最近在许多识别任务中取得了成功，但目前仅限于视觉特征或单个视觉概念，如对象。在本文中，我们提出了第一种从图像句子对中学习以提取图像中局部对象及其关系的图形表示的方法，称为场景图。为了弥合图像和文本之间的鸿沟，我们利用现成的对象检测器来识别和定位对象实例，将检测到的区域的标签与从标题解析的概念相匹配，从而创建用于学习场景图的“伪”标签。此外，我们还设计了一个基于转换器的模型，通过屏蔽令牌预测任务来预测这些“伪”标签。我们的模型仅从图像-句子对中学习，与使用人类注释的非定域场景图训练的最新方法相比，获得了30%的相对增益。我们的模型还显示了弱和完全监督场景图生成的强大结果。此外，我们还探索了一种用于检测场景图的开放词汇表设置，并给出了开放集场景图生成的第一个结果。

A neural radiance field (NeRF) is a scene model supporting high-quality view synthesis, optimized per scene. In this paper, we explore enabling user editing of a category-level NeRF trained on a shape category. Specifically, we propose a method for propagating coarse 2D user scribbles to the 3D space, to modify the color or shape of a local region. First, we propose a conditional radiance field that incorporates new modular network components, including a branch that is shared across object instances in the category. Observing multiple instances of the same category, our model learns underlying part semantics without any supervision, thereby allowing the propagation of coarse 2D user scribbles to the entire 3D region (e.g., chair seat) in a consistent fashion. Next, we investigate for the editing tasks which components of our network require updating. We propose a hybrid network update strategy that targets the later network components, which balances efficiency and accuracy. During user interaction, we formulate an optimization problem that both satisfies the user's constraints and preserves the original object structure. We demonstrate our approach on a variety of editing tasks over three shape datasets and show that it outperforms prior neural editing approaches. Finally, we edit the appearance and shape of a real photograph and show that the edit propagates to extrapolated novel views.

神经辐射场 (NeRF) 是支持高质量视图合成的场景模型，针对每个场景进行优化。在本文中，我们将探讨如何启用用户编辑在形状类别上训练的类别级别NeRF。具体来说，我们提出了一种将粗糙的2D用户涂鸦传播到3D空间的方法，以修改局部区域的颜色或形状。首先，我们提出了一个条件辐射场，它包含新的模块化网络组件，包括一个在类别中的对象实例之间共享的分支。通过观察同一类别的多个实例，我们的模型在没有任何监督的情况下学习基础零件语义，从而允许以一致的方式将粗略的2D用户涂鸦传播到整个3D区域（例如，座椅）。接下来，我们调查编辑任务中需要更新的网络组件。我们提出了一种混合网络更新策略，该策略以后期网络组件为目标，平衡了效率和准确性。在用户交互过程中，我们提出了一个既满足用户约束又保持原始对象结构的优化问题。我们在三个形状数据集的各种编辑任务上演示了我们的方法，并表明它优于以前的神经编辑方法。最后，我们编辑了一张真实照片的外观和形状，并显示编辑传播到外推的新视图。

In this paper, we challenge the common assumption that collapsing the spatial dimensions of a 3D (spatial-channel) tensor in a convolutional neural network (CNN) into a vector via global pooling removes all spatial information. Specifically, we demonstrate that positional information is encoded based on the ordering of the channel dimensions, while semantic information is largely not. Following this demonstration, we show the real world impact of these findings by applying them to two applications. First, we propose a simple yet effective data augmentation strategy and loss function which improves the translation invariance of a CNN's output. Second, we propose a method to efficiently determine which channels in the latent representation are responsible for (i) encoding overall position information or (ii) region-specific positions. We first show that semantic segmentation has a significant reliance on the overall position channels to make predictions. We then show for the first time that it is possible to perform a 'region-specific' attack, and degrade a network's performance in a particular part of the input. We believe our findings and demonstrated applications will benefit research areas concerned with understanding the characteristics of CNNs.

在本文中，我们挑战了一个普遍的假设，即通过全局合并将卷积神经网络 (CNN) 中的三维 (空间通道) 张量的空间维度压缩为一个向量会删除所有空间信息。具体来说，我们证明了位置信息是基于通道维度的顺序编码的，而语义信息在很大程度上不是。在本演示之后，我们通过将这些发现应用于两个应用程序来展示它们对现实世界的影响。首先，我们提出了一种简单而有效的数据增强策略和损失函数，以提高CNN输出的平移不变性。其次，我们提出了一种方法来有效地确定潜在表示中的哪些通道负责 (i) 编码总体位置信息或 (ii) 区域特定位置。我们首先表明，语义分割在很大程度上依赖于整体位置

通道来进行预测。然后，我们第一次展示了有可能执行“特定于区域”的攻击，并在输入的特定部分降低网络的性能。我们相信，我们的发现和演示的应用将有助于了解CNN特征的相关研究领域。

We present in this paper a new architecture, named Convolutional vision Transformer (CvT), that improves Vision Transformer (ViT) in performance and efficiency by introducing convolutions into ViT to yield the best of both designs. This is accomplished through two primary modifications: a hierarchy of Transformers containing a new convolutional token embedding, and a convolutional Trasnsformer block leveraging a convolutional projection. These changes introduce desirable properties of convolutional neural networks (CNNs) to the ViT architecture (i.e. shift, scale, and distortion invariance) while maintaining the merits of Transformers (i.e. dynamic attention, global context, and better generalization). We validate CvT by conducting extensive experiments, showing that this approach achieves state-of-the-art performance over other Vision Transformers and ResNets on ImageNet-1k, with less parameters and lower FLOPs. In addition, performance gains are maintained when pretrained on larger datasets (e.g. ImageNet-22k) and fine-tuned to downstream tasks. Finally, our results show that the positional encoding, a crucial component in existing Vision Transformers, can be safely removed in our model, simplifying the design for higher resolution vision tasks. Code will be released at <https://github.com/microsoft/CvT>.

本文提出了一种新的结构，称为卷积视觉变换器（CvT），它通过在ViT中引入卷积来提高视觉变换器（ViT）的性能和效率。这是通过两个主要修改来实现的：一个包含新卷积令牌嵌入的变压器层次结构，以及一个利用卷积投影的卷积变换器块。这些变化将卷积神经网络（CNN）的理想特性引入ViT体系结构（即平移、缩放和失真不变性），同时保持变压器的优点（即动态注意、全局上下文和更好的泛化）。我们通过大量实验证明了CvT，结果表明，与ImageNet-1k上的其他视觉转换器和RESNET相比，该方法实现了最先进的性能，参数更少，触发器更少。此外，在对更大的数据集（如ImageNet-22k）进行预训练并对下游任务进行微调时，可以保持性能提升。最后，我们的结果表明，位置编码，一个在现有的视觉变压器的关键组成部分，可以在我们的模型中安全地删除，简化了高分辨率视觉任务的设计。守则将于<https://github.com/microsoft/CvT>。

Most weakly supervised semantic segmentation (WSSS) methods follow the pipeline that generates pseudo-masks initially and trains the segmentation model with the pseudo-masks in fully supervised manner after. However, we find some matters related to the pseudo-masks, including high quality pseudo-masks generation from class activation maps (CAMS), and training with noisy pseudo-mask supervision. For these matters, we propose the following designs to push the performance to new state-of-art: (i) Coefficient of Variation Smoothing to smooth the CAMS adaptively; (ii) Proportional Pseudo-mask Generation to project the expanded CAMS to pseudo-mask based on a new metric indicating the importance of each class on each location, instead of the scores trained from binary classifiers. (iii) Pretended Under-Fitting strategy to suppress the influence of noise in pseudo-mask; (iv) Cyclic Pseudo-mask to boost the pseudo-masks during training of fully supervised semantic segmentation (FSSS). Experiments based on our methods achieve new state-of-art results on two challenging weakly supervised semantic segmentation datasets, pushing the mIoU to 70.0% and 40.2% on PAS-CAL VOC 2012 and MS COCO 2014 respectively. Codes including segmentation framework are released at <https://github.com/Eli-YiLi/PMM>

大多数弱监督语义切分（WSSS）方法遵循最初生成伪掩码的管道，然后以完全监督的方式使用伪掩码训练切分模型。然而，我们发现了一些与伪掩模相关的问题，包括从类激活映射（CAM）生成高质量的伪掩模，以及使用噪声伪掩模监督进行训练。针对这些问题，我们提出了以下设计，以将性能提升到新的水平：(i) 变异系数平滑以自适应平滑凸轮；(ii) 按比例生成伪掩码，以基于指示每个位置上每个类别重要性的新度量，而不是从二进制分类器训练的分数，将扩展的CAM投影到伪掩码。(iii) 在拟合策

略下假装，以抑制伪掩模中噪声的影响；(iv) 循环伪掩码，用于在全监督语义切分(FSSS)训练期间增强伪掩码。基于我们的方法的实验在两个转换弱监督语义分割数据集上获得了最新的结果，在PASCAL VOC 2012和MS COCO 2014上，mIoU分别达到70.0%和40.2%。包括分段框架在内的代码在<https://github.com/Eli-YiLi/PMM>

We propose a novel framework for finding correspondences in images based on a deep neural network that, given two images and a query point in one of them, finds its correspondence in the other. By doing so, one has the option to query only the points of interest and retrieve sparse correspondences, or to query all points in an image and obtain dense mappings. Importantly, in order to capture both local and global priors, and to let our model relate between image regions using the most relevant among said priors, we realize our network using a transformer. At inference time, we apply our correspondence network by recursively zooming in around the estimates, yielding a multi-scale pipeline able to provide highly-accurate correspondences. Our method significantly outperforms the state-of-the-art on both sparse and dense correspondence problems on multiple datasets and tasks, ranging from wide-baseline stereo to optical flow, without any retraining for a specific dataset.

我们提出了一种新的基于深度神经网络的图像匹配框架，该框架在给定两幅图像和其中一幅图像中的一个查询点的情况下，在另一幅图像中查找其对应关系。通过这样做，可以选择只查询感兴趣的点并检索稀疏的对应关系，或者查询图像中的所有点并获得密集映射。重要的是，为了捕获局部和全局先验，并让我们的模型使用所述先验中最相关的先验在图像区域之间建立关联，我们使用变压器来实现我们的网络。在推断时，我们通过递归放大估计值来应用我们的对应网络，产生一个能够提供高度精确对应的多尺度管道。我们的方法在多个数据集和任务（从宽基线立体到光流）上的稀疏和密集对应问题上显著优于最新技术，而无需对特定数据集进行任何再培训。

Previous pseudo-label approaches for semi-supervised object detection typically follow a multi-stage schema, with the first stage to train an initial detector on a few labeled data, followed by the pseudo labeling and re-training stage on unlabeled data. These multi-stage methods complicate the training, and also hinder the use of improved detectors for more accurate pseudo-labeling. In this paper, we propose an end-to-end approach to simultaneously improve the detector and pseudo labels gradually for semi-supervised object detection. The pseudo labels are generated on the fly by a teacher model which is an aggregated version of the student detector at different steps. As the detector becomes stronger during the training, the teacher detector's performance improves and the pseudo labels tend to be more accurate, which further benefits the detector training. Within the end-to-end training, we present two simple yet effective techniques: weigh the classification loss of unlabeled images through soft teacher and select reliable pseudo boxes for regression through box jittering. Experimentally, the proposed approach outperforms the state-of-the-art methods by a large margin on MS-COCO benchmark by using Faster R-CNN with ResNet-50 and FPN, reaching 20.5 mAP, 30.7 mAP and 34.0 mAP with 1%, 5%, 10% labeled data, respectively. Moreover, the proposed approach also proves to improve this detector trained on the COCO full set by +1.8 mAP by leveraging additional unlabelled data of COCO, achieving 42.7 mAP.

以前用于半监督对象检测的伪标记方法通常遵循多阶段模式，第一阶段是在几个标记数据上训练初始检测器，然后是在未标记数据上的伪标记和重新训练阶段。这些多阶段方法使训练复杂化，也阻碍了使用改进的检测器进行更精确的伪标记。在本文中，我们提出了一种端到端的方法来同时改进检测器和伪标签，以实现半监督目标检测。伪标签由教师模型动态生成，教师模型是学生检测器在不同步骤的聚合版本。随着检测器在培训过程中变得更强，教师检测器的性能提高，伪标签趋于更准确，这进一步有利于检测器培训。在端到端训练中，我们提出了两种简单而有效的技术：通过软教师衡量未标记图像的分类损失，并通过框抖动选择可靠的伪框进行回归。实验表明，该方法在MS-COCO基准上比现有方法有很大

的优势，使用更快的R-CNN和ResNet-50和FPN，分别达到20.5 mAP、30.7 mAP和34.0 mAP，标记数据分别为1%、5%和10%。此外，所提出的方法还证明，通过利用额外的COCO未标记数据，通过+1.8 mAP改进了COCO全集上训练的检测器，实现了42.7 mAP。

Accuracy predictor is a key component in Neural Architecture Search (NAS) for ranking architectures. Building a high-quality accuracy predictor usually costs enormous computation. To address this issue, instead of using an accuracy predictor, we propose a novel zero-shot index dubbed Zen-Score to rank the architectures. The Zen-Score represents the network expressivity and positively correlates with the model accuracy. The calculation of Zen-Score only takes a few forward inferences through a randomly initialized network, without training network parameters. Built upon the Zen-Score, we further propose a new NAS algorithm, termed as Zen-NAS, by maximizing the Zen-Score of the target network under given inference budgets. Within less than half GPU day, Zen-NAS is able to directly search high performance architectures in a data-free style. Comparing with previous NAS methods, the proposed Zen-NAS is magnitude times faster on multiple server-side and mobile-side GPU platforms with state-of-the-art accuracy on ImageNet. Searching and training code as well as pre-trained models are available from <https://github.com/idstcv/ZenNAS>.

准确度预测器是神经结构搜索（NAS）中用于排序结构的关键部件。建立一个高质量的精度预测通常需要大量的计算。为了解决这个问题，我们提出了一种称为Zen Score的新的零炮指数来对架构进行排序，而不是使用精度预测。Zen分数代表网络表现力，并与模型精度呈正相关。Zen分数的计算只需要通过随机初始化的网络进行一些正向推断，而不需要训练网络参数。在Zen评分的基础上，我们进一步提出了一种新的NAS算法，称为Zen NAS，该算法通过在给定的推理预算下最大化目标网络的Zen评分。在不到半天的GPU时间内，Zen NAS能够以无数据的方式直接搜索高性能体系结构。与以前的NAS方法相比，所提出的Zen NAS在多个服务器端和移动端GPU平台上的速度快了数倍，在ImageNet上具有最先进的精度。搜索和培训代码以及预先培训的模型可从<https://github.com/idstcv/ZenNAS>.

The light transport matrix (LTM) is an instrumental tool in line-of-sight (LOS) imaging, describing how light interacts with the scene and enabling applications such as relighting or separation of illumination components. We introduce a framework to estimate the LTM of non-line-of-sight (NLOS) scenarios, coupling recent virtual forward light propagation models for NLOS imaging with the LOS light transport equation. We design computational projector-camera setups, and use these virtual imaging systems to estimate the transport matrix of hidden scenes. We introduce the specific illumination functions to compute the different elements of the matrix, overcoming the challenging wide-aperture conditions of NLOS setups. Our NLOS light transport matrix allows us to (re)illuminate specific locations of a hidden scene, and separate direct, first-order indirect, and higher-order indirect illumination of complex cluttered hidden scenes, similar to existing LOS techniques.

光传输矩阵（LTM）是视线（LOS）成像中的一种工具，它描述了光如何与场景交互，并支持重新照明或照明组件分离等应用。我们引入一个框架来估计非视距（NLOS）场景的LTM，将最近的NLOS成像虚拟前向光传播模型与LOS光传输方程耦合起来。我们设计了计算投影仪相机装置，并使用这些虚拟成像系统来估计隐藏场景的传输矩阵。我们引入了特定的照明函数来计算矩阵的不同元素，克服了NLOS设置的挑战性宽孔径条件。我们的NLOS光传输矩阵允许我们（重新）照亮隐藏场景的特定位置，并分离复杂杂乱隐藏场景的直接、一阶间接和高阶间接照明，类似于现有的LOS技术。

The scale of deep learning nowadays calls for efficient distributed training algorithms. Decentralized momentum SGD (DmSGD), in which each node averages only with its neighbors, is more communication efficient than vanilla Parallel momentum SGD that incurs global average across all computing nodes. On the other hand, the large-batch training has been demonstrated critical to achieve runtime speedup. This motivates us to investigate how DmSGD performs in the large-batch scenario. In this work, we find the momentum term can amplify the inconsistency bias in DmSGD. Such bias becomes more evident as batch-size grows large and hence results in severe performance degradation. We next propose DecentLaM, a novel decentralized large-batch momentum SGD to remove the momentum-incurred bias. The convergence rate for both strongly convex and non-convex scenarios is established. Our theoretical results justify the superiority of DecentLaM to DmSGD especially in the large-batch scenario. Experimental results on a variety of computer vision tasks and models show that DecentLaM promises both efficient and high-quality training.

当今深度学习的规模要求高效的分布式训练算法。分散动量SGD (DmSGD) , 其中每个节点仅与其邻居进行平均，比在所有计算节点上产生全局平均的普通并行动量SGD具有更高的通信效率。另一方面，大批量培训对于实现运行时加速至关重要。这促使我们研究DmSGD在大批量场景中的表现。在这项工作中，我们发现动量项可以放大DmSGD中的不一致性偏差。随着批量的增大，这种偏差变得更加明显，从而导致严重的性能下降。接下来，我们提出DecentLaM，一种新型的分散式大批量动量SGD，以消除动量偏差。建立了强凸和非凸情形下的收敛速度。我们的理论结果证明了DecentLaM相对于DmSGD的优越性，特别是在大批量情况下。在各种计算机视觉任务和模型上的实验结果表明，DecentLaM能够保证高效和高质量的训练。

In this paper, we propose a novel solution for object-matching based semi-supervised video object segmentation, where the target object masks in the first frame are provided. Existing object-matching based methods focus on the matching between the raw object features of the current frame and the first/previous frames. However, two issues are still not solved by these object-matching based methods. As the appearance of the video object changes drastically over time, 1) unseen parts/details of the object present in the current frame, resulting in incomplete annotation in the first annotated frame (e.g., view/scale changes). 2) even for the seen parts/details of the object in the current frame, their positions change relatively (e.g., pose changes/camera motion), leading to a misalignment for the object matching. To obtain the complete information of the target object, we propose a novel object-based dynamic memory network that exploits visual contents of all the past frames. To solve the misalignment problem caused by position changes of visual contents, we propose an adaptive object alignment module by incorporating a region translation function that aligns object proposals towards templates in the feature space. Our method achieves state-of-the-art results on latest benchmark datasets DAVIS 2017 (J of 81.4% and F of 87.5% on the validation set) and YouTube-VOS (the overall score of 82.7% on the validation set) with a very efficient inference time (0.16 second/frame on DAVIS 2017 validation set). Code is available at:  
<https://github.com/liang4sx/DMN-AOA>.

本文提出了一种新的基于目标匹配的半监督视频对象分割方法，该方法在第一帧中提供了目标对象模板。现有的基于对象匹配的方法侧重于当前帧的原始对象特征与第一帧/前一帧之间的匹配。然而，这些基于对象匹配的方法仍然没有解决两个问题。随着视频对象的外观随时间急剧变化，1) 当前帧中存在的对象的不可见部分/细节，导致第一个带注释帧中的注释不完整（例如，视图/比例变化）。2) 即使对于当前帧中对象的可见部分/细节，它们的位置也会发生相对变化（例如，姿势变化/摄影机运动），从而导致对象匹配不对齐。为了获得目标对象的完整信息，我们提出了一种新的基于对象的动态记忆网络，该网络利用了所有过去帧的视觉内容。为了解决视觉内容位置变化引起的错位问题，我们提出了一种自适应对象对齐模块，该模块结合了区域平移功能，将对象建议与特征空间中的模板对齐。我们的方法在

最新的基准数据集DAVIS 2017（验证集的J值为81.4%，F值为87.5%）和YouTube VOS（验证集的总分为82.7%）上获得了最先进的结果，并且推理时间非常高效（DAVIS 2017验证集为0.16秒/帧）。代码可以从以下网址获取：<https://github.com/liang4sx/DMN-AOA>.

Adversarial attack algorithms are dominated by penalty methods, which are slow in practice, or more efficient distance-customized methods, which are heavily tailored to the properties of the considered distance. We propose a white-box attack algorithm to generate minimally perturbed adversarial examples based on Augmented Lagrangian principles. We bring several algorithmic modifications, which have a crucial effect on performance. Our attack enjoys the generality of penalty methods and the computational efficiency of distance-customized algorithms, and can be readily used for a wide set of distances. We compare our attack to state-of-the-art methods on three datasets and several models, and consistently obtain competitive performances with similar or lower computational complexity.

对抗性攻击算法主要是惩罚方法，这种方法在实践中很慢，或者是更有效的距离定制方法，它们根据所考虑的距离的特性进行了大量定制。我们提出了一种基于增广拉格朗日原理的白盒攻击算法来生成最小扰动对抗性示例。我们对算法进行了一些修改，这些修改对性能有着至关重要的影响。我们的攻击具有惩罚方法的通用性和距离定制算法的计算效率，并且可以很容易地用于广泛的距离集。我们在三个数据集和几个模型上将我们的攻击与最先进的方法进行比较，并一致地获得具有类似或更低计算复杂度的竞争性能。

This paper proposes a method for representation learning of multimodal data using contrastive losses. A traditional approach is to contrast different modalities to learn the information shared between them. However, that approach could fail to learn the complementary synergies between modalities that might be useful for downstream tasks. Another approach is to concatenate all the modalities into a tuple and then contrast positive and negative tuple correspondences. However, that approach could consider only the stronger modalities while ignoring the weaker ones. To address these issues, we propose a novel contrastive learning objective, TupleInfoNCE. It contrasts tuples based not only on positive and negative correspondences, but also by composing new negative tuples using modalities describing different scenes. Training with these additional negatives encourages the learning model to examine the correspondences among modalities in the same tuple, ensuring that weak modalities are not ignored. We provide a theoretical justification based on mutual-information for why this approach works, and we propose a sample optimization algorithm to generate positive and negative samples to maximize training efficacy. We find that TupleInfoNCE significantly outperforms previous state of the arts on three different downstream tasks.

提出了一种基于对比损失的多模态数据表示学习方法。传统的方法是对比不同的模式来学习它们之间共享的信息。然而，这种方法可能无法了解可能对下游任务有用的模式之间的互补协同作用。另一种方法是将所有模态连接成一个元组，然后对比正元组和负元组对应关系。然而，这种方法只能考虑更强的模态而忽略较弱的模态。为了解决这些问题，我们提出了一个新的对比学习目标，TupleInfo。它不仅基于正负对应，而且通过使用描述不同场景的模式组合新的负元组来对比元组。使用这些附加否定项的训练鼓励学习模型检查同一元组中模式之间的对应关系，确保不忽略弱模式。我们基于互信息为这种方法的工作原理提供了理论依据，并提出了一种样本优化算法来生成正样本和负样本，以最大限度地提高训练效率。我们发现，TupleInfo在三种不同的下游任务上显著优于以前的技术水平。

We propose a deep reparametrization of the maximum a posteriori formulation commonly employed in multi-frame image restoration tasks. Our approach is derived by introducing a learned error metric and a latent representation of the target image, which transforms the MAP objective to a deep feature space. The deep reparametrization allows us to directly model the image formation process in the latent space, and to integrate learned image priors into the prediction. Our approach thereby leverages the advantages of deep learning, while also benefiting from the principled multi-frame fusion provided by the classical MAP formulation. We validate our approach through comprehensive experiments on burst denoising and burst super-resolution datasets. Our approach sets a new state-of-the-art for both tasks, demonstrating the generality and effectiveness of the proposed formulation.

我们对多帧图像恢复任务中常用的最大后验公式提出了一种深度再参数化方法。我们的方法是通过引入学习的误差度量和目标图像的潜在表示，将地图目标转换为深层特征空间。深度再参数化允许我们直接在潜在空间中建模图像形成过程，并将学到的图像先验知识集成到预测中。因此，我们的方法利用了深度学习的优势，同时也受益于经典MAP公式提供的原则性多帧融合。我们通过对突发去噪和突发超分辨率数据集的综合实验来验证我们的方法。我们的方法为这两项任务设定了一个新的最先进水平，证明了所提议公式的通用性和有效性。

We present pure-transformer based models for video classification, drawing upon the recent success of such models in image classification. Our model extracts spatio-temporal tokens from the input video, which are then encoded by a series of transformer layers. In order to handle the long sequences of tokens encountered in video, we propose several, efficient variants of our model which factorise the spatial- and temporal-dimensions of the input. Although transformer-based models are known to only be effective when large training datasets are available, we show how we can effectively regularise the model during training and leverage pretrained image models to be able to train on comparatively small datasets. We conduct thorough ablation studies, and achieve state-of-the-art results on multiple video classification benchmarks including Kinetics 400 and 600, Epic Kitchens, Something-Something v2 and Moments in Time, outperforming prior methods based on deep 3D convolutional networks. To facilitate further research, we will release code and models.

我们提出了纯变压器为基础的视频分类模型，借鉴了最近成功的图像分类模型。我们的模型从输入视频中提取时空标记，然后通过一系列变换层对其进行编码。为了处理视频中遇到的长序列标记，我们提出了几个有效的模型变体，用于分解输入的空间和时间维度。虽然已知基于转换器的模型仅在大的训练数据集可用时有效，但我们展示了如何在训练期间有效地正则化模型，并利用预训练图像模型在相对较小的数据集上进行训练。我们进行了彻底的消融研究，并在多个视频分类基准上取得了最先进的结果，包括Kinetics 400和600、Epic Kitchens、Something v2和Moments in Time，优于以前基于深3D卷积网络的方法。为了便于进一步研究，我们将发布代码和模型。

Inspired by the ability of StyleGAN to generate highly realistic images in a variety of domains, much recent work has focused on understanding how to use the latent spaces of StyleGAN to manipulate generated and real images. However, discovering semantically meaningful latent manipulations typically involves painstaking human examination of the many degrees of freedom, or an annotated collection of images for each desired manipulation. In this work, we explore leveraging the power of recently introduced Contrastive Language-Image Pre-training (CLIP) models in order to develop a text-based interface for StyleGAN image manipulation that does not require such manual effort. We first introduce an optimization scheme that utilizes a CLIP-based loss to modify an input latent vector in response to a user-provided text prompt. Next, we describe a latent mapper that infers a text-guided latent manipulation step for a given input image, allowing faster and more stable text-based manipulation. Finally, we present a method for mapping a text prompt to input-agnostic directions in StyleGAN's style space, enabling interactive text-driven image manipulation. Extensive results and comparisons demonstrate the effectiveness of our approaches.

受StyleGAN在不同领域生成高度真实图像的能力的启发，最近的许多工作都集中在理解如何使用StyleGAN的潜在空间来处理生成的和真实的图像。然而，发现语义上有意义的潜在操作通常需要人类对多个自由度进行艰苦的检查，或者为每个期望的操作收集带注释的图像。在这项工作中，WeeExplore利用最近引入的对比语言图像预训练（CLIP）模型的功能，为StyleGAN图像处理开发一个基于文本的界面，而不需要手动操作。我们首先介绍一种优化方案，该方案利用基于剪辑的丢失来修改输入潜在向量，以响应用户提供的文本提示。接下来，我们描述了一个潜在映射器，它为给定的输入图像推断出一个文本引导的潜在操作步骤，允许更快、更稳定的基于文本的操作。最后，我们提出了一种将文本提示映射为ping的方法，以在StyleGAN的样式空间中输入不可知方向，从而实现交互式文本驱动的图像处理。大量的结果和比较证明了我们方法的有效性。

Modern deep neural networks are often vulnerable to adversarial examples. Most exist attack methods focus on crafting adversarial examples in the digital domain, while only limited works study physical adversarial attack. However, it is more challenging to generate effective adversarial examples in the physical world due to many uncontrollable physical dynamics. Most current physical attack methods aim to generate robust physical adversarial examples by simulating all possible physical dynamics. When attacking new images or new DNN models, they require expensive manually efforts for simulating physical dynamics and considerable time for iteratively optimizing for each image. To tackle these issues, we propose a class-agnostic and model-agnostic physical adversarial attack model (Meta-Attack), which is able to not only generate robust physical adversarial examples by simulating color and shape distortions, but also generalize to attacking novel images and novel DNN models by accessing a few digital and physical images. To the best of our knowledge, this is the first work to formulate the physical attack as a few-shot learning problem. Here, the training task is redefined as the composition of a support set, a query set, and a target DNN model. Under the few-shot setting, we design a novel class-agnostic and model-agnostic meta-learning algorithm to enhance the generalization ability of our method. Extensive experimental results on two benchmark datasets with four challenging experimental settings verify the superior robustness and generalization of our method by comparing to state-of-the-art physical attack methods.

现代深层神经网络往往容易受到敌对例子的攻击。现有的攻击方法大多侧重于在数字领域中制作对抗性示例，而研究物理对抗性攻击的工作却非常有限。然而，由于许多不可控的物理动力学，在物理世界中生成有效的对抗示例更具挑战性。当前大多数物理攻击方法的目标是通过模拟所有可能的物理动力学来生成健壮的物理对抗示例。当攻击新图像或新的DNN模型时，它们需要花费昂贵的手动工作来模拟物理动力学，并且需要相当长的时间来迭代优化每个图像。为了解决这些问题，我们提出了一个类不可知和

模型不可知的物理对抗攻击模型（元攻击），该模型不仅能够通过模拟颜色和形状扭曲生成健壮的物理对抗示例，但也可以通过访问一些数字和物理图像来攻击新图像和新的DNN模型。据我们所知，这是第一个将物理攻击表述为几个射击学习问题的工作。这里，训练任务被重新定义为支持集、查询集和目标DNN模型的组成部分。在少数镜头设置下，我们设计了一种新的类不可知和模型不可知元学习算法，以增强方法的泛化能力。在两个基准数据集和四个具有挑战性的实验设置上的大量实验结果通过与最先进的物理攻击方法的比较，验证了我们的方法优越的鲁棒性和泛化性。

This presentation addresses the problem of reconstructing a high-resolution image from multiple lower-resolution snapshots captured from slightly different viewpoints in space and time. Key challenges for solving this super-resolution problem include (i) aligning the input pictures with sub-pixel accuracy, (ii) handling raw (noisy) images for maximal faithfulness to native camera data, and (iii) designing/learning an image prior (regularizer) well suited to the task. We address these three challenges with a hybrid algorithm building on the insight from Wronski et al. that aliasing is an ally in this setting, with parameters that can be learned end to end, while retaining the interpretability of classical approaches to inverse problems. The effectiveness of our approach is demonstrated on synthetic and real image bursts, setting a new state of the art on several benchmarks and delivering excellent qualitative results on real raw bursts captured by smartphones and prosumer cameras.

本演示介绍如何从从空间和时间上略有不同的视点捕获的多个低分辨率快照重建高分辨率图像。解决此超分辨率问题的关键挑战包括 (i) 将输入图片与亚像素精度对齐，(ii) 处理原始（含噪）图像以最大程度地忠实于本机相机数据，以及 (iii) 设计/学习非常适合此任务的图像先验（正则化器）。我们基于 Wronski 等人的见解，采用混合算法解决了这三个难题。在这种情况下，混叠是一个盟友，参数可以端到端学习，同时保留了反问题经典方法的可解释性。我们的方法的有效性在合成和真实图像突发上得到了证明，在多个基准上开创了新的技术水平，并在智能手机和 prosumer 相机捕获的真实原始突发上提供了出色的定性结果。

RGB-D based 6D pose estimation has recently achieved remarkable progress, but still suffers from two major limitations: (1) ineffective representation of depth data and (2) insufficient integration of different modalities. This paper proposes a novel deep learning approach, namely Graph Convolutional Network with Point Refinement (PR-GCN), to simultaneously address the issues above in a unified way. It first introduces the Point Refinement Network (PRN) to polish 3D point clouds, recovering missing parts with noise removed. Subsequently, the Multi-Modal Fusion Graph Convolutional Network (MMF-GCN) is presented to strengthen RGB-D combination, which captures geometry-aware inter-modality correlation through local information propagation in the graph convolutional network. Extensive experiments are conducted on three widely used benchmarks, and state-of-the-art performance is reached. Besides, it is also shown that the proposed PRN and MMF-GCN modules are well generalized to other frameworks.

基于RGB-D的6D位姿估计最近取得了显著的进展，但仍然存在两个主要局限性：(1) 深度数据的无效表示；(2) 不同模式的集成不足。本文提出了一种新的深度学习方法，即带点细化的图卷积网络 (PR-GCN)，以统一的方式同时解决上述问题。它首先引入点细化网络 (PRN) 来抛光3D点云，在去除噪声的情况下恢复缺失的部分。随后，提出了多模态融合图卷积网络 (MMF-GCN) 来增强RGB-D组合，该网络通过图卷积网络中的局部信息传播来捕获几何感知的模态间相关性。在三个广泛使用的基准上进行了大量实验，达到了最先进的性能。此外，还表明所提出的PRN和MMF-GCN模块可以很好地推广到其他框架。

As a fundamental building block in computer vision, edges can be categorised into four types according to the discontinuity in surface-Reflectance, illumination, surface-Normal or Depth. While great progress has been made in detecting generic or individual types of edges, it remains under-explored to comprehensively study all four edge types together. In this paper, we propose a novel neural network solution, RINDNet, to jointly detect all four types of edges. Taking into consideration the distinct attributes of each type of edges and the relationship between them, RINDNet learns effective representations for each of them and works in three stages. In stage I, RINDNet uses a common backbone to extract features shared by all edges. Then in stage II it branches to prepare discriminative features for each edge type by the corresponding decoder. In stage III, an independent decision head for each type aggregates the features from previous stages to predict the initial results. Additionally, an attention module learns attention maps for all types to capture the underlying relations between them, and these maps are combined with initial results to generate the final edge detection results. For training and evaluation, we construct the first public benchmark, BSDS-RIND, with all four types of edges carefully annotated. In our experiments, RINDNet yields promising results in comparison with state-of-the-art methods. Additional analysis is presented in supplementary material.

边缘作为计算机视觉的基本组成部分，根据表面反射率、照度、表面法线或深度的不连续性可分为四种类型。虽然在检测一般或个别类型的边缘方面已经取得了很大的进展，但综合研究所有四种边缘类型仍处于探索阶段。在本文中，我们提出了一种新的神经网络解决方案RINDNet，用于联合检测所有四种类型的边缘。考虑到每种边类型的不同属性以及它们之间的关系，RINDNet学习每种边的有效表示，并分三个阶段工作。在第一阶段，RINDNet使用公共主干提取所有边共享的特征。然后在第二阶段，它分支，通过相应的解码器为每种边缘类型准备鉴别特征。在第三阶段，每种类型的独立决策负责人汇总前一阶段的特征，以预测初始结果。此外，注意力模块学习所有类型的注意力贴图，以捕获它们之间的潜在关系，并将这些贴图与初始结果相结合，以生成最终的边缘检测结果。为了进行培训和评估，我们构建了第一个公共基准BSDS-RIND，并对所有四种类型的边进行了仔细注释。在我们的实验中，与最先进的方法相比，RINDNet产生了有希望的结果。补充材料中提供了额外的分析。

Pseudo-LiDAR-based methods for monocular 3D object detection have received considerable attention in the community due to the performance gains exhibited on the KITTI3D benchmark, in particular on the commonly reported validation split. This generated a distorted impression about the superiority of Pseudo-LiDAR-based (PL-based) approaches over methods working with RGB images only. Our first contribution consists in rectifying this view by pointing out and showing experimentally that the validation results published by PL-based methods are substantially biased. The source of the bias resides in an overlap between the KITTI3D object detection validation set and the training/validation sets used to train depth predictors feeding PL-based methods. Surprisingly, the bias remains also after geographically removing the overlap. This leaves the test set as the only reliable set for comparison, where published PL-based methods do not excel. Our second contribution brings PL-based methods back up in the ranking with the design of a novel deep architecture which introduces a 3D confidence prediction module. We show that 3D confidence estimation techniques derived from RGB-only 3D detection approaches can be successfully integrated into our framework and, more importantly, that improved performance can be obtained with a newly designed 3D confidence measure, leading to state-of-the-art performance on the KITTI3D benchmark.

基于伪激光雷达的单目3D目标检测方法由于在KITTI3D基准上表现出的性能提升，特别是在通常报告的验证分割上，在社区中受到了相当大的关注。这给人们留下了一个扭曲的印象，即基于伪激光雷达（PL）的方法优于仅处理RGB图像的方法。我们的第一个贡献在于纠正这一观点，指出并通过实验证明基于PL的方法发布的验证结果存在严重偏差。偏差的来源在于KITTI3D对象检测验证集和用于训练基于

PL的方法的深度预测器的训练/验证集之间的重叠。令人惊讶的是，在地理上消除重叠后，这种偏见仍然存在。这使得测试集成为唯一可靠的比较集，而已发布的基于PL的方法并不擅长于此。我们的第二个贡献是设计了一种新的深度体系结构，引入了一个3D置信度预测模块，从而使基于PL的方法重新回到了排名中。我们表明，从仅RGB 3D检测方法衍生出的3D置信度估计技术可以成功地集成到我们的框架中，更重要的是，通过新设计的3D置信度度量可以获得更高的性能，从而在KITTI3D基准上实现最先进的性能。

Benchmark datasets that measure camera pose accuracy have driven progress in visual re-localisation research. To obtain poses for thousands of images, it is common to use a reference algorithm to generate pseudo ground truth. Popular choices include Structure-from-Motion (SfM) and Simultaneous-Localisation-and-Mapping (SLAM) using additional sensors like depth cameras if available. Re-localisation benchmarks thus measure how well each method replicates the results of the reference algorithm. This begs the question whether the choice of the reference algorithm favours a certain family of re-localisation methods. This paper analyzes two widely used re-localisation datasets and shows that evaluation outcomes indeed vary with the choice of the reference algorithm. We thus question common beliefs in the re-localisation literature, namely that learning-based scene coordinate regression outperforms classical feature-based methods, and that RGB-D-based methods outperform RGB-based methods. We argue that any claims on ranking re-localisation methods should take the type of the reference algorithm, and the similarity of the methods to the reference algorithm, into account.

测量相机姿态精度的基准数据集推动了视觉重新定位研究的进展。为了获得数千张图像的姿势，通常使用参考算法生成伪地面真值。流行的选择包括运动结构（SfM）和使用其他传感器（如深度摄像机）的同步定位和映射（SLAM）。因此，重新定位基准测量每种方法复制参考算法结果的效果。这就引出了一个问题：参考算法的选择是否有利于某种重新定位方法。本文分析了两个广泛使用的重新定位数据集，表明评估结果确实随参考算法的选择而变化。因此，我们质疑重新定位文献中的共同信念，即基于学习的场景坐标回归优于经典的基于特征的方法，基于RGB-D的方法优于基于RGB的方法。我们认为，任何关于排名再定位方法的主张都应该考虑参考算法的类型，以及方法与参考算法的相似性。

The minimal geodesic models based on the Eikonal equations are capable of finding suitable solutions in various image segmentation scenarios. Existing geodesic-based segmentation approaches usually exploit the image features in conjunction with geometric regularization terms (such as curve length or elastica length) for computing geodesic paths. In this paper, we consider a more complicated problem: finding simple and closed geodesic curves which are imposed a convexity shape prior. The proposed approach relies on an orientation-lifting strategy, by which a planar curve can be mapped to an high-dimensional orientation space. The convexity shape prior serves as a constraint for the construction of local metrics. The geodesic curves in the lifted space then can be efficiently computed through the fast marching method. In addition, we introduce a way to incorporate region-based homogeneity features into the proposed geodesic model so as to solve the region-based segmentation issues with shape prior constraints.

基于Eikonal方程的最小测地线模型能够在各种图像分割场景中找到合适的解。现有的基于测地线的分割方法通常利用图像特征和几何正则化项（如曲线长度或弹性长度）来计算测地线路径。在本文中，我们考虑一个更复杂的问题：找到简单的和封闭的测地线，这是强加的凸形状之前。所提出的方法依赖于方向提升策略，通过该策略可以将平面曲线映射到高维方向空间。凸性形状先验作为构造局部度量的约束。然后，可以通过快速行进法有效地计算提升空间中的测地线曲线。此外，我们还介绍了一种将基于区域的同质性特征融入到所提出的测地线模型中的方法，以解决具有形状先验约束的基于区域的分割问题。

Fast arbitrary neural style transfer has attracted widespread attention from academic, industrial and art communities due to its flexibility in enabling various applications. Existing solutions either attentively fuse deep style feature into deep content feature without considering feature distributions, or adaptively normalize deep content feature according to the style such that their global statistic information is matched. Although effective, leaving shallow feature unexplored or without locally considering feature statistics, they are prone to suffer from unnatural output with unpleasing local distortions. To alleviate this problem, in this paper, we propose a novel Adaptive Attention Normalization (AdaAttN) module to adaptively perform attentive normalization on per-point basis. Specifically, spatial attention score is learnt from both shallow and deep features of content and style images. Then per-point weighted statistics are calculated by regarding a style feature point as a distribution of attention-weighted output of all style feature points. Finally, the content feature is normalized so that they demonstrate the same local feature statistics as the calculated per-point weighted style feature statistics. Besides, a novel local feature loss is derived based on AdaAttN to enhance local visual quality. We also extend AdaAttN to be ready for video style transfer with slight modifications. Extensive experiments demonstrate that our method achieves state-of-the-art arbitrary image/video style transfer. Codes and models will be available.

快速任意神经风格转换因其在各种应用中的灵活性而引起学术界、工业界和艺术界的广泛关注。现有的解决方案要么在不考虑特征分布的情况下仔细地将深度样式特征融合到深度内容特征中，要么根据样式自适应地规范化深度内容特征，使其全局统计信息相匹配。虽然有效，留下浅层特征未被探索或没有局部考虑特征统计，但它们容易遭受非自然输出和令人不快的局部扭曲。为了缓解这一问题，本文提出了一种新的自适应注意规范化（AdaAttN）模块，该模块可以在每个点的基础上自适应地执行注意规范化。具体来说，空间注意分数是从内容和风格图像的浅层和深层特征中学习的。然后，通过将一个样式特征点视为所有样式特征点的注意力加权输出的分布来计算每点加权统计。最后，对内容特征进行规范化，以便它们显示与计算的逐点加权样式特征统计相同的局部特征统计。此外，为了提高局部视觉质量，基于AdaAttN提出了一种新的局部特征丢失算法。我们还对AdaAttN进行了扩展，以便在稍作修改的情况下进行视频样式转换。大量实验表明，我们的方法实现了最先进的任意图像/视频样式传输。将提供代码和型号。

Arbitrary shape text detection is a challenging task due to the high complexity and variety of scene texts. In this work, we propose a novel adaptive boundary proposal network for arbitrary shape text detection, which can learn to directly produce accurate boundary for arbitrary shape text without any post-processing. Our method mainly consists of a boundary proposal model and an innovative adaptive boundary deformation model. The boundary proposal model constructed by multi-layer dilated convolutions is adopted to produce prior information (including classification map, distance field, and direction field) and coarse boundary proposals. The adaptive boundary deformation model is an encoder-decoder network, in which the encoder mainly consists of a Graph Convolutional Network (GCN) and a Recurrent Neural Network (RNN). It aims to perform boundary deformation in an iterative way for obtaining text instance shape guided by prior information from the boundary proposal model. In this way, our method can directly and efficiently generate accurate text boundaries without complex post-processing. Extensive experiments on publicly available datasets demonstrate the state-of-the-art performance of our method.

由于场景文本的高度复杂性和多样性，任意形状文本检测是一项具有挑战性的任务。在这项工作中，我们提出了一种新的用于任意形状文本检测的自适应边界建议网络，它可以学习直接为任意形状文本生成精确的边界，而无需任何后处理。我们的方法主要包括一个边界建议模型和一个创新的自适应边界变形模型。采用多层扩张卷积构造的边界建议模型产生先验信息（包括分类图、距离场和方向场）和粗边界

建议。自适应边界变形模型是一个编解码网络，其中编码器主要由一个图卷积网络（GCN）和一个递归神经网络（RNN）组成。它的目的是以迭代的方式执行边界变形，以获得由边界建议模型的先验信息引导的文本实例形状。这样，我们的方法可以直接有效地生成精确的文本边界，而无需复杂的后处理。在公开数据集上的大量实验证明了我们方法的最新性能。

Learning temporally consistent foreground opacity from videos, i.e., video matting, has drawn great attention due to the blossoming of video conferencing. Previous approaches are built on top of image matting models, which fail in maintaining the temporal coherence when being adapted to videos. They either utilize the optical flow to smooth frame-wise prediction, where the performance is dependent on the selected optical flow model; or naively combine feature maps from multiple frames, which does not model well the correspondence of pixels in adjacent frames. In this paper, we propose to enhance the temporal coherence by Consistency-Regularized Graph Neural Networks (CRGNN) with the aid of a synthesized video matting dataset. CRGNN utilizes Graph Neural Networks (GNN) to relate adjacent frames such that pixels or regions that are incorrectly predicted in one frame can be corrected by leveraging information from its neighboring frames. To generalize our model from synthesized videos to real-world videos, we propose a consistency regularization technique to enforce the consistency on the alpha and foreground when blending them with different backgrounds. To evaluate the efficacy of CRGNN, we further collect a real-world dataset with annotated alpha mattes. Compared with state-of-the-art methods that require hand-crafted trimaps or backgrounds for modeling training, CRGNN generates favorably results with the help of unlabeled real training dataset.

由于视频会议的蓬勃发展，从视频中学习时间一致的前景不透明度，即视频抠图，已经引起了极大的关注。以前的方法是建立在图像抠图模型的基础上的，在适应视频时，该模型无法保持时间一致性。它们或者利用光流平滑帧预测，其中性能取决于所选光流模型；或者天真地组合来自多个帧的特征映射，这不能很好地模拟相邻帧中像素的对应关系。在这篇文章中，我们提出了利用一致性正则化图神经网络（CRGNN）和合成的视频抠图数据集来增强时间相干性。CRGNN利用图形神经网络（GNN）来关联相邻帧，这样，在一个帧中错误预测的像素或区域可以通过利用其相邻帧的信息进行校正。为了将我们的模型从合成视频推广到现实世界的视频，我们提出了一种一致性正则化技术，用于在混合不同背景时增强alpha和前景的一致性。为了评估CRGNN的有效性，我们进一步收集了一个带有注释alpha mattes的真实数据集。与需要手工制作TRIMAP或背景进行建模训练的最新方法相比，CRGNN借助未标记的真实训练数据集生成了令人满意的结果。

Rolling shutter (RS) images can be viewed as the result of the row-wise combination of global shutter (GS) images captured by a virtual moving GS camera over the period of camera readout time. The RS effect brings tremendous difficulties for the downstream applications. In this paper, we propose to invert the above RS imaging mechanism, i.e., recovering a high framerate GS video from consecutive RS images to achieve RS temporal super-resolution (RSSR). This extremely challenging problem, e.g., recovering 1440 GS images from two 720-height RS images, is far from being solved end-to-end. To address this challenge, we exploit the geometric constraint in the RS camera model, thus achieving geometry-aware inversion. Specifically, we make three contributions in resolving the above difficulties: (i) formulating the bidirectional RS undistortion flows under the constant velocity motion model, (ii) building the connection between the RS undistortion flow and optical flow via a scaling operation, and (iii) developing a mutual conversion scheme between varying RS undistortion flows that correspond to different scanlines. Building upon these formulations, we propose the first RS temporal super-resolution network in a cascaded structure to extract high framerate global shutter video. Our method explores the underlying spatio-temporal geometric relationships within a deep learning framework, where no extra supervision besides the middle-scanline ground truth GS image is needed. Essentially, our method can be very efficient for explicit propagation to generate GS images under any scanline. Experimental results on both synthetic and real data show that our method can produce high-quality GS image sequences with rich details, outperforming state-of-the-art methods.

滚动快门 (RS) 图像可以被视为在相机读出时间段内由虚拟移动的GS相机捕获的全局快门 (GS) 图像的行组合的结果。RS效应给下游应用带来了极大的困难。在本文中，我们建议反转上述RS成像机制，即从连续RS图像恢复高帧率GS视频，以实现RS时间超分辨率 (RSSR)。这一极具挑战性的问题，例如，从两幅720高的遥感图像中恢复1440 GS图像，远未得到端到端的解决。为了应对这一挑战，我们利用遥感相机模型中的几何约束，从而实现几何感知反演。具体来说，我们在解决上述困难方面做出了三个贡献：(i) 在恒定速度运动模型下建立双向RS不失真流，(ii) 通过缩放操作建立RS不失真流和光流之间的连接，和 (iii) 在对应于不同扫描线的不同RS不失真流之间开发相互转换方案。在这些公式的基础上，我们提出了第一个级联结构的RS时间超分辨率网络来提取高帧率全局快门视频。我们的方法在深度学习框架内探索潜在的时空几何关系，在深度学习框架中，除了中间扫描线地面真实GS图像之外，不需要额外的监督。本质上，我们的方法可以非常有效地进行显式传播，以生成任意扫描线下的GS图像。在合成数据和真实数据上的实验结果表明，我们的方法可以生成具有丰富细节的高质量GS图像序列，优于现有的方法。

Self-supervised detection and segmentation of foreground objects aims for accuracy without annotated training data. However, existing approaches predominantly rely on restrictive assumptions on appearance and motion. For scenes with dynamic activities and camera motion, we propose a multi-camera framework in which geometric constraints are embedded in the form of multi-view consistency during training via coarse 3D localization in a voxel grid and fine-grained offset regression. In this manner, we learn a joint distribution of proposals over multiple views. At inference time, our method operates on single RGB images. We outperform state-of-the-art techniques both on images that visually depart from those of standard benchmarks and on those of the classical Human3.6M dataset.

前景目标的自监督检测和分割旨在提高精度，而无需标注训练数据。然而，现有的方法主要依赖于对外观和运动的限制性假设。对于具有动态活动和摄像机运动的场景，我们提出了一种多摄像机框架，其中通过体素网格中的粗略三维定位和细粒度偏移回归，在训练过程中以多视图一致性的形式嵌入几何约束。通过这种方式，我们了解到提案在多个视图上的联合分布。在推断时，我们的方法对单个RGB图像

进行操作。在视觉上偏离标准基准的图像和经典Human3的图像上，我们的表现都优于最先进的技术。6M数据集。

Model compression techniques are recently gaining explosive attention for obtaining efficient AI models for various real time applications. Channel pruning is one important compression strategy, and widely used in slimming various DNNs. Previous gate-based or importance-based pruning methods aim to remove channels whose "importance" are smallest. However, it remains unclear what criteria the channel importance should be measured on, leading to various channel selection heuristics. Some other sampling-based pruning methods deploy sampling strategy to train sub-nets, which often causes the training instability and the compressed model's degraded performance. In view of the research gaps, we present a new module named Gates with Differentiable Polarization (GDP), inspired by principled optimization ideas. GDP can be plugged before convolutional layers without bells and whistles, to control the on-and-off of each channel or whole layer block. During the training process, the polarization effect will drive a subset of gates to smoothly decrease to exactly zero, while other gates gradually stay away from zero by a large margin. When training terminates, those zero-gated channels can be painlessly removed, while other non-zero gates can be absorbed into the succeeding convolution kernel, causing completely no interruption to training nor damage to the trained model. Experiments conducted over CIFAR-10 and ImageNet datasets show that the proposed GDP algorithm achieves the state-of-the-art performance on various benchmark DNNs at a broad range of pruning ratios. We also apply GDP to DeepLabV3Plus-ResNet50 on the challenging Pascal VOC segmentation task, whose test performance sees no drop (even slightly improved) with over 60% FLOPS saving.

模型压缩技术最近获得了爆炸性的关注，以获得各种实时应用的有效人工智能模型。通道剪枝是一种重要的压缩策略，广泛应用于各种DNN的瘦身。以前基于门或基于重要性的修剪方法旨在删除“重要性”最小的通道。然而，目前尚不清楚衡量渠道重要性的标准，导致了各种渠道选择启发式。其他一些基于采样的剪枝方法采用采样策略来训练子网，这往往会导致训练的不稳定性和压缩模型的性能下降。鉴于研究的空白，我们在原则优化思想的启发下，提出了一个新的模块，名为可微极化门（GDP）。GDP可以在没有钟声和哨声的卷积层之前插入，以控制每个通道或整个层块的开启和关闭。在训练过程中，极化效应将促使一部分门平滑地减小到零，而其他门则逐渐远离零。当训练终止时，这些零门通道可以被无痛地移除，而其他非零门可以被吸收到后续的卷积核中，从而完全不会中断训练，也不会损坏训练模型。在CIFAR-10和ImageNet数据集上进行的实验表明，所提出的GDP算法在各种基准DNN上以广泛的剪枝率实现了最先进的性能。我们还将GDP应用于具有挑战性的Pascal VOC分段任务中的DeepLabV3Plus-ResNet50，该任务的测试性能没有下降（甚至略有改善），节省了60%以上的触发器。

The popularity of multimodal sensors and the accessibility of the Internet have brought us a massive amount of unlabeled multimodal data. Since existing datasets and well-trained models are primarily unimodal, the modality gap between a unimodal network and unlabeled multimodal data poses an interesting problem: how to transfer a pre-trained unimodal network to perform the same task on unlabeled multimodal data? In this work, we propose multimodal knowledge expansion (MKE), a knowledge distillation-based framework to effectively utilize multimodal data without requiring labels. Opposite to traditional knowledge distillation, where the student is designed to be lightweight and inferior to the teacher, we observe that the multimodal student model consistently rectifies pseudo labels and generalizes better than its teacher. Extensive experiments on four tasks and different modalities verify this finding. Furthermore, we connect the mechanism of MKE to semi-supervised learning and offer both empirical and theoretical explanations to understand the expansion capability of a multimodal student.

多模传感器的普及和互联网的可访问性为我们带来了大量未标记的多模数据。由于现有的数据集和经过良好训练的模型主要是单峰数据，单峰网络和未标记的多峰数据之间的模态差异提出了一个有趣的问题：如何传输预先训练的单峰网络，以便对未标记的多峰数据执行相同任务？在这项工作中，我们提出了多模态知识扩展（MKE），这是一个基于知识提取的框架，可以在不需要标签的情况下有效地利用多模态数据。与传统的知识提炼相反，在传统知识提炼中，学生被设计成轻量级的，不如老师，我们观察到，多模态学生模型始终纠正伪标签，并且比老师概括得更好。对四项任务和不同方式的广泛实验验证了这一发现。此外，我们将MKE机制与半监督学习联系起来，并提供实证和理论解释，以了解多模态学生的扩展能力。

Bundle adjustment (BA) occupies a large portion of SfM and visual SLAM's total execution time. Local BA over the latest several keyframes plays a crucial role in visual SLAM. Its execution time should be sufficiently short for robust tracking; this is especially critical for embedded systems with a limited computational resource. This study proposes a learning-based method using a graph network that can replace conventional optimization-based BA and works faster. The graph network operates on a graph consisting of the nodes of keyframes and landmarks and the edges of the latter's visibility from the former. The graph network receives the parameters' initial values as inputs and predicts the updates to their optimal values. We design an intermediate representation of inputs inspired by the normal equation of the Levenberg-Marquardt method. We use the sum of reprojection errors as a loss function to train the graph network. The experiments show that the proposed method outputs parameter estimates with slightly inferior accuracy in 1/60-1/10 of time compared with the conventional BA.

束调整（BA）占据SfM和visual SLAM总执行时间的很大一部分。最近几个关键帧上的局部BA在视觉SLAM中起着至关重要的作用。其执行时间应足够短，以便进行鲁棒跟踪；这对于计算资源有限的嵌入式系统尤其重要。本研究提出一种基于学习的方法，使用图形网络，可以取代传统的基于优化的BA，并且工作速度更快。图形网络在一个图形上运行，该图形由关键帧和地标的节点以及后者相对于前者的可见性的边组成。图形网络接收参数的初始值作为输入，并预测其最佳值的更新。受Levenberg-Marquardt方法的正规方程启发，我们设计了输入的中间表示。我们使用重投影误差之和作为损失函数来训练图形网络。实验表明，与传统的BA相比，该方法在1/60-1/10的时间内输出的参数估计精度稍差。

Many objects do not appear frequently enough in complex scenes (e.g., certain handbags in living rooms) for training an accurate object detector, but are often found frequently by themselves (e.g., in product images). Yet, these object-centric images are not effectively leveraged for improving object detection in scene-centric images. In this paper, we propose Mosaic of Object-centric images as Scene-centric images (Mosaicos), a simple and novel framework that is surprisingly effective at tackling the challenges of long-tailed object detection. Keys to our approach are three-fold: (i) pseudo scene-centric image construction from object-centric images for mitigating domain differences, (ii) high-quality bounding box imputation using the object-centric images' class labels, and (iii) a multi-stage training procedure. On LVIS object detection (and instance segmentation), Mosaicos leads to a massive 60% (and 23%) relative improvement in average precision for rare object categories. We also show that our framework can be compatibly used with other existing approaches to achieve even further gains. Our pre-trained models are publicly available at <https://github.com/czhang0528/Mosaicos/>.

许多物体在复杂场景（例如，客厅中的某些手提包）中出现的频率不足以训练精确的物体检测器，但它们往往是自己经常发现的（例如，在产品图像中）。然而，这些以对象为中心的图像并没有有效地用于改进以场景为中心的图像中的对象检测。在本文中，我们提出了以对象为中心的图像拼接作为以场景为中心的图像（Mosaicos），这是一个简单而新颖的框架，在解决长尾目标检测的挑战方面非常有效。我

们的方法的关键有三个方面：(i) 从以对象为中心的图像构建伪场景为中心的图像，以减少域差异；(ii) 使用以对象为中心的图像的类标签进行高质量边界盒插补；(iii) 多阶段训练过程。在LVIS对象检测（和实例分割）方面，MosaicOS使稀有对象类别的平均精度相对提高了60%（和23%）。我们还表明，我们的框架可以与其他现有方法兼容使用，以实现进一步的收益。我们预先培训的模型可在<https://github.com/czhang0528/MosaicOS/>。

Recently, video deblurring has attracted considerable research attention, and several works suggest that events at high time rate can benefit deblurring. In this paper, we develop a principled framework D2Nets for video deblurring to exploit non-consecutively blurry frames, and propose a flexible event fusion module (EFM) to bridge the gap between event-driven and video deblurring. In D2Nets, we propose to first detect nearest sharp frames (NSFs) using a bidirectional LSTM detector, and then perform deblurring guided by NSFs. Furthermore, the proposed EFM is flexible to be incorporated into D2Nets, in which events can be leveraged to notably boost the deblurring performance. EFM can also be easily incorporated into existing deblurring networks, making event-driven deblurring task benefit from state-of-the-art deblurring methods. On synthetic and real-world blurry datasets, our methods achieve better results than competing methods, and EFM not only benefits D2Nets but also significantly improves the competing deblurring networks.

最近，视频去模糊引起了相当多的研究关注，一些研究表明，高时间速率的事件有利于去模糊。在本文中，我们开发了一个用于视频去模糊的原则框架D2NET，以利用非连续模糊帧，并提出了一个灵活的事件融合模块（EFM）来弥补事件驱动和视频去模糊之间的差距。在D2NET中，我们建议首先使用双向LSTM检测器检测最近尖锐帧（NSF），然后在NSF的引导下执行去模糊。此外，建议的EFM可以灵活地并入D2NET，在D2NET中可以利用事件显著提高去模糊性能。EFM也可以很容易地并入现有的去模糊网络，使事件驱动的去模糊任务受益于最先进的去模糊方法。在合成的和真实的模糊数据集上，我们的方法比竞争方法取得更好的结果，EFM不仅有利于D2NET，而且还显著改善了竞争的去模糊网络。

To meet the space limitation of optical elements, free-form surfaces or high-order aspherical lenses are adopted in mobile cameras to compress volume. However, the application of free-form surfaces also introduces the problem of image quality mutation. Existing model-based deconvolution methods are inefficient in dealing with the degradation that shows a wide range of spatial variants over regions. And the deep learning techniques in low-level and physics-based vision suffer from a lack of accurate data. To address this issue, we develop a degradation framework to estimate the spatially variant point spread functions (PSFs) of mobile cameras. When input extreme-quality digital images, the proposed framework generates degraded images sharing a common domain with real-world photographs. Supplied with the synthetic image pairs, we design a Field-of-view shared kernel prediction network (FOV-KPN) to perform spatial-adaptive reconstruction on real degraded photos. Extensive experiments demonstrate that the proposed approach achieves extreme-quality computational imaging and outperforms the state-of-the-art methods. Furthermore, we illustrate that our technique can be integrated into existing postprocessing systems, resulting in significantly improved visual quality.

为了满足光学元件的空间限制，移动相机采用自由曲面或高阶非球面透镜来压缩体积。然而，自由曲面的应用也带来了图像质量突变的问题。现有的基于模型的反褶积方法在处理退化方面效率低下，这种退化显示出区域间广泛的空间变化。低水平和基于物理的视觉深度学习技术缺乏准确的数据。为了解决这个问题，我们开发了一个退化框架来估计移动摄像机的空间变化点扩展函数（PSF）。当输入高质量的数字图像时，该框架生成与真实世界照片共享一个公共域的退化图像。在合成图像对的支持下，我们设计了一个视场共享核预测网络（FOV-KPN）对真实退化照片进行空间自适应重建。大量实验表明，该方

法实现了高质量的计算成像，并优于现有的方法。此外，我们还说明，我们的技术可以集成到现有的后处理系统中，从而显著提高视觉质量。

In unconstrained real-world surveillance scenarios, person re-identification (Re-ID) models usually suffer from different low-level perceptual variations, e.g., cross-resolution and insufficient lighting. Due to the limited variation range of training data, existing models are difficult to generalize to scenes with unknown perceptual interference types. To address the above problem, in this paper, we propose two disjoint data-generation ways to complement existing training samples to improve the robustness of Re-ID models. Firstly, considering the sparsity and imbalance of samples in the perceptual space, a dense resampling method from the estimated perceptual distribution is performed. Secondly, to dig more representative generated samples for identity representation learning, we introduce a graph-based white-box attacker to guide the data generation process with intra-batch ranking and discriminative attention. In addition, two synthetic-to-real feature constraints are introduced into the Re-ID training to prevent the generated data from bringing domain bias. Our method is effective, easy-to-implement, and independent of the specific network architecture. Applying our approach to a ResNet-50 baseline can already achieve competitive results, surpassing state-of-the-art methods by +1.2% at Rank-1 on the MLR-CUHK03 dataset.

在无约束的现实监控场景中，人员重新识别（re-ID）模型通常会遇到不同的低级感知变化，例如交叉分辨率和照明不足。由于训练数据变化范围有限，现有模型难以推广到感知干扰类型未知的场景。为了解决上述问题，本文提出了两种不相交的数据生成方法来补充现有的训练样本，以提高Re-ID模型的鲁棒性。首先，考虑到感知空间中样本的稀疏性和不平衡性，从估计的感知分布中进行稠密重采样。其次，为了挖掘更具代表性的生成样本进行身份表示学习，我们引入了一种基于图的白盒攻击者，通过批内排序和区分注意来指导数据生成过程。此外，在Re-ID训练中引入了两个合成到真实的特征约束，以防止生成的数据带来域偏差。我们的方法是有效的，易于实现，并且独立于特定的网络架构。将我们的方法应用于ResNet-50基线已经可以取得有竞争力的结果，在MLR-CUHK03数据集排名第一的情况下超过最先进的方法+1.2%。

Curvilinear structure segmentation (CSS) is under semantic segmentation, whose applications include crack detection, aerial road extraction, and biomedical image segmentation. In general, geometric topology and pixel-wise features are two critical aspects of CSS. However, most semantic segmentation methods only focus on enhancing feature representations while existing CSS techniques emphasize preserving topology alone. In this paper, we present a Joint Topology-preserving and Feature-refinement Network (JTFN) that jointly models global topology and refined features based on an iterative feedback learning strategy. Specifically, we explore the structure of objects to help preserve corresponding topologies of predicted masks, thus design a reciprocal two-stream module for CSS and boundary detection. In addition, we introduce such topology-aware predictions as feedback guidance that refines attentive features by supplementing and enhancing saliences. To the best of our knowledge, this is the first work that jointly addresses topology preserving and feature refinement for CSS. We evaluate JTFN on four datasets of diverse applications: Crack500, CrackTree200, Roads, and DRIVE. Results show that JTFN performs best in comparison with alternative methods. Code is available.

曲线结构分割（CSS）属于语义分割，其应用包括裂纹检测、空中道路提取和生物医学图像分割。一般来说，几何拓扑和像素特征是CSS的两个关键方面。然而，大多数语义分割方法只关注增强特征表示，而现有的CSS技术只强调保持拓扑结构。在本文中，我们提出了一种联合拓扑保持和特征细化网络（JTFN），该网络基于迭代反馈学习策略联合建模全局拓扑和细化特征。具体来说，我们探索对象的结构以帮助保留预测掩码的相应拓扑，从而设计用于CSS和边界检测的双向流模块。此外，我们还引入了

拓扑感知预测作为反馈指导，通过补充和增强显著性来细化关注特征。据我们所知，这是第一个联合处理CSS拓扑保持和特征细化的工作。我们在不同应用的四个数据集上评估JTFN：Crack500、CrackTree200、Roads和DRIVE。结果表明，与其他方法相比，JTFN的性能最好。代码是可用的。

we introduce the task of weakly supervised learning for detecting human and object interactions in videos. Our task poses unique challenges as a system does not know what types of human-object interactions are present in a video or the actual spatiotemporal location of the human and object. To address these challenges, we introduce a contrastive weakly supervised training loss that aims to jointly associate spatiotemporal regions in a video with an action and object vocabulary and encourage temporal continuity of the visual appearance of moving objects as a form of self-supervision. To train our model, we introduce a dataset comprising over 6.5k videos with human-object interaction annotations that have been semi-automatically curated from sentence captions associated with the videos. We demonstrate improved performance over weakly supervised baselines adapted to our task on our video dataset.

我们介绍了弱监督学习的任务，用于检测视频中的人与对象交互。我们的任务带来了独特的挑战，因为系统不知道视频中存在什么类型的人-物交互，也不知道人和物的实际时空位置。为了应对这些挑战，我们引入了一种对比弱监督训练损失，旨在将视频中的时空区域与动作和对象词汇联合起来，并鼓励运动对象视觉外观的时间连续性，作为自我监督的一种形式。为了训练我们的模型，我们引入了一个数据集，该数据集包含超过6.5k个视频，这些视频带有人机交互注释，这些注释是由与视频相关的句子标题半自动策划的。我们在视频数据集上展示了与我们的任务相适应的弱监督基线相比的改进性能。

We propose a new generative model for layout generation. We generate layouts in three steps. First, we generate the layout elements as nodes in a layout graph. Second, we compute constraints between layout elements as edges in the layout graph. Third, we solve for the final layout using constrained optimization. For the first two steps, we build on recent transformer architectures. The layout optimization implements the constraints efficiently. We show three practical contributions compared to the state of the art: our work requires no user input, produces higher quality layouts, and enables many novel capabilities for conditional layout generation.

我们提出了一种新的布局生成模型。我们分三步生成布局。首先，我们将布局元素生成为布局图中的节点。其次，我们将布局元素之间的约束计算为布局图中的边。第三，我们使用约束优化方法求解最终布局。对于前两个步骤，我们以最新的transformer架构为基础。布局优化有效地实现了约束。与最新技术相比，我们展示了三个实际贡献：我们的工作不需要用户输入，生成更高质量的布局，并支持许多新的条件布局生成功能。

Image-based virtual try-on involves synthesizing perceptually convincing images of a model wearing a particular garment and has garnered significant research interest due to its immense practical applicability. Recent methods involve a two-stage process: i) warping of the garment to align with the model ii) texture fusion of the warped garment and target model to generate the try-on output. Issues arise due to the non-rigid nature of garments and the lack of geometric information about the model or the garment. It often results in improper rendering of granular details. We propose ZFlow, an end-to-end framework, which seeks to alleviate these concerns regarding geometric and textural integrity (such as pose, depth-ordering, skin and neckline reproduction) through a combination of gated aggregation of hierarchical flow estimates termed Gated Appearance Flow, and dense structural priors at various stages of the network. ZFlow achieves state-of-the-art results as observed qualitatively, and on benchmark image quality measures (PSNR, SSIM, and FID scores). The paper also presents extensive comparisons with existing state-of-the-art including a detailed user study and ablation studies to gauge the effectiveness of each of our contributions on multiple datasets

基于图像的虚拟试穿涉及合成穿着特定服装的模特的具有感知说服力的图像，并且由于其巨大的实用性而引起了重大的研究兴趣。最近的方法涉及两个阶段：i) 服装翘曲以与模型对齐ii) 翘曲服装和目标模型的纹理融合以生成试穿输出。由于服装的非刚性以及缺乏关于模型或服装的几何信息，会出现问题。它通常会导致不正确地呈现颗粒细节。我们提出了ZFlow，一种端到端的框架，旨在通过称为门控外观流的分层流估计的门控聚集组合，缓解对几何和纹理完整性（如姿势、深度排序、皮肤和领口再现）的担忧，在网络的各个阶段都有密集的结构先验。ZFlow通过定性观察和基准图像质量度量（PSNR、SSIM和FID分数）获得最先进的结果。本文还与现有的最新技术进行了广泛的比较，包括详细的用户研究和消融研究，以评估我们在多个数据集上的每个贡献的有效性

Internet video delivery has undergone a tremendous explosion of growth over the past few years. However, the quality of video delivery system greatly depends on the Internet bandwidth. Deep Neural Networks (DNNs) are utilized to improve the quality of video delivery recently. These methods divide a video into chunks, and stream LR video chunks and corresponding content-aware models to the client. The client runs the inference of models to super-resolve the LR chunks. Consequently, a large number of models are streamed in order to deliver a video. In this paper, we first carefully study the relation between models of different chunks, then we tactfully design a joint training framework along with the Content-aware Feature Modulation (CaFM) layer to compress these models for neural video delivery. With our method, each video chunk only requires less than 1% of original parameters to be streamed, achieving even better SR performance. We conduct extensive experiments across various SR backbones, video time length, and scaling factors to demonstrate the advantages of our method. Besides, our method can be also viewed as a new approach of video coding. Our primary experiments achieve better video quality compared with the commercial H.264 and H.265 standard under the same storage cost, showing the great potential of the proposed method. Code is available at: <https://github.com/Neural-video-delivery/CaFM-Pytorch-ICCV2021>

互联网视频传输在过去几年经历了巨大的爆炸式增长。然而，视频传输系统的质量在很大程度上取决于互联网带宽。近年来，人们利用深度神经网络（DNN）来提高视频传输质量。这些方法将视频划分为块，并将LR视频块和相应的内容感知模型流到客户端。客户机运行模型推理来超级解析LR块。因此，为了传输视频，大量模型被流式传输。在本文中，我们首先仔细研究了不同块的模型之间的关系，然后巧妙地设计了一个联合训练框架以及内容感知特征调制（CaFM）层来压缩这些模型用于神经视频传输。通过我们的方法，每个视频块只需要不到原始参数的1%的数据流，从而实现更好的SR性能。我们在各种SR主干、视频时间长度和缩放因子上进行了广泛的实验，以证明我们的方法的优势。此外，我们的方法也可以看作是一种新的视频编码方法。我们的初步实验在相同的存储成本下实现了比商用H.264和H.265

标准更好的视频质量，显示了该方法的巨大潜力。代码可从以下网址获取：<https://github.com/Neural-video-delivery/CaFM-Pytorch-ICCV2021>

Vision-and-language (V&L) reasoning necessitates perception of visual concepts such as objects and actions, understanding semantics and language grounding, and reasoning about the interplay between the two modalities. One crucial aspect of visual reasoning is spatial understanding, which involves understanding relative locations of objects, i.e. implicitly learning the geometry of the scene. In this work, we evaluate the faithfulness of V&L models to such geometric understanding, by formulating the prediction of pair-wise relative locations of objects as a classification as well as a regression task. Our findings suggest that state-of-the-art transformer-based V&L models lack sufficient abilities to excel at this task. Motivated by this, we design two objectives as proxies for 3D spatial reasoning (SR) -- object centroid estimation, and relative position estimation, and train V&L with weak supervision from off-the-shelf depth estimators. This leads to considerable improvements in accuracy for the "GQA" visual question answering challenge (in fully supervised, few-shot, and O.O.D settings) as well as improvements in relative spatial reasoning. Code and data will be released here.

视觉和语言（V&L）推理需要感知视觉概念，如物体和动作，理解语义和语言基础，并对两种模式之间的相互作用进行推理。视觉推理的一个关键方面是空间理解，它涉及到理解对象的相对位置，即隐含地学习场景的几何结构。在这项工作中，我们评估了V&L模型对此类几何理解的忠实性，将对象成对相对位置的预测作为分类和回归任务。我们的研究结果表明，最先进的基于变压器的V&L模型缺乏足够的能力来胜任这项任务。基于此，我们设计了两个目标作为3D空间推断（SR）的代理——对象质心估计和相对位置估计，并在现有深度估计器的弱监督下训练V&L。这将大大提高“GQA”视觉问答挑战的准确性（在完全监督、少镜头和O.O.D设置中）以及相对空间推断的改进。代码和数据将在此处发布。

Image matting refers to the estimation of the opacity of foreground objects. It requires correct contours and fine details of foreground objects for the matting results. To better accomplish human image matting tasks, we propose the Cascade Image Matting Network with Deformable Graph Refinement (CasDGR), which can automatically predict precise alpha mattes from single human images without any additional inputs. We adopt a network cascade architecture to perform matting from low-to-high resolution, which corresponds to coarse-to-fine optimization. We also introduce the Deformable Graph Refinement (DGR) module based on graph neural networks (GNNs) to overcome the limitations of convolutional neural networks (CNNs). The DGR module can effectively capture long-range relations and obtain more global and local information to help produce finer alpha mattes. We also reduce the computation complexity of the DGR module by dynamically predicting the neighbors and apply DGR module to higher-resolution features. Experimental results demonstrate the ability of our CasDGR to achieve state-of-the-art performance on synthetic datasets and produce good results on real human images.

图像抠图是指对前景对象不透明度的估计。它需要正确的轮廓和前景对象的精细细节，以获得消光效果。为了更好地完成人体图像抠图任务，我们提出了带有变形图细化的级联图像抠图网络（CasDGR），该网络可以从单个人体图像中自动预测精确的alpha抠图，而无需任何额外输入。我们采用网络级联结构从低分辨率到高分辨率执行抠图，这对应于从粗到精的优化。为了克服卷积神经网络（CNN）的局限性，我们还引入了基于图神经网络（GNN）的变形图细化（DGR）模块。DGR模块可以有效地捕获远程关系，并获得更多全局和局部信息，以帮助生成更精细的阿尔法蒙版。我们还通过动态预测邻域来降低DGR模块的计算复杂度，并将DGR模块应用于更高分辨率的特征。实验结果表明，我们的CasDGR能够在合成数据集上实现最先进的性能，并在真实人体图像上产生良好的效果。

Preserving maximal information is the basic principle of designing self-supervised learning methodologies. To reach this goal, contrastive learning adopts an implicit way which is contrasting image pairs. However, we believe it is not fully optimal to simply use the contrastive estimation for preservation. Moreover, it is necessary and complementary to introduce an explicit solution to preserve more information. From this perspective, we introduce Preservational Learning to reconstruct diverse image contexts in order to preserve more information in learned representations. Together with the contrastive loss, we present Preservational Contrastive Representation Learning (PCRL) for learning self-supervised medical representations. PCRL provides very competitive results under the pretraining-finetuning protocol, outperforming both self-supervised and supervised counterparts in 5 classification/segmentation tasks substantially.

保存最大信息是设计自监督学习方法的基本原则。为了达到这个目的，对比学习采用了一种内隐的方式，即对比图像对。然而，我们认为，仅仅使用对比评估进行保存并非完全最优。此外，引入一个明确的解决方案来保存更多的信息是必要的和补充的。从这个角度出发，我们引入保留学习来重建不同的图像上下文，以便在学习的表征中保留更多的信息。结合对比损失，我们提出了用于学习自我监督医学表征的保留对比表征学习（PCRL）。PCRL在预训练微调协议下提供了非常有竞争力的结果，在5项分类/分割任务中大大优于自我监督和监督的结果。

Localizing individuals in crowds is more in accordance with the practical demands of subsequent high-level crowd analysis tasks than simply counting. However, existing localization based methods relying on intermediate representations (i.e., density maps or pseudo boxes) serving as learning targets are counter-intuitive and error-prone. In this paper, we propose a purely point-based framework for joint crowd counting and individual localization. For this framework, instead of merely reporting the absolute counting error at image level, we propose a new metric, called density Normalized Average Precision (nAP), to provide more comprehensive and more precise performance evaluation. Moreover, we design an intuitive solution under this framework, which is called Point to Point Network (P2PNet). P2PNet discards superfluous steps and directly predicts a set of point proposals to represent heads in an image, being consistent with the human annotation results. By thorough analysis, we reveal the key step towards implementing such a novel idea is to assign optimal learning targets for these proposals. Therefore, we propose to conduct this crucial association in an one-to-one matching manner using the Hungarian algorithm. The P2PNet not only significantly surpasses state-of-the-art methods on popular counting benchmarks, but also achieves promising localization accuracy. The codes will be available at: <https://github.com/TencentYoutuResearch/CrowdCounting-P2PNet>.

在人群中定位个体比简单地计数更符合后续高级人群分析任务的实际需求。然而，现有的基于定位的方法依赖于作为学习目标的中间表示（即密度图或伪框），这是违反直觉和容易出错的。在本文中，我们提出了一个纯粹基于点的联合人群计数和个体定位框架。对于该框架，我们提出了一种新的度量，称为密度归一化平均精度（nAP），以提供更全面和更精确的性能评估，而不是仅报告图像级的绝对计数误差。此外，我们在此框架下设计了一个直观的解决方案，称为点对点网络（P2PNet）。P2PNet摒弃了多余的步骤，直接预测一组点建议来表示图像中的头部，与人类注释结果一致。通过深入的分析，我们发现实现这种新想法的关键步骤是为这些建议分配最佳的学习目标。因此，我们建议使用匈牙利算法以一对一的匹配方式进行这一关键关联。P2PNet不仅在流行的计数基准上大大超过了最先进的方法，而且还实现了很好的定位精度。代码将在以下位置提供：<https://github.com/TencentYoutuResearch/CrowdCounting-P2PNet>。

Turn-taking has played an essential role in structuring the regulation of a conversation. The task of identifying the main speaker (who is properly taking his/her turn of speaking) and the interrupters (who are interrupting or reacting to the main speaker's utterances) remains a challenging task. Although some prior methods have partially addressed this task, there still remain some limitations. Firstly, a direct association of Audio and Visual features may limit the correlations to be extracted due to different modalities. Secondly, the relationship across temporal segments helping to maintain the consistency of localization, separation and conversation contexts is not effectively exploited. Finally, the interactions between speakers that usually contain the tracking and anticipatory decisions about transition to a new speaker is usually ignored. Therefore, this work introduces a new Audio-Visual Transformer approach to the problem of localization and highlighting the main speaker in both audio and visual channels of a multi-speaker conversation video in the wild. The proposed method exploits different types of correlations presented in both visual and audio signals. The temporal audio-visual relationships across spatial-temporal space are anticipated and optimized via the self-attention mechanism in a Transformer structure. Moreover, a newly collected dataset is introduced for the main speaker detection. To the best of our knowledge, it is one of the first studies that is able to automatically localize and highlight the main speaker in both visual and audio channels in multi-speaker conversation videos.

话轮转换在构建会话规则中起着至关重要的作用。识别主要演讲者（正确轮到他/她发言）和打断者（打断或回应主要演讲者的话语）的任务仍然是一项具有挑战性的任务。尽管以前的一些方法已经部分解决了这一任务，但仍然存在一些局限性。首先，由于模式不同，音频和视频特征的直接关联可能会限制要提取的相关性。其次，跨时间段的关系有助于保持本地化、分离和会话上下文的一致性，但没有得到有效利用。最后，说话人之间的互动通常被忽略，这些互动通常包含关于向新说话人过渡的跟踪和预期决策。因此，这项工作引入了一种新的视听转换器方法来解决野外多人对话视频中的定位问题，并在音频和视频通道中突出主说话人。该方法利用了视觉和音频信号中呈现的不同类型的相关性。通过变压器结构中的自我注意机制，预测并优化跨时空的时间视听关系。此外，还引入了一个新收集的数据集用于主说话人检测。据我们所知，这是第一个能够在多人对话视频的视觉和音频通道中自动定位和突出显示主讲人的研究。

Most existing convolution neural network (CNN) based super-resolution (SR) methods generate their paired training dataset by artificially synthesizing low-resolution (LR) images from the high-resolution (HR) ones. However, this dataset preparation strategy harms the application of these CNNs in real-world scenarios due to the inherent domain gap between the training and testing data. A popular attempts towards the challenge is unpaired generative adversarial networks, which generate "real" LR counterparts from real HR images using image-to-image translation and then perform super-resolution from "real" LR->SR. Despite great progress, it is still difficult to synthesize perfect "real" LR images for super-resolution. In this paper, we firstly consider the real-world SR problem from the traditional domain adaptation perspective. We propose a novel unpaired SR training framework based on feature distribution alignment, with which we can obtain degradation-indistinguishable feature maps and then map them to HR images. In order to generate better SR images for target LR domain, we introduce several regularization losses to force the aligned feature to locate around the target domain. Our experiments indicate that our SR network obtains the state-of-the-art performance over both blind and unpaired SR methods on diverse datasets.

现有的大多数基于卷积神经网络 (CNN) 的超分辨率 (SR) 方法都是通过人工合成高分辨率 (HR) 图像中的低分辨率 (LR) 图像来生成成对的训练数据集。然而，由于训练数据和测试数据之间固有的领域差距，这种数据集准备策略损害了这些CNN在现实场景中的应用。针对这一挑战的一种流行尝试是不成对的生成性对抗网络，该网络使用图像到图像的转换从真实HR图像生成“真实”LR对应物，然后从“真实”

LR->SR执行超分辨率。尽管取得了很大进展，但仍然难以合成完美的“真实”LR图像以实现超分辨率。在本文中，我们首先考虑现实世界的SR问题从传统的域适应的角度。我们提出了一种新的基于特征分布对齐的非成对SR训练框架，利用该框架可以获得退化不可区分的特征映射，然后将其映射到HR图像。为了为目标LR域生成更好的SR图像，我们引入了一些正则化损失，以强制对齐特征定位在目标域周围。我们的实验表明，在不同的数据集上，我们的SR网络在盲SR方法和非配对SR方法上都获得了最先进的性能。

Black-box attacks aim to generate adversarial noise to fail the victim deep neural network in the black box. The central task in black-box attack method design is to estimate and characterize the victim model in the high-dimensional model space based on feedback results of queries submitted to the victim network. The central performance goal is to minimize the number of queries needed for successful attack. Existing attack methods directly search and refine the adversarial noise in an extremely high-dimensional space, requiring hundreds or even thousands queries to the victim network. To address this challenge, we propose to explore a consistency and sensitivity guided ensemble attack (CSEA) method in a low-dimensional space. Specifically, we estimate the victim model in the black box using a learned linear composition of an ensemble of surrogate models with diversified network structures. Using random block masks on the input image, these surrogate models jointly construct and submit randomized and sparsified queries to the victim model. Based on these query results and guided by a consistency constraint, the surrogate models can be trained using a very small number of queries such that their learned composition is able to accurately approximate the victim model in the high-dimensional space. The randomized and sparsified queries also provide important information for us to construct an attack sensitivity map for the input image, with which the adversarial attack can be locally refined to further increase its success rate. Our extensive experimental results demonstrate that our proposed approach significantly reduces the number of queries to the victim network while maintaining very high success rates, outperforming existing black-box attack methods by large margins.

黑盒攻击的目的是产生对抗性噪声，将受害者的深层神经网络置于黑盒中。黑箱攻击方法设计的核心任务是根据向受害者网络提交的查询的反馈结果，在高维模型空间中估计和描述受害者模型。中心性能目标是使成功的attack所需的查询数量最小化。现有的攻击方法直接在极高维空间中搜索和细化广告噪声，需要对受害者网络进行数百甚至数千次查询。为了应对这一挑战，我们建议在低维空间中探索一种一致性和敏感性引导的集成攻击（CSEA）方法。具体地说，我们在黑盒中使用具有多种网络结构的代理模型集合的学习线性组合来估计受害者模型。使用输入图像上的随机块掩码，这些代理模型共同构造并向victim model提交随机和稀疏查询。基于这些查询结果并在一致性约束的指导下，可以使用非常少的查询来训练代理模型，从而使其所学习的合成能够在高维空间中准确地近似victim model。随机和稀疏查询也为我们构建输入图像的攻击敏感度图提供了重要信息，利用该图可以局部细化对抗攻击，进一步提高其成功率。我们的大量实验结果表明，我们提出的方法显著减少了对受害者网络的查询数量，同时保持了很高的成功率，大大优于现有的黑盒攻击方法。

Transformers with powerful global relation modeling abilities have been introduced to fundamental computer vision tasks recently. As a typical example, the Vision Transformer (ViT) directly applies a pure transformer architecture on image classification, by simply splitting images into tokens with a fixed length, and employing transformers to learn relations between these tokens. However, such naive tokenization could destruct object structures, assign grids to uninterested regions such as background, and introduce interference signals. To mitigate the above issues, in this paper, we propose an iterative and progressive sampling strategy to locate discriminative regions. At each iteration, embeddings of the current sampling step are fed into a transformer encoder layer, and a group of sampling offsets is predicted to update the sampling locations for the next step. The progressive sampling is differentiable. When combined with the Vision Transformer, the obtained PS-ViT network can adaptively learn where to look. The proposed PS-ViT is both effective and efficient. When trained from scratch on ImageNet, PS-ViT performs 3.8% higher than the vanilla ViT in terms of top-1 accuracy with about 4x fewer parameters and 10x fewer FLOPs. Code is available at <https://github.com/yuexy/PS-ViT>.

具有强大全局关系建模能力的变换器最近被引入到基本的计算机视觉任务中。作为一个典型示例，Vision Transformer (ViT) 直接将纯Transformer体系结构应用于图像分类，方法是简单地将图像分割为具有固定长度的标记，并使用Transformer来学习这些标记之间的关系。然而，这种幼稚的标记化可能破坏对象结构，将网格分配给背景等不感兴趣的区域，并引入干扰信号。为了缓解上述问题，在本文中，我们提出了一种迭代渐进采样策略来定位区分区域。在每次迭代中，当前采样步骤的嵌入被馈入变压器编码器层，并且预测一组采样偏移以更新下一步的采样位置。渐进抽样是可微的。当与视觉变换器相结合时，获得的PS ViT网络可以自适应地学习看哪里。建议的PS ViT既有效又高效。当在ImageNet上从头开始训练时，PS ViT在顶级精度方面比香草ViT高3.8%，参数少4倍，触发器少10倍。代码可在[http://github.com/yuexy/PS-ViT](https://github.com/yuexy/PS-ViT).

Learning maps between data samples is fundamental. Applications range from representation learning, image translation and generative modeling, to the estimation of spatial deformations. Such maps relate feature vectors, or map between feature spaces. Well-behaved maps should be regular, which can be imposed explicitly or may emanate from the data itself. We explore what induces regularity for spatial transformations, e.g., when computing image registrations. Classical optimization-based models compute maps between pairs of samples and rely on an appropriate regularizer for well-posedness. Recent deep learning approaches have attempted to avoid using such regularizers altogether by relying on the sample population instead. We explore if it is possible to obtain spatial regularity using an inverse consistency loss only and elucidate what explains map regularity in such a context. We find that deep networks combined with an inverse consistency loss and randomized off-grid interpolation yield well behaved, approximately diffeomorphic, spatial transformations. Despite the simplicity of this approach, our experiments present compelling evidence, on both synthetic and real data, that regular maps can be obtained without carefully tuned explicit regularizers and competitive registration performance.

学习数据样本之间的映射是基础。应用范围从表示学习、图像翻译和生成建模，到空间变形的估计。这种映射关系到特征向量或特征空间之间的映射。表现良好的映射应该是规则的，可以明确地施加，也可以来自数据本身。我们探讨了空间变换的规律性，例如，在计算图像注册时。经典的基于优化的模型计算样本对之间的映射，并依赖于适当的正则化器来实现适定性。最近的深度学习方法试图通过依赖样本总体来避免完全使用此类正则化器。我们探讨是否有可能仅使用逆一致性损失来获得空间规则性，并阐明在这种情况下如何解释地图规则性。我们发现，深度网络与逆一致性损失和随机离网插值相结合，可以产生性能良好的近似微分同胚空间变换。尽管这种方法很简单，但我们的实验在合成数据和真实数据上都提供了令人信服的证据，即无需仔细调整显式正则化器和竞争性注册性能，就可以获得规则映射。

single image deraining is important for many high-level computer vision tasks since the rain streaks can severely degrade the visibility of images, thereby affecting the recognition and analysis of the image. Recently, many CNN-based methods have been proposed for rain removal. Although these methods can remove part of the rain streaks, it is difficult for them to adapt to real-world scenarios and restore high-quality rain-free images with clear and accurate structures. To solve this problem, we propose a Structure-Preserving Deraining Network (SPDNet) with RCP guidance. SPDNet directly generates high-quality rain-free images with clear and accurate structures under the guidance of RCP but does not rely on any rain-generating assumptions. Specifically, we found that the RCP of images contains more accurate structural information than rainy images. Therefore, we introduced it to our deraining network to protect structure information of the rain-free image. Meanwhile, a Wavelet-based Multi-Level Module (WMLM) is proposed as the backbone for learning the background information of rainy images and an Interactive Fusion Module (IFM) is designed to make full use of RCP information. In addition, an iterative guidance strategy is proposed to gradually improve the accuracy of RCP, refining the result in a progressive path. Extensive experimental results on both synthetic and real-world datasets demonstrate that the proposed model achieves new state-of-the-art results. Code: <https://github.com/Joyies/SPDNet>

单幅图像去噪对于许多高级计算机视觉任务都很重要，因为雨纹会严重降低图像的可见性，从而影响图像的识别和分析。最近，许多基于CNN的雨水清除方法被提出。虽然这些方法可以去除部分雨纹，但它们很难适应真实场景，恢复清晰准确的高质量无雨图像。为了解决这个问题，我们提出了一种具有RCP制导的结构保持降额网络（SPDNet）。SPDNet在RCP的指导下直接生成结构清晰准确的高质量无雨图像，但不依赖任何降雨假设。具体来说，我们发现图像的RCP比雨图像包含更准确的结构信息。所以，我们将其引入到我们的除雨网络中，以保护无雨图像的结构信息。同时，提出了一种基于小波的多级模块（WMLM）作为雨天图像背景信息学习的主干，并设计了一种交互式融合模块（IFM）来充分利用RCP信息。此外，还提出了一种迭代制导策略，以逐步提高RCP的精度，并在渐进路径中细化结果。在合成数据集和真实数据集上的大量实验结果表明，该模型取得了最新的结果。代码：<https://github.com/Joyies/SPDNet>

Deep neural networks (DNNs) have been widely used recently while their hardware deployment optimizations are very time-consuming and the historical deployment knowledge is not utilized efficiently. In this paper, to accelerate the optimization process and find better deployment configurations, we propose a novel transfer learning method based on deep Gaussian processes (DGPs). Firstly, a deep Gaussian process (DGP) model is built on the historical data to learn empirical knowledge. Secondly, to transfer knowledge to a new task, a tuning set is sampled for the new task under the guidance of the DGP model. Then DGP is tuned according to the tuning set via maximum-a-posteriori (MAP) estimation to accommodate for the new task and finally used to guide the deployments of the task. The experiments show that our method achieves the best inference latencies of convolutions while accelerating the optimization process significantly, compared with previous arts.

深度神经网络（Deep neural networks, DNNs）近年来得到了广泛的应用，但其硬件部署优化非常耗时，并且没有有效地利用历史部署知识。本文提出了一种基于深度高斯过程（DGPs）的迁移学习方法，以加快优化过程，找到更好的部署配置。首先，基于历史数据建立深度高斯过程（DGP）模型，学习经验知识。其次，在DGP模型的指导下，为新任务采样一个调整集，以将知识转移到新任务。然后根据调整集通过最大后验概率（MAP）估计调整DGP以适应新任务，并最终用于指导任务的部署。实验表明，与以往的方法相比，该方法在显著加快优化过程的同时，获得了最佳的卷积推理延迟。

Existing RGB-D saliency detection models do not explicitly encourage RGB and depth to achieve effective multi-modal learning. In this paper, we introduce a novel multi-stage cascaded learning framework via mutual information minimization to explicitly model the multi-modal information between RGB image and depth data. Specifically, we first map the feature of each mode to a lower dimensional feature vector, and adopt mutual information minimization as a regularizer to reduce the redundancy between appearance features from RGB and geometric features from depth. We then perform multi-stage cascaded learning to impose the mutual information minimization constraint at every stage of the network. Extensive experiments on benchmark RGB-D saliency datasets illustrate the effectiveness of our framework. Further, to prosper the development of this field, we contribute the largest (7x larger than NJU2K) COME20K dataset, which contains 15,625 image pairs with high quality polygon-/scribble-/object-/instance-/rank-level annotations. Based on these rich labels, we additionally construct four new benchmarks (code, results, and benchmarks will be made publicly available.) with strong baselines and observe some interesting phenomena, which can motivate future model design.

现有的RGB-D显著性检测模型没有明确鼓励RGB和深度来实现有效的多模式学习。本文提出了一种新的基于互信息最小化的多阶段级联学习框架，对RGB图像和深度数据之间的多模态信息进行显式建模。具体地说，我们首先将每个模式的特征映射到一个低维特征向量，并采用互信息最小化作为正则化器来减少RGB的外观特征和深度的几何特征之间的冗余。然后，我们执行多级级联学习，在网络的每个阶段施加互信息最小化约束。在基准RGB-D显著性数据集上的大量实验表明了我们的框架的有效性。此外，为了促进该领域的发展，我们提供了最大（比NJU2K大7倍）的COME20K数据集，其中包含15625个图像对，具有高质量的多边形/涂鸦/对象/实例/等级标注。基于这些丰富的标签，我们还构建了四个新的基准测试（代码、结果和基准测试将公开提供）具有强大的基线，并观察到一些有趣的现象，这可以激励未来的模型设计。

This paper addresses weakly supervised amodal instance segmentation, where the goal is to segment both visible and occluded (amodal) object parts, while training provides only ground-truth visible (modal) segmentations. Following prior work, we use data manipulation to generate occlusions in training images and thus train a segmenter to predict amodal segmentations of the manipulated data. The resulting predictions on training images are taken as the pseudo-ground truth for the standard training of Mask-RCNN, which we use for amodal instance segmentation of test images. For generating the pseudo-ground truth, we specify a new Amodal Segmenter based on Boundary Uncertainty estimation (ASBU) and make two contributions. First, while prior work uses the occluder's mask, our ASBU uses the occlusion boundary as input. Second, ASBU estimates an uncertainty map of the prediction. The estimated uncertainty regularizes learning such that lower segmentation loss is incurred on regions with high uncertainty. ASBU achieves significant performance improvement relative to the state of the art on the COCOA and KINS datasets in three tasks: amodal instance segmentation, amodal completion, and ordering recovery.

本文讨论弱监督amodal实例分割，其目标是分割可见和遮挡（amodal）对象部分，而训练只提供地面真实可见（模态）分割。在之前的工作之后，我们使用数据操纵在训练图像中生成遮挡，从而训练分割器来预测操纵数据的amodal分割。将训练图像的预测结果作为Mask-RCNN标准训练的伪地面真值，我们将其用于测试图像的amodal实例分割。为了生成伪地面真值，我们设计了一种新的基于边界不确定性估计（ASBU）的Amodal分割器，并做出了两个贡献。首先，之前的工作使用遮罩，而我们的ASBU使用遮罩边界作为输入。其次，ASBU估计预测的不确定性图。估计的不确定性使学习规则化，从而在具有高不确定性的区域上产生较低的分割损失。相对于COCOA和KINS数据集的最新技术，ASBU在三项任务中实现了显著的性能改进：amodal实例分割、amodal完成和排序恢复。

We present "Cross-Camera Convolutional Color Constancy" (C5), a learning-based method, trained on images from multiple cameras, that accurately estimates a scene's illuminant color from raw images captured by a new camera previously unseen during training. C5 is a hypernetwork-like extension of the convolutional color constancy (CCC) approach: C5 learns to generate the weights of a CCC model that is then evaluated on the input image, with the CCC weights dynamically adapted to different input content. Unlike prior cross-camera color constancy models, which are usually designed to be agnostic to the spectral properties of test-set images from unobserved cameras, C5 approaches this problem through the lens of transductive inference: additional unlabeled images are provided as input to the model at test time, which allows the model to calibrate itself to the spectral properties of the test-set camera during inference. C5 achieves state-of-the-art accuracy for cross-camera color constancy on several datasets, is fast to evaluate (7 and 90 ms per image on a GPU or CPU, respectively), and requires little memory (2 MB), and thus is a practical solution to the problem of calibration-free automatic white balance for mobile photography.

我们提出了“交叉摄像机卷积颜色恒常性”(C5)，这是一种基于学习的方法，对来自多个摄像机的图像进行训练，该方法可以从新摄像机捕获的原始图像中准确估计场景的光源颜色，而新摄像机在训练过程中从未见过。C5是卷积颜色恒定性(CCC)方法的一种类似超网络的扩展：C5学习生成CCC模型的权重，然后对输入图像进行评估，CCC权重动态适应不同的输入内容。与先前的交叉摄像机颜色恒定性模型不同，交叉摄像机颜色恒定性模型通常被设计为对未观测到的摄像机的测试集图像的光谱特性不可知，C5通过传导性推理的镜头来处理这个问题：在测试时，额外的未标记图像被提供作为模型的输入，这允许模型在推理过程中根据测试集摄像机的光谱特性进行自我校准。C5在多个数据集上实现了最先进的跨摄像机颜色恒定性精度，评估速度快（在GPU或CPU上，每个图像分别为7和90毫秒），并且需要很少的内存（2 MB），因此是移动摄影无校准自动白平衡问题的实际解决方案。

We introduce N-ImageNet, a large-scale dataset targeted for robust, fine-grained object recognition with event cameras. The dataset is collected using programmable hardware in which an event camera consistently moves around a monitor displaying images from ImageNet. N-ImageNet serves as a challenging benchmark for event-based object recognition, due to its large number of classes and samples. We empirically show that pretraining on N-ImageNet improves the performance of event-based classifiers and helps them learn with few labeled data. In addition, we present several variants of N-ImageNet to test the robustness of event-based classifiers under diverse camera trajectories and severe lighting conditions, and propose a novel event representation to alleviate the performance degradation. To the best of our knowledge, we are the first to quantitatively investigate the consequences caused by various environmental conditions on event-based object recognition algorithms. N-ImageNet and its variants are expected to guide practical implementations for deploying event-based object recognition algorithms in the real world.

我们介绍了N-ImageNet，这是一个大型数据集，用于使用事件摄影机进行健壮的细粒度对象识别。使用可编程硬件收集数据集，其中事件摄影机始终在显示ImageNet图像的监视器周围移动。N-ImageNet是基于事件的对象识别的一个具有挑战性的基准，因为它有大量的类和样本。我们的经验表明，在N-ImageNet上进行预训练可以提高基于事件的分类器的性能，并帮助它们在标记数据较少的情况下进行学习。此外，我们还提出了N-ImageNet的几种变体，以测试基于事件的分类器在不同摄像机轨迹和恶劣光照条件下的鲁棒性，并提出了一种新的事件表示方法来缓解性能下降。据我们所知，我们是第一个定量研究各种环境条件对基于事件的目标识别算法造成后果的人。N-ImageNet及其变体有望指导在现实世界中部署基于事件的对象识别算法的实际实现。

We present a task and benchmark dataset for person-centric visual grounding, the problem of linking between people named in a caption and people pictured in an image. In contrast to prior work in visual grounding, which is predominantly object-based, our new task masks out the names of people in captions in order to encourage methods trained on such image--caption pairs to focus on contextual cues (such as rich interactions between multiple people), rather than learning associations between names and appearances. To facilitate this task, we introduce a new dataset, Who's Waldo, mined automatically from image--caption data on Wikimedia Commons. We propose a Transformer-based method that outperforms several strong baselines on this task, and are releasing our data to the research community to spur work on contextual models that consider both vision and language.

我们提出了一个任务和基准数据集，用于以人为主的视觉基础，即标题中命名的人和图像中的人物之间的链接问题。与之前的视觉基础研究（主要基于对象）不同，我们的新任务是在字幕中隐藏人名，以鼓励对此类图像进行训练的方法——字幕对专注于上下文线索（如多人之间的丰富互动），而不是学习名字和外表之间的联系。为了简化这项任务，我们引入了一个新的数据集，Who's Waldo，它是从维基媒体共享空间上的图像——字幕数据中自动挖掘出来的。我们提出了一种基于变压器的方法，它优于这个任务上的几个强基线，并将我们的数据发布到研究社区，以刺激工作的上下文模型考虑视觉和语言。

Optimizing the K-class hyperplanes in the latent space has become the standard paradigm for efficient representation learning. However, it's almost impossible to find an optimal K-class hyperplane to accurately describe the latent space of massive noisy data. For this potential problem, we constructively propose a new method, named Switchable K-class Hyperplanes (SKH), to sufficiently describe the latent space by the mixture of K-class hyperplanes. It can directly replace the conventional single K-class hyperplane optimization as the new paradigm for noise-robust representation learning. When collaborated with the popular ArcFace on million-level data representation learning, we found that the switchable manner in SKH can effectively eliminate the gradient conflict generated by real-world label noise on a single K-class hyperplane. Moreover, combined with the margin-based loss functions (e.g. ArcFace), we propose a simple Posterior Data Clean strategy to reduce the model optimization deviation on clean dataset caused by the reduction of valid categories in each K-class hyperplane. Extensive experiments demonstrate that the proposed SKH easily achieves new state-of-the-art on IJB-B and IJB-C by encouraging noise-robust representation learning.

在潜在空间中优化K类超平面已成为有效表征学习的标准范例。然而，几乎不可能找到一个最优的K类超平面来准确描述海量噪声数据的潜在空间。针对这一潜在问题，我们建设性地提出了一种新的方法，称为切换K类超平面（SKH），用K类超平面的混合来充分描述潜在空间。它可以直接取代传统的单K类超平面优化，成为噪声鲁棒表征学习的新范式。当与流行的ArcFace合作进行百万级数据表示学习时，我们发现SKH中的切换方式可以有效地消除真实世界标签噪声在单个K类超平面上产生的梯度冲突。此外，结合基于边缘的损失函数（如ArcFace），我们提出了一种简单的后验数据清理策略，以减少由于每个K类超平面中有效类别的减少而导致的干净数据集上的模型优化偏差。大量的实验表明，通过鼓励噪声鲁棒表征学习，所提出的SKH可以很容易地在IJB-B和IJB-C上实现新的技术水平。

City modeling is the foundation for computational urban planning, navigation, and entertainment. In this work, we present the first generative model of city blocks named BlockPlanner, and showcase its ability to synthesize valid city blocks with varying land lots configurations. We propose a novel vectorized city block representation utilizing a ring topology and a two-tier graph to capture the global and local structures of a city block. Each land lot is abstracted into a vector representation covering both its 3D geometry and land use semantics. Such vectorized representation enables us to deploy a lightweight network to capture the underlying distribution of land lots configuration in a city block. To enforce intrinsic spatial constraints of a valid city block, a set of effective loss functions are imposed to shape rational results. We contribute a pilot city block dataset to demonstrate the effectiveness and efficiency of our representation and framework over the state-of-the-art. Notably, our BlockPlanner is also able to edit and manipulate city blocks, enabling several useful applications, e.g., topology refinement and footprint generation.

城市建模是计算城市规划、导航和娱乐的基础。在这项工作中，我们提出了第一个名为BlockPlanner的城市街区生成模型，并展示了其合成具有不同地块配置的有效城市街区的能力。我们提出了一种新的矢量化城市街区表示方法，利用环形拓扑和两层图来捕捉城市街区的全局和局部结构。每个地块都被抽象为一个向量表示，涵盖其三维几何结构和土地使用语义。这种矢量化表示使我们能够部署一个轻量级网络，以捕获城市街区中地块配置的基本分布。为了加强有效城市街区的内在空间约束，一组有效的损失函数被用来形成合理的结果。我们提供了一个试点城市街区数据集，以证明我们的表示和框架在最新水平上的有效性和效率。值得注意的是，我们的BlockPlanner还能够编辑和操作城市街区，从而实现一些有用的应用，例如拓扑优化和足迹生成。

Cognitive grammar suggests that the acquisition of language grammar is grounded within visual structures. While grammar is an essential representation of natural language, it also exists ubiquitously in vision to represent the hierarchical part-whole structure. In this work, we study grounded grammar induction of vision and language in a joint learning framework. Specifically, we present VLGrammar, a method that uses compound probabilistic context-free grammars (compound PCFGs) to induce the language grammar and the image grammar simultaneously. We propose a novel contrastive learning framework to guide the joint learning of both modules. To provide a benchmark for the grounded grammar induction task, we collect a large-scale dataset, PartIt, which contains human-written sentences that describe part-level semantics for 3D objects. Experiments on the PartIt dataset show that VLGrammar outperforms all baselines in image grammar induction and language grammar induction. The learned VLGrammar naturally benefits related downstream tasks. Specifically, it improves the image unsupervised clustering accuracy by 30%, and performs well in image retrieval and text retrieval. Notably, the induced grammar shows superior generalizability by easily generalizing to unseen categories.

认知语法认为语言语法的习得是以视觉结构为基础的。虽然语法是自然语言的一种基本表征，但它也普遍存在于视觉中，用来表征层次结构的部分和整体结构。在这项工作中，我们研究了视觉和语言在联合学习框架下的扎根语法归纳。特别地，我们提出了VLGRAMR，一种使用复合概率上下文无关语法（复合PCFGs）同时归纳语言语法和图像语法的方法。我们提出了一个新的对比学习框架来指导两个模块的联合学习。为了为扎根语法归纳任务提供一个基准，我们收集了一个大规模的数据集PartIt，其中包含描述三维对象的部分级语义的人类书面语句。在PartIt数据集上的实验表明，VLGrammar在图像语法归纳和语言语法归纳方面优于所有基线。所学的语法自然有利于相关的下游任务。具体来说，它将图像无监督聚类的准确率提高了30%，并且在图像检索和文本检索中表现良好。值得注意的是，归纳语法通过容易地归纳到看不见的类别而显示出优越的归纳能力。

Representing human-made objects as a collection of base primitives has a long history in computer vision and reverse engineering. In the case of high-resolution point cloud scans, the challenge is to be able to detect both large primitives as well as those explaining the detailed parts. While the classical RANSAC approach requires case-specific parameter tuning, state-of-the-art networks are limited by memory consumption of their backbone modules such as PointNet++, and hence fail to detect the fine-scale primitives. We present Cascaded Primitive Fitting Networks (CPFN) that relies on an adaptive patch sampling network to assemble detection results of global and local primitive detection networks. As a key enabler, we present a merging formulation that dynamically aggregates the primitives across global and local scales. Our evaluation demonstrates that CPFN improves the state-of-the-art SPFN performance by 13-14% on high-resolution point cloud datasets and specifically improves the detection of fine-scale primitives by 20-22%. Our code is available at: <https://github.com/erictuanle/CPFN>

在计算机视觉和逆向工程中，将人造物体表示为基本原语的集合有着悠久的历史。在高分辨率点云扫描的情况下，挑战在于能够检测大型原语以及解释详细部分的原语。虽然经典的RANSAC方法需要特定于具体情况的参数调整，但最先进的网络受到其主干模块（如PointNet++）内存消耗的限制，因此无法检测到精细规模的原语。我们提出了级联原始拟合网络（CPFN），该网络依赖于自适应面片采样网络来汇集全局和局部原始检测网络的检测结果。作为一个关键的使能因素，我们提出了一个合并公式，可以在全局和局部范围内动态聚合原语。我们的评估表明，在高分辨率点云数据集上，CPFN将最先进的SPFN性能提高了13-14%，特别是将精细尺度基元的检测提高了20-22%。我们的代码可从以下网址获得：<https://github.com/erictuanle/CPFN>

Recent works on implicit neural representations have shown promising results for multi-view surface reconstruction. However, most approaches are limited to relatively simple geometries and usually require clean object masks for reconstructing complex and concave objects. In this work, we introduce a novel neural surface reconstruction framework that leverages the knowledge of stereo matching and feature consistency to optimize the implicit surface representation. More specifically, we apply a signed distance field (SDF) and a surface light field to represent the scene geometry and appearance respectively. The SDF is directly supervised by geometry from stereo matching, and is refined by optimizing the multi-view feature consistency and the fidelity of rendered images. Our method is able to improve the robustness of geometry estimation and support reconstruction of complex scene topologies. Extensive experiments have been conducted on DTU, EPFL and Tanks and Temples datasets. Compared to previous state-of-the-art methods, our method achieves better mesh reconstruction in wide open scenes without masks as input.

最近关于隐式神经表示的研究表明，隐式神经表示对于多视图曲面重建具有良好的效果。然而，大多数方法仅限于相对简单的几何图形，通常需要干净的对象遮罩来重建复杂和凹面对象。在这项工作中，我们介绍了一种新的神经曲面重建框架，该框架利用立体匹配和特征一致性的知识来优化隐式曲面表示。更具体地说，我们分别应用有符号距离场（SDF）和曲面光场来表示场景几何体和外观。SDF由立体匹配中的几何图形直接监督，并通过优化多视图特征的一致性和渲染图像的保真度来细化。我们的方法能够提高几何估计的鲁棒性，并支持复杂场景拓扑的重建。在DTU、EPFL和储罐和庙宇数据集上进行了广泛的实验。与以前的最新方法相比，我们的方法在完全开放的场景中实现了更好的网格重建，无需使用遮罩作为输入。

Generalization on out-of-distribution (OOD) test data is an essential but underexplored topic in visual question answering. Current state-of-the-art VQA models often exploit the biased correlation between data and labels, which results in a large performance drop when the test and training data have different distributions. Inspired by the fact that humans can recognize novel concepts by composing existed concepts and capsule network's ability of representing part-whole hierarchies, we propose to use capsules to represent parts and introduce "Linguistically Routing" to merge parts with human-prior hierarchies. Specifically, we first fuse visual features with a single question word as atomic parts. Then we introduce the "Linguistically Routing" to reweight the capsule connections between two layers such that: 1) the lower layer capsules can transfer their outputs to the most compatible higher capsules, and 2) two capsules can be merged if their corresponding words are merged in the question parse tree. The routing process maximizes the above unary and binary potentials across multiple layers and finally carves a tree structure inside the capsule network. We evaluate our proposed routing method on the CLEVR compositional generation test, the VQA-CP2 dataset and the VQAv2 dataset. The experimental results show that our proposed method can improve current VQA models on OOD split without losing performance on the in-domain test data.

非分布（OOD）测试数据的泛化是视觉问答中一个重要但尚未得到充分研究的课题。当前最先进的VQA模型通常利用数据和标签之间的偏差相关性，当测试和训练数据具有不同的分布时，这会导致性能大幅下降。受人类可以通过组合已有概念来识别新概念以及胶囊网络表示部分-整体层次结构的能力的启发，我们建议使用胶囊来表示部分，并引入“语言路由”将部分与人类先前的层次结构合并。具体来说，我们首先将视觉特征与单个疑问词作为原子部分进行融合。然后我们引入“语言路由”来重新加权两层之间的胶囊连接，这样：1) 下层胶囊可以将其输出传输到最兼容的高层胶囊，2) 如果在问题解析树中合并了两个胶囊对应的单词，则可以合并两个胶囊。路由过程使上述一元和二元电位跨多个层最大化，并最终在胶囊网络内雕刻出树形结构。我们在CLEVR合成生成测试、VQA-CP2数据集和VQAv2数据集上评估了我们提出的路由方法。实验结果表明，该方法可以在不损失域内测试数据的情况下，改进现有的面向对象分割的VQA模型。

We present Neural Articulated Radiance Field (NARF), a novel deformable 3D representation for articulated objects learned from images. While recent advances in 3D implicit representation have made it possible to learn models of complex objects, learning pose-controllable representations of articulated objects remains a challenge, as current methods require 3D shape supervision and are unable to render appearance. In formulating an implicit representation of 3D articulated objects, our method considers only the rigid transformation of the most relevant object part in solving for the radiance field at each 3D location. In this way, the proposed method represents pose-dependent changes without significantly increasing the computational complexity. NARF is fully differentiable and can be trained from images with pose annotations. Moreover, through the use of an autoencoder, it can learn appearance variations over multiple instances of an object class. Experiments show that the proposed method is efficient and can generalize well to novel poses. The code is available for research purposes at <https://github.com/nogu-atsu/NARF>

我们提出了神经关节辐射场（NARF），一种新的可变形三维表示从图像中学习的关节对象。虽然3D隐式表示的最新进展使得学习复杂对象的模型成为可能，但学习关节对象的姿势可控表示仍然是一个挑战，因为当前的方法需要3D形状监控，并且无法呈现外观。在建立三维关节对象的隐式表示时，我们的方法只考虑了在求解每个三维位置的辐射场时最相关对象部分的刚性变换。这样，所提出的方法在不显著增加计算复杂度的情况下表示姿势相关的变化。NARF是完全可微的，可以从带有姿势注释的图像中进行训练。此外，通过使用自动编码器，它可以了解对象类的多个实例的外观变化。实验表明，该方法是有效的，可以很好地推广到新的姿态。该守则可在<https://github.com/nogu-atsu/NARF>

Knowledge distillation (KD) transfers the dark knowledge from cumbersome networks (teacher) to lightweight (student) networks and expects the student to achieve more promising performance than training without the teacher's knowledge. However, a counter-intuitive argument is that better teachers do not make better students due to the capacity mismatch. To this end, we present a novel adaptive knowledge distillation method to complement traditional approaches. The proposed method, named as Student Customized Knowledge Distillation (SCKD), examines the capacity mismatch between teacher and student from the perspective of gradient similarity. We formulate the knowledge distillation as a multi-task learning problem so that the teacher transfers knowledge to the student only if the student can benefit from learning such knowledge. We validate our methods on multiple datasets with various teacher-student configurations on image classification, object detection, and semantic segmentation.

知识提炼 (KD) 将黑暗的知识从笨重的网络 (教师) 转移到轻量级 (学生) 网络，并期望学生取得比没有教师知识的培训更有希望的成绩。然而，一个与直觉相反的论点是，由于能力不匹配，更好的教师不会造就更好的学生。为此，我们提出了一种新的自适应知识提取方法来补充传统方法。该方法称为学生定制知识提取 (SCKD)，从梯度相似性的角度考察了教师和学生之间的能力失配。我们将知识提炼描述为一个多任务学习问题，这样，只有当学生能够从学习中受益时，教师才能将知识传授给学生。我们在多个数据集上验证了我们的方法，这些数据集在图像分类、目标检测和语义分割方面具有不同的师生配置。

We consider the scalable recognition problem in the fine-grained expert domain where large-scale data collection is easy whereas annotation is difficult. Existing solutions are typically based on semi-supervised or self-supervised learning. We propose an alternative new framework, MEMORABLE, based on machine teaching and online crowdsourcing platforms. A small amount of data is first labeled by experts and then used to teach online annotators for the classes of interest, who finally label the entire dataset. Preliminary studies show that the accuracy of classifiers trained on the final dataset is a function of the accuracy of the student annotators. A new machine teaching algorithm, CMaxGrad, is then proposed to enhance this accuracy by introducing explanations in a state-of-the-art machine teaching algorithm. For this, CMaxGrad leverages counterfactual explanations, which take into account student predictions, thereby proving feedback that is student-specific, explicitly addresses the causes of student confusion, and adapts to the level of competence of the student. Experiments show that both MEMORABLE and CMaxGrad outperform existing solutions to their respective problems.

我们考虑可扩展的识别问题在细粒度的专家领域，大规模数据收集是容易的，而注释是困难的。现有的解决方案通常基于半监督或自监督学习。我们基于机器教学和在线众包平台，提出了另一种新的框架，令人难忘。一小部分数据首先由专家标记，然后用于教授感兴趣的类的在线注释员，最后由他们标记整个数据集。初步研究表明，在最终数据集上训练的分类器的准确性是学生注释者准确性的函数。然后提出了一种新的机器教学算法CMaxGrad，通过在最先进的机器教学算法中引入解释来提高这种精度。为此，CMaxGrad利用反事实解释，考虑到学生的预测，从而证明反馈是针对学生的，明确解决了学生困惑的原因，并适应学生的能力水平。实验表明，对于各自的问题，Membrable和CMaxGrad都优于现有的解决方案。

Video super-resolution (VSR) aims to improve the spatial resolution of low-resolution (LR) videos. Existing VSR methods are mostly trained and evaluated on synthetic datasets, where the LR videos are uniformly downsampled from their high-resolution (HR) counterparts by some simple operators (e.g., bicubic downsampling). Such simple synthetic degradation models, however, cannot well describe the complex degradation processes in real-world videos, and thus the trained VSR models become ineffective in real-world applications. As an attempt to bridge the gap, we build a real-world video super-resolution (RealVSR) dataset by capturing paired LR-HR video sequences using the multi-camera system of iPhone 11 Pro Max. Since the LR-HR video pairs are captured by two separate cameras, there are inevitably certain misalignment and luminance/color differences between them. To more robustly train the VSR model and recover more details from the LR inputs, we convert the LR-HR videos into YCbCr space and decompose the luminance channel into a Laplacian pyramid, and then apply different loss functions to different components. Experiments validate that VSR models trained on our RealVSR dataset demonstrate better visual quality than those trained on synthetic datasets under real-world settings. They also exhibit good generalization capability in cross-camera tests. The dataset and code can be found at <https://github.com/IanYeung/RealVSR>.

视频超分辨率 (VSR) 旨在提高低分辨率 (LR) 视频的空间分辨率。现有的VSR方法大多在合成数据集上进行训练和评估，其中LR视频通过一些简单的操作（例如，双三次下采样）从其高分辨率 (HR) 对应物中均匀下采样。然而，这种简单的综合退化模型不能很好地描述现实世界视频中的复杂退化过程，因此训练的VSR模型在实际应用中变得无效。为了弥补这一差距，我们通过使用iPhone 11 Pro Max的多摄像头系统捕获成对的LR-HR视频序列，构建了一个真实世界的视频超分辨率 (RealVSR) 数据集。由于LR-HR视频对由两个单独的摄像头捕获，因此它们之间不可避免地存在一定的错位和亮度/颜色差异。为了更稳健地训练VSR模型并从LR输入中恢复更多细节，我们将LR-HR视频转换为YCbCr空间，并将亮度通道分解为拉普拉斯金字塔，然后对不同的分量应用不同的损失函数。实验证明，在真实环境下，在RealVSR数据集上训练的VSR模型比在合成数据集上训练的VSR模型具有更好的视觉质量。它们在交叉摄像机测试中也表现出良好的泛化能力。数据集和代码可在以下位置找到：<https://github.com/IanYeung/RealVSR>。

We study self-supervised video representation learning, which is a challenging task due to 1) sufficient labels for supervision; 2) unstructured and noisy visual information. Existing methods mainly use contrastive loss with video clips as the instances and learn visual representation by discriminating instances from each other, but they need a careful treatment of negative pairs by either relying on large batch sizes, memory banks, extra modalities or customized mining strategies, which inevitably includes noisy data. In this paper, we observe that the consistency between positive samples is the key to learn robust video representation. Specifically, we propose two tasks to learn appearance and speed consistency, respectively. The appearance consistency task aims to maximize the similarity between two clips of the same video with different playback speeds. The speed consistency task aims to maximize the similarity between two clips with the same playback speed but different appearance information. We show that optimizing the two tasks jointly consistently improves the performance on downstream tasks, e.g., action recognition and video retrieval. Remarkably, for action recognition on the UCF-101 dataset, we achieve 90.8% accuracy without using any extra modalities or negative pairs for unsupervised pre-training, which outperforms the ImageNet supervised pre-trained model. Codes and models will be available.

我们研究了自监督视频表示学习，这是一项具有挑战性的任务，因为1) 有足够的标签用于监督；2) 非结构化和嘈杂的视觉信息。现有的方法主要使用视频片段的对比损失作为实例，通过相互区分实例来学习视觉表示，但它们需要通过依赖大批量、内存库、额外模式或定制挖掘策略仔细处理负对，这不可避免地包含了嘈杂的数据。在本文中，我们观察到正样本之间的一致性是学习鲁棒视频表示的关键。具体来说，我们提出了两个任务，分别学习外观和速度一致性。外观一致性任务旨在最大限度地提高相同视频的两个片段之间的相似性，且播放速度不同。速度一致性任务旨在最大化具有相同播放速度但不同外观信息的两个剪辑之间的相似性。我们发现，联合优化这两个任务可以持续提高下游任务的性能，例如动作识别和视频检索。值得注意的是，对于UCF-101数据集上的动作识别，我们在不使用任何额外模式或负对进行无监督预训练的情况下实现了90.8%的准确率，这优于ImageNet监督预训练模型。将提供代码和型号。

Batch Whitening is a technique that accelerates and stabilizes training by transforming input features to have a zero mean (Centering) and a unit variance (Scaling), and by removing linear correlation between channels (Decorrelation). In commonly used structures, which are empirically optimized with Batch Normalization, the normalization layer appears between convolution and activation function. Following Batch Whitening studies have employed the same structure without further analysis; even Batch whitening was analyzed on the premise that the input of a linear layer is whitened. To bridge the gap, we propose a new Convolutional Unit that in line with the theory, and our method generally improves the performance of Batch whitening. Moreover, we show the inefficacy of the original Convolutional Unit by investigating rank and correlation of features. As our method is employable off-the-shelf whitening modules, we use Iterative Normalization (IterNorm), the state-of-the-art whitening module, and obtain significantly improved performance on five image classification datasets: CIFAR-10, CIFAR-100, CUB-200-2011, Stanford Dogs, and ImageNet. Notably, we verify that our method improves stability and performance of whitening when using large learning rate, group size, and iteration number.

批处理白化是一种通过将输入特征转换为零均值（居中）和单位方差（缩放）以及消除通道之间的线性相关性（去相关）来加速和稳定训练的技术。在通常使用的结构中，通过批量归一化进行经验优化，归一化层出现在卷积和激活函数之间。以下批量增白研究采用了相同的结构，无需进一步分析；在线性层输入被白化的前提下，分析了均匀批量白化。为了弥补这一差距，我们提出了一种新的符合理论的卷积单元，并且我们的方法总体上提高了批处理白化的性能。此外，我们通过研究特征的秩和相关性来证明原始卷积单元的无效性。由于我们的方法可用于现成的增白模块，因此我们使用了迭代归一化(IterNorm)这一最先进的增白模块，并在五个图像分类数据集(CIFAR-10、CIFAR-100、CUB-200-2011、斯坦福狗和ImageNet)上获得了显著改进的性能。值得注意的是，我们验证了当使用较大的学习率、组大小和迭代次数时，我们的方法提高了白化的稳定性和性能。

The growing use of deep learning for a wide range of data problems has highlighted the need to understand and diagnose these models appropriately, making deep learning interpretation techniques an essential tool for data analysts. The numerous model interpretation methods proposed in recent years are generally based on heuristics, with little or no theoretical guarantees. Here we present a statistical framework for saliency estimation for black-box computer vision models. Our proposed model-agnostic estimation procedure, which is statistically consistent and capable of passing saliency checks, has polynomial-time computational efficiency since it only requires solving a linear program. An upper bound is established on the number of model evaluations needed to recover regions of importance with high probability through our theoretical analysis. Furthermore, a new perturbation scheme is presented for the estimation of local gradients that is more efficient than commonly used random perturbation schemes. The validity and excellence of our new method are demonstrated experimentally using sensitivity analysis on multiple datasets.

深度学习在广泛的数据问题中的应用日益广泛，这突出了正确理解和诊断这些模型的必要性，使深度学习解释技术成为数据分析人员的一个重要工具。近年来提出的许多模型解释方法通常基于启发式，很少或没有理论保证。在这里，我们提出了一个统计框架的显著性估计黑箱计算机视觉模型。我们提出的模型不可知估计方法在统计上是一致的，并且能够通过显著性检查，由于它只需要求解一个线性规划，因此具有多项式时间的计算效率。通过我们的理论分析，建立了以高概率恢复重要区域所需的模型评估数量上限。此外，本文还提出了一种新的局部梯度估计的摄动格式，它比常用的随机摄动格式更有效。通过对多个数据集的灵敏度分析，验证了新方法的有效性和优越性。

ultra-high resolution image segmentation has raised increasing interests in recent years due to its realistic applications. In this paper, we innovate the widely used high-resolution image segmentation pipeline, in which an ultra-high resolution image is partitioned into regular patches for local segmentation and then the local results are merged into a high-resolution semantic mask. In particular, we introduce a novel locality-aware contextual correlation based segmentation model to process local patches, where the relevance between local patch and its various contexts are jointly and complementarily utilized to handle the semantic regions with large variations. Additionally, we present a contextual semantics refinement network that associates the local segmentation result with its contextual semantics, and thus is endowed with the ability of reducing boundary artifacts and refining mask contours during the generation of final high-resolution mask. Furthermore, in comprehensive experiments, we demonstrate that our model outperforms other state-of-the-art methods in public benchmarks.

超高分辨率图像分割由于其现实的应用，近年来引起了人们越来越多的兴趣。在本文中，我们创新了广泛使用的高分辨率图像分割流水线，将超高分辨率图像分割成规则块进行局部分割，然后将局部结果合并到高分辨率语义掩码中。特别是，我们引入了一种新的基于位置感知上下文相关的分割模型来处理局部补丁，其中局部补丁与其不同上下文之间的相关性被联合和互补地用于处理变化较大的语义区域。此外，我们提出了一个上下文语义细化网络，该网络将局部分割结果与其上下文语义相关联，从而在最终高分辨率掩模生成过程中具有减少边界伪影和细化掩模轮廓的能力。此外，在综合实验中，我们证明了我们的模型在公共基准测试中优于其他最先进的方法。

Knowledge distillation (KD) has been proven a simple and effective tool for training compact dense prediction models. Lightweight student networks are trained by extra supervision transferred from large teacher networks. Most previous KD variants for dense prediction tasks align the activation maps from the student and teacher network in the spatial domain, typically by normalizing the activation values on each spatial location and minimizing point-wise and/or pair-wise discrepancy. Different from the previous methods, here we propose to normalize the activation map of each channel to obtain a soft probability map. By simply minimizing the Kullback--Leibler (KL) divergence between the channel-wise probability map of the two networks, the distillation process pays more attention to the most salient regions of each channel, which are valuable for dense prediction tasks. We conduct experiments on a few dense prediction tasks, including semantic segmentation and object detection. Experiments demonstrate that our proposed method outperforms state-of-the-art distillation methods considerably, and can require less computational cost during training. In particular, we improve the RetinaNet detector (ResNet50 backbone) by 3.4% in mAP on the COCO dataset and spent (ResNet18 backbone) by 5.81% in mIoU on the cityscapes dataset. Code is available at: <https://git.io/Distiller>.

知识提取 (KD) 已被证明是训练紧凑密集预测模型的一种简单而有效的工具。轻量级学生网络通过从大型教师网络转移过来的额外监督进行培训。大多数以前用于密集预测任务的KD变体在空间域中对齐来自学生和教师网络的激活图，通常是通过规范化每个空间位置上的激活值并最小化逐点和/或成对差异。与以前的方法不同，这里我们建议对每个通道的激活图进行规范化，以获得软概率图。通过简单地最小化

两个网络的通道概率图之间的Kullback--Leibler (KL) 发散，蒸馏过程更加关注每个通道的最显著区域，这对于密集预测任务非常有价值。我们在一些密集的预测任务上进行了实验，包括语义分割和目标检测。实验表明，我们提出的方法大大优于现有的蒸馏方法，并且在训练过程中需要较少的计算量。特别是，我们将视网膜网检测器 (RESNET50主干) 改进了3。COCO数据集上的mAP为4%，花费 (ResNet18主干) 为5%。城市景观数据集上81%的百万富翁。代码可从以下网址获取：<https://git.io/Distiller>。

Deep CNN-based methods have so far achieved the state of the art results in multi-view 3D object reconstruction. Despite the considerable progress, the two core modules of these methods - view feature extraction and multi-view fusion, are usually investigated separately, and the relations among multiple input views are rarely explored. Inspired by the recent great success in Transformer models, we reformulate the multi-view 3D reconstruction as a sequence-to-sequence prediction problem and propose a framework named 3D Volume Transformer. Unlike previous CNN-based methods using a separate design, we unify the feature extraction and view fusion in a single Transformer network. A natural advantage of our design lies in the exploration of view-to-view relationships using self-attention among multiple unordered inputs. On ShapeNet - a large-scale 3D reconstruction benchmark, our method achieves a new state-of-the-art accuracy in multi-view reconstruction with fewer parameters (70% less) than CNN-based methods. Experimental results also suggest the strong scaling capability of our method. Our code will be made publicly available.

到目前为止，基于深度CNN的方法在多视图3D对象重建方面取得了最新成果。尽管取得了相当大的进展，但这些方法的两个核心模块——视图特征提取和多视图融合通常是分开研究的，而对多个输入视图之间的关系研究很少。受变压器模型最近取得的巨大成功的启发，我们将多视图三维重建转化为一个序列到序列的预测问题，并提出了一个名为3D Volume Transformer的框架。与以前基于CNN的方法使用单独的设计不同，我们在单个变压器网络中统一了特征提取和视图融合。我们的设计的一个自然优势在于探索在多个无序输入中使用自我关注的视图到视图关系。在大规模三维重建基准ShapeNet上，我们的方法在多视图重建中实现了新的最先进的精度，与基于CNN的方法相比，参数更少（减少70%）。实验结果还表明，我们的方法具有很强的标度能力。我们的代码将公开展示。

Finding shape correspondences can be formulated as an NP-hard quadratic assignment problem (QAP) that becomes infeasible for shapes with high sampling density. A promising research direction is to tackle such quadratic optimization problems over binary variables with quantum annealing, which allows for some problems a more efficient search in the solution space. Unfortunately, enforcing the linear equality constraints in QAPs via a penalty significantly limits the success probability of such methods on currently available quantum hardware. To address this limitation, this paper proposes Q-Match, i.e., a new iterative quantum method for QAPs inspired by the alpha-expansion algorithm, which allows solving problems of an order of magnitude larger than current quantum methods. It implicitly enforces the QAP constraints by updating the current estimates in a cyclic fashion. Further, Q-Match can be applied iteratively, on a subset of well-chosen correspondences, allowing us to scale to real-world problems. Using the latest quantum annealer, the D-Wave Advantage, we evaluate the proposed method on a subset of QAPLIB as well as on isometric shape matching problems from the FAUST dataset.

寻找形状对应关系可以表述为一个NP难的二次分配问题 (QAP)，对于高采样密度的形状不可行。一个很有前途的研究方向是利用量子退火技术解决二元变量上的二次优化问题，这使得某些问题在解空间中的搜索更加有效。不幸的是，通过惩罚强制QAP中的线性等式约束极大地限制了此类方法在当前可用的量子硬件上的成功概率。为了解决这一限制，本文提出了Q-Match，即受alpha展开算法启发的一种新的qap迭代量子方法，它允许解决比当前量子方法大一个数量级的问题。它以循环方式更新当前估计值，

从而隐式实施QAP约束。此外，Q-匹配可以迭代地应用于精心选择的对应关系的子集，使我们能够扩展到现实世界的问题。我们使用最新的量子退火机D-Wave Advantage，对QAPLIB子集以及FAUST数据集中的等距形状匹配问题评估了所提出的方法。

The recently proposed Detection Transformer (DETR) model successfully applies Transformer to objects detection and achieves comparable performance with two-stage object detection frameworks, such as Faster-RCNN. However, DETR suffers from its slow convergence. Training DETR from scratch needs 500 epochs to achieve a high accuracy. To accelerate its convergence, we propose a simple yet effective scheme for improving the DETR framework, namely Spatially Modulated Co-Attention (SMCA) mechanism. The core idea of SMCA is to conduct location-aware co-attention in DETR by constraining co-attention responses to be high near initially estimated bounding box locations. Our proposed SMCA increases DETR's convergence speed by replacing the original co-attention mechanism in the decoder while keeping other operations in DETR unchanged. Furthermore, by integrating multi-head and scale-selection attention designs into SMCA, our fully-fledged SMCA can achieve better performance compared to DETR with a dilated convolution-based backbone (45.6 mAP at 108 epochs vs. 43.3 mAP at 500 epochs). We perform extensive ablation studies on COCO dataset to validate SMCA. Code is released at <https://github.com/gaopengcuhk/SMCA-DETR>.

最近提出的检测变压器（DETR）模型成功地将变压器应用于对象检测，并实现了与两阶段对象检测框架（如更快的RCNN）相当的性能。然而，DETR收敛缓慢。从头开始的训练需要500个纪元才能达到高精度。为了加快其收敛速度，我们提出了一种简单而有效的改进DETR框架的方案，即空间调制共注意（SMCA）机制。SMCA的核心思想是在DETR中通过将共同注意反应限制在初始估计的边界框位置附近，从而实现位置感知的共同注意。我们提出的SMCA在保持DETR中其他操作不变的情况下，通过替换解码器中原有的共同注意机制，提高了DETR的收敛速度。此外，通过将多头和比例选择注意设计集成到SMCA中，我们成熟的SMCA可以实现比基于扩展卷积主干的DETR更好的性能（108个历年时45.6 mAP，500个历年时43.3 mAP）。我们对COCO数据集进行了广泛的消融研究，以验证SMCA。代码发布于<https://github.com/gaopengcuhk/SMCA-DETR>。

Lesion segmentation in medical imaging has been an important topic in clinical research. Researchers have proposed various detection and segmentation algorithms to address this task. Recently, deep learning-based approaches have significantly improved the performance over conventional methods. However, most state-of-the-art deep learning methods require the manual design of multiple network components and training strategies. In this paper, we propose a new automated machine learning algorithm, T-AutoML, which not only searches for the best neural architecture, but also finds the best combination of hyper-parameters and data augmentation strategies simultaneously. The proposed method utilizes the modern transformer model, which is introduced to adapt to the dynamic length of the search space embedding and can significantly improve the ability of the search. We validate T-AutoML on several large-scale public lesion segmentation data-sets and achieve state-of-the-art performance.

医学影像学中的病灶分割一直是临床研究的重要课题。研究人员提出了各种检测和分割算法来解决这一问题。最近，基于深度学习的方法与传统方法相比显著提高了性能。然而，大多数最先进的深度学习方法需要手动设计多个网络组件和培训策略。在本文中，我们提出了一种新的自动机器学习算法T-AutoML，它不仅可以搜索最佳的神经结构，而且可以同时找到超参数和数据扩充策略的最佳组合。该方法利用现代变压器模型，引入变压器模型以适应搜索空间嵌入的动态长度，可以显著提高搜索能力。我们在几个大型公共病变分割数据集上验证了T-AutoML，并实现了最先进的性能。

The ability to localize and segment objects from unseen classes would open the door to new applications, such as autonomous object learning in active vision. Nonetheless, improving the performance on unseen classes requires additional training data, while manually annotating the objects of the unseen classes can be labor-extensive and expensive. In this paper, we explore the use of unlabeled video sequences to automatically generate training data for objects of unseen classes. It is in principle possible to apply existing video segmentation methods to unlabeled videos and automatically obtain object masks, which can then be used as a training set even for classes with no manual labels available. However, our experiments show that these methods do not perform well enough for this purpose. We therefore introduce a Bayesian method that is specifically designed to automatically create such a training set: Our method starts from a set of object proposals and relies on (non-realistic) analysis-by-synthesis to select the correct ones by performing an efficient optimization over all the frames simultaneously. Through extensive experiments, we show that our method can generate a high-quality training set which significantly boosts the performance of segmenting objects of unseen classes. We thus believe that our method could open the door for open-world instance segmentation by exploiting abundant Internet videos.

从看不见的类中定位和分割对象的能力将为新的应用打开大门，例如主动视觉中的自主对象学习。尽管如此，改进不可见类的性能需要额外的培训数据，而手动注释不可见类的对象可能需要大量的人力和昂贵的成本。在本文中，我们探讨了使用未标记的视频序列来自动生成未知类对象的训练数据。原则上，可以将现有的视频分割方法应用于未标记的视频，并自动获取对象掩码，然后将其用作训练集，即使对于没有手动标签的课程也是如此。然而，我们的实验表明，这些方法不能很好地实现这一目的。因此，我们引入了一种专门设计用于自动创建此类训练集的贝叶斯方法：我们的方法从一组对象建议开始，依靠（非现实的）综合分析，通过对所有帧同时执行有效优化来选择正确的建议。通过大量的实验，我们证明了我们的方法能够生成高质量的训练集，显著提高了对未知类对象的分割性能。因此，我们相信我们的方法可以通过利用丰富的互联网视频为开放世界的实例分割打开大门。

This paper presents a new task, point cloud object co-segmentation, aiming to segment the common 3D objects in a set of point clouds. We formulate this task as an object point sampling problem, and develop two techniques, the mutual attention module and co-contrastive learning, to enable it. The proposed method employs two point samplers based on deep neural networks, the object sampler and the background sampler. The former targets at sampling points of common objects while the latter focuses on the rest. The mutual attention module explores point-wise correlation across point clouds. It is embedded in both samplers and can identify points with strong cross-cloud correlation from the rest. After extracting features for points selected by the two samplers, we optimize the networks by developing the co-contrastive loss, which minimizes feature discrepancy of the estimated object points while maximizing feature separation between the estimated object and background points. Our method works on point clouds of an arbitrary object class. It is end-to-end trainable and does not need point-level annotations. It is evaluated on the ScanObjectNN and S3DIS datasets and achieves promising results.

本文提出了一种新的点云目标协同分割方法，旨在对一组点云中常见的三维目标进行分割。我们将此任务描述为一个对象点抽样问题，并开发了两种技术，即相互注意模块和对比学习。该方法采用基于深度神经网络的两点采样器，即目标采样器和背景采样器。前者的目标是普通物体的采样点，而后者的目标是其他物体的采样点。相互注意模块探索点云之间的点相关。它嵌入在两个采样器中，可以从其他采样器中识别出具有强跨云相关性的点。在为两个采样器选择的点提取特征后，我们通过发展对比损失来优化网络，从而最小化估计目标点的特征差异，同时最大化估计目标点和背景点之间的特征分离。我们的

方法适用于任意对象类的点云。它是端到端可培训的，不需要点级注释。在ScanObjectNN和S3DIS数据集上对其进行了评估，并取得了令人满意的结果。

Unsupervised person re-identification (Re-ID) remains challenging due to the lack of ground-truth labels. Existing methods often rely on estimated pseudo labels via iterative clustering and classification, and they are unfortunately highly susceptible to performance penalties incurred by the inaccurate estimated number of clusters. Alternatively, we propose the Meta Pairwise Relationship Distillation (MPRD) method to estimate the pseudo labels of sample pairs for unsupervised person Re-ID. Specifically, it consists of a Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN), in which the GCN estimates the pseudo labels of sample pairs based on the current features extracted by CNN, and the CNN learns better features by involving high-fidelity positive and negative sample pairs imposed by GCN. To achieve this goal, a small amount of labeled samples are used to guide GCN training, which can distill meta knowledge to judge the difference in the neighborhood structure between positive and negative sample pairs. Extensive experiments on Market-1501, DukeMTMC-reID and MSMT17 datasets show that our method outperforms the state-of-the-art approaches.

由于缺乏地面真相标签，无监督人员重新识别（re ID）仍然具有挑战性。现有的方法通常依赖于通过迭代聚类和分类估计的伪标签，不幸的是，它们很容易受到由于估计的聚类数不准确而导致的性能损失。或者，我们提出元成对关系提取（MPRD）方法来估计无监督人员Re-ID样本对的伪标签。具体而言，它由卷积神经网络（CNN）和图卷积网络（GCN）组成，其中，GCN根据CNN提取的当前特征估计样本对的伪标签，CNN通过使用GCN施加的高保真正样本对和负样本对学习更好的特征。为了实现这一目标，使用少量标记样本指导GCN训练，提取元知识判断正负样本对之间邻域结构的差异。在Market-1501、DukeMTMC reID和MSMT17数据集上的大量实验表明，我们的方法优于最先进的方法。

The use of deep 3D point cloud models in safety-critical applications, such as autonomous driving, dictates the need to certify the robustness of these models to real-world transformations. This is technically challenging, as it requires a scalable verifier tailored to point cloud models that handles a wide range of semantic 3D transformations. In this work, we address this challenge and introduce 3DCertify, the first verifier able to certify the robustness of point cloud models. 3DCertify is based on two key insights: (i) a generic relaxation based on first-order Taylor approximations, applicable to any differentiable transformation, and (ii) a precise relaxation for global feature pooling, which is more complex than pointwise activations (e.g., ReLU or sigmoid) but commonly employed in point cloud models. We demonstrate the effectiveness of 3DCertify by performing an extensive evaluation on a wide range of 3D transformations (e.g., rotation, twisting) for both classification and part segmentation tasks. For example, we can certify robustness against rotations by  $\pm 60^\circ$  for 95.7% of point clouds, and our max pool relaxation increases certification by up to 15.6%.

在安全关键应用程序（如自动驾驶）中使用深度3D点云模型，需要证明这些模型对真实世界转换的鲁棒性。这在技术上是具有挑战性的，因为它需要一个可伸缩的验证器，该验证器是为处理广泛的语义3D转换的点云模型定制的。在这项工作中，我们解决了这个挑战，并介绍了3DCertify，第一个能够证明点云模型健壮性的验证器。3DCERTIFIED基于两个关键见解：(i) 基于一阶泰勒近似的通用松弛，适用于任何可微变换；(ii) 全局特征池的精确松弛，这比点态激活（如ReLU或sigmoid）更复杂，但通常用于点云模型。我们通过对分类和零件分割任务的各种3D变换（例如旋转、扭曲）进行广泛评估，证明了3DCERTIFIED的有效性。例如，对于95.7%的点云，我们可以证明其抗旋转 $\pm 60^\circ$ 的鲁棒性，而我们的“最大池松弛”将认证提高了15.6%。

Normalizing flows have recently demonstrated promising results for low-level vision tasks. For image super-resolution (SR), it learns to predict diverse photo-realistic high-resolution (HR) images from the low-resolution (LR) image rather than learning a deterministic mapping. For image rescaling, it achieves high accuracy by jointly modelling the downscaling and upscaling processes. While existing approaches employ specialized techniques for these two tasks, we set out to unify them in a single formulation. In this paper, we propose the hierarchical conditional flow (HCFflow) as a unified framework for image SR and image rescaling. More specifically, HCFflow learns a bijective mapping between HR and LR image pairs by modelling the distribution of the LR image and the rest high-frequency component simultaneously. In particular, the high-frequency component is conditional on the LR image in a hierarchical manner. To further enhance the performance, other losses such as perceptual loss and GAN loss are combined with the commonly used negative log-likelihood loss in training. Extensive experiments on general image SR, face image SR and image rescaling have demonstrated that the proposed HCFflow achieves state-of-the-art performance in terms of both quantitative metrics and visual quality.

规范化流程最近在低级别视觉任务中显示了有希望的结果。对于图像超分辨率 (SR)，它学习从低分辨率 (LR) 图像预测不同的照片真实高分辨率 (HR) 图像，而不是学习确定性映射。对于图像重缩放，它通过联合建模降尺度和升尺度过程来实现高精度。虽然现有的方法对这两项任务使用专门的技术，但我们将它们统一到一个公式中。在本文中，我们提出了分层条件流 (HCFflow) 作为图像SR和图像重缩放的统一框架。更具体地说，HCFflow通过同时建模LR图像和其余高频分量的分布来学习HR和LR图像对之间的双射映射。具体地，高频分量以分层方式取决于LR图像。为了进一步提高性能，将感知损失和GAN损失等其他损失与训练中常用的负对数似然损失相结合。对普通图像SR、人脸图像SR和图像重缩放的大量实验表明，所提出的HCFflow在定量度量和视觉质量方面都达到了最先进的性能。

Visible infrared person re-identification (VI-REID) aims to match pedestrian images between the daytime visible and nighttime infrared camera views. The large cross-modality discrepancies have become the bottleneck which limits the performance of VI-REID. Existing methods mainly focus on capturing cross-modality sharable representations by learning an identity classifier. However, the heterogeneous pedestrian images taken by different spectrum cameras differ significantly in image styles, resulting in inferior discriminability of feature representations. To alleviate the above problem, this paper explores the correlation between two modalities and proposes a novel syncretic modality collaborative learning (SMCL) model to bridge the cross-modality gap. A new modality that incorporates features of heterogeneous images is constructed automatically to steer the generation of modality-invariant representations. Challenge enhanced homogeneity learning (CEHL) and auxiliary distributional similarity learning (ADSL) are integrated to project heterogeneous features on a unified space and enlarge the inter-class disparity, thus strengthening the discriminative power. Extensive experiments on two cross-modality benchmarks demonstrate the effectiveness and superiority of the proposed method. Especially, on SYSU-MM01 dataset, our SMCL model achieves 67.39% rank-1 accuracy and 61.78% mAP, surpassing the cutting-edge works by a large margin.

可见红外人员重新识别 (VI-REID) 旨在在白天可见光和夜间红外摄像机视图之间匹配行人图像。大规模的跨模态差异已经成为限制VI-REID性能的瓶颈。现有的方法主要是通过学习身份分类器来获取跨模态的共享表示。然而，不同光谱相机拍摄的异质行人图像在图像样式上存在显著差异，导致特征表示的可分辨性较差。为了缓解上述问题，本文探讨了两种模式之间的相关性，并提出了一种新的融合模式协作学习 (SMCL) 模型来弥补跨模式的差距。自动构造一种融合异质图像特征的新模态，以指导模态不变表示的生成。将挑战增强同质性学习 (CEHL) 和辅助分布相似性学习 (ADSL) 相结合，在统一空间上投影异质特征，扩大类间差异，增强识别能力。在两个跨模态基准上的大量实验证明了该方法的有效性和优

越性。特别是在SYSU-MM01数据集上，我们的SMCL模型达到了67.39%的秩-1精度和61.78%的mAP，大大超过了最前沿的工作。

Face age transformation aims to synthesize past or future face images by reflecting the age factor on given faces. Ideally, this task should synthesize natural-looking faces across various age groups while maintaining identity. However, most of the existing work has focused on only one of these or is difficult to train while unnatural artifacts still appear. In this work, we propose Re-Aging GAN (RAGAN), a novel single framework considering all the critical factors in age transformation. Our framework achieves state-of-the-art personalized face age transformation by compelling the input identity to perform the self-guidance of the generation process. Specifically, RAGAN can learn the personalized age features by using high-order interactions between given identity and target age. Learned personalized age features are identity information that is recalibrated according to the target age. Hence, such features encompass identity and target age information that provides important clues on how an input identity should be at a certain age. Experimental result shows the lowest FID and KID scores and the highest age recognition accuracy compared to previous methods. The proposed method also demonstrates the visual superiority with fewer artifacts, identity preservation, and natural transformation across various age groups.

人脸年龄变换的目的是通过在给定的人脸上反映年龄因素来合成过去或未来的人脸图像。理想情况下，这项任务应该综合不同年龄组的自然面孔，同时保持身份。然而，现有的大多数工作只关注其中一个，或者在非自然工件仍然出现的情况下很难进行训练。在这项工作中，我们提出了重新老化GAN (RAGAN)，这是一种考虑了年龄转换中所有关键因素的新型单一框架。我们的框架通过强制输入身份执行生成过程的自我指导，实现了最先进的个性化面部年龄转换。具体来说，RAGAN可以通过使用给定身份和目标年龄之间的高阶交互来学习个性化的年龄特征。学到的个性化年龄特征是根据目标年龄重新校准的身份信息。因此，这些特征包含身份和目标年龄信息，这些信息提供了输入身份在特定年龄应该如何的重要线索。实验结果表明，与以前的方法相比，FID和KID分数最低，年龄识别准确率最高。所提出的方法还显示了视觉优势，减少了人工制品、身份保护和跨不同年龄组的自然转换。

Light-field (LF) imaging is appealing to the mobile devices market because of its capability for intuitive post-capture processing. Acquiring LF data with high angular, spatial and temporal resolution poses significant challenges, especially with space constraints preventing bulky optics. At the same time, stereo video capture, now available on many consumer devices, can be interpreted as a sparse LF-capture. We explore the application of small baseline stereo videos for reconstructing high fidelity LF videos. We propose a self-supervised learning-based algorithm for LF video reconstruction from stereo video. The self-supervised LF video reconstruction is guided via the geometric information from the individual stereo pairs and the temporal information from the video sequence. LF estimation is further regularized by a low-rank constraint based on layered LF displays. The proposed self-supervised algorithm facilitates advantages such as post-training fine-tuning on test sequences and variable angular view interpolation and extrapolation. Quantitatively the LF videos show higher fidelity than previously proposed unsupervised approaches for LF reconstruction. We demonstrate our results via LF videos generated from stereo videos acquired from commercially available stereoscopic cameras. Finally, we demonstrate that our reconstructed LF videos allow applications such as post-capture focus control and ROI-based focus tracking for videos.

光场 (LF) 成像因其直观的捕获后处理能力而受到移动设备市场的青睐。获取具有高角度、空间和时间分辨率的LF数据带来了重大挑战，尤其是在空间限制阻止笨重光学元件的情况下。同时，立体声视频捕获，现在可在许多消费类设备上使用，可以解释为稀疏LF捕获。我们探索了小基线立体视频在重建高保真LF视频中的应用。提出了一种基于自监督学习的立体视频LF视频重建算法。自监督LF视频重建通过来自单个立体对的几何信息和来自视频序列的时间信息来引导。LF估计通过基于分层LF显示的低秩约束进一步正则化。所提出的自监督算法具有训练后测试序列微调、可变角度视图插值和外推等优点。定量地，LF视频显示出比先前提出的无监督LF重建方法更高的保真度。我们通过从商用立体摄像机获取的立体视频生成LF视频来演示我们的结果。最后，我们演示了我们重建的LF视频允许应用程序，如捕获后焦点控制和基于RoI的视频焦点跟踪。

In order to train 3D gaze estimators without too many annotations, we propose an unsupervised learning framework, Cross-Encoder, to leverage the unlabeled data to learn suitable representation for gaze estimation. To address the issue that the feature of gaze is always intertwined with the appearance of the eye, Cross-Encoder disentangles the features using a latent-code-swapping mechanism on eye-consistent image pairs and gaze-similar ones. Specifically, each image is encoded as a gaze feature and an eye feature. Cross-Encoder is trained to reconstruct each image in the eye-consistent pair according to its gaze feature and the other's eye feature, but to reconstruct each image in the gaze-similar pair according to its eye feature and the other's gaze feature. Experimental results show the validity of our work. First, using the Cross-Encoder-learned gaze representation, the gaze estimator trained with very few samples outperforms the ones using other unsupervised learning methods, under both within-dataset and cross-dataset protocol. Second, ResNet18 pretrained by Cross-Encoder is competitive with state-of-the-art gaze estimation methods. Third, ablation study shows that Cross-Encoder disentangles the gaze feature and eye feature.

为了在没有太多注释的情况下训练三维凝视估计器，我们提出了一种无监督的学习框架，即交叉编码器，以利用未标记的数据学习合适的凝视估计器表示。为了解决凝视特征总是与眼睛外观交织在一起的问题，交叉编码器在眼睛一致的图像对和凝视相似的图像对上使用潜在的代码交换机制来分离特征。具体地说，每个图像被编码为凝视特征和眼睛特征。交叉编码器被训练成根据眼睛的注视特征和对方的眼睛特征重建眼睛一致对中的每个图像，但根据眼睛特征和对方的注视特征重建注视相似对中的每个图像。实验结果表明了我们工作的有效性。首先，使用交叉编码器学习的凝视表示，在数据集内和跨数据集协议下，使用很少样本训练的凝视估计器优于使用其他无监督学习方法的凝视估计器。其次，交叉编码器预训练的ResNet18与最先进的凝视估计方法具有竞争力。第三，消融研究表明，交叉编码器分离了凝视特征和眼睛特征。

Event cameras can report scene movements as an asynchronous stream of data called the events. Unlike traditional cameras, event cameras have very low latency (microseconds vs milliseconds) very high dynamic range (140dB vs 60 dB), and low power consumption, as they report changes of a scene and not a complete frame. As they report per pixel feature-like events and not the whole intensity frame they are immune to motion blur. However, event cameras require movement between the scene and camera to fire events ,i.e., they have no output when the scene is relatively static. Traditional cameras, however, report the whole frame of pixels at once in fixed intervals but have lower dynamic range and are prone to motion blur in case of rapid movements. We get the best from both worlds and use events and intensity images together in our complementary design and estimate dense disparity from this combination. The proposed end-to-end design combines events and images in a sequential manner and correlates them to estimate dense depth values. Our various experimental settings in real-world and simulated scenarios exploit the superiority of our method in predicting accurate depth values with fine details. We further extend our method to extreme cases of missing the left or right event or stereo pair and also investigate stereo depth estimation with inconsistent dynamic ranges or event thresholds on the left and right pairs

事件摄影机可以将场景移动报告为称为事件的异步数据流。与传统摄影机不同，事件摄影机具有非常低的延迟（微秒vs毫秒）、非常高的动态范围（140dB vs 60dB）和低功耗，因为它们报告场景的变化而不是完整的帧。由于它们报告了每像素的功能，例如事件，而不是整个强度帧，因此它们对运动模糊免疫。但是，事件摄影机需要在场景和摄影机之间移动以触发事件，即，当场景相对静止时，它们没有输出。然而，传统相机以固定的间隔一次报告整个像素帧，但动态范围较低，并且在快速移动时容易出现运动模糊。我们从这两个世界中得到最好的结果，在互补设计中同时使用事件和强度图像，并从这种组合中估计密集的视差。建议的端到端设计以顺序方式组合事件和图像，并将其关联以估计密集深度值。我们在真实世界和模拟场景中的各种实验设置充分利用了我们的方法在预测精确深度值和精细节方面的优势。我们进一步将我们的方法扩展到丢失左、右事件或立体对的极端情况，并且还研究了左、右事件对上具有不一致动态范围或事件阈值的立体深度估计

Do GANs replicate training images? Previous studies have shown that GANs do not seem to replicate training data without significant change in the training procedure. This leads to a series of research on the exact condition needed for GANs to overfit to the training data. Although a number of factors has been theoretically or empirically identified, the effect of dataset size and complexity on GANs replication is still unknown. With empirical evidence from BigGAN and StyleGAN2, on datasets CelebA, Flower and LSUN-bedroom, we show that dataset size and its complexity play an important role in GANs replication and perceptual quality of the generated images. We further quantify this relationship, discovering that replication percentage decays exponentially with respect to dataset size and complexity, with a shared decaying factor across GAN-dataset combinations. Meanwhile, the perceptual image quality follows a U-shape trend w.r.t dataset size. This finding leads to a practical tool for one-shot estimation on minimal dataset size to prevent GAN replication which can be used to guide datasets construction and selection.

GANs复制训练图像吗？先前的研究表明，如果训练过程没有显著变化，GANs似乎不会复制训练数据。这导致了一系列关于GANs过度拟合训练数据所需确切条件的研究。尽管已经从理论或经验上确定了许多因素，但数据集大小和复杂性对GANs复制的影响仍然未知。根据BigGAN和StyleGAN2的经验证据，在CelebA、Flower和LSUN卧室数据集上，我们发现数据集大小及其复杂性在生成图像的GANs复制和感知质量中起着重要作用。我们进一步量化了这种关系，发现复制百分比随数据集大小和复杂性呈指数衰减，在数据集组合中有一个共同的衰减因子。同时，感知图像质量遵循U形趋势w.r.t数据集大小。这一发现为最小数据集大小的一次性估计提供了一个实用工具，以防止GAN复制，从而可用于指导数据集的构建和选择。

The video-based action recognition task has been extensively studied in recent years. In this paper, we study the structural vulnerability of deep learning-based action recognition models against the adversarial attack using the one frame attack that adds an inconspicuous perturbation to only a single frame of a given video clip. Our analysis shows that the models are highly vulnerable against the one frame attack due to their structural properties. Experiments demonstrate high fooling rates and inconspicuous characteristics of the attack. Furthermore, we show that strong universal one frame perturbations can be obtained under various scenarios. Our work raises the serious issue of adversarial vulnerability of the state-of-the-art action recognition models in various perspectives.

近年来，基于视频的动作识别任务得到了广泛的研究。在本文中，我们研究了基于深度学习的动作识别模型在对抗性攻击时的结构脆弱性，使用单帧攻击只在给定视频片段的单帧上添加不明显的扰动。我们的分析表明，由于其结构特性，这些模型非常容易受到单帧攻击。实验证明了高愚弄率和攻击的不明显特征。此外，我们还证明了在各种情况下都可以得到强的普适单帧扰动。我们的工作从不同角度提出了最先进的动作识别模型的对抗脆弱性这一严重问题。

Self-supervised pretraining followed by supervised fine-tuning has seen success in image recognition, especially when labeled examples are scarce, but has received limited attention in medical image analysis. This paper studies the effectiveness of self-supervised learning as a pretraining strategy for medical image classification. We conduct experiments on two distinct tasks: dermatology condition classification from digital camera images and multi-label chest X-ray classification, and demonstrate that self-supervised learning on ImageNet, followed by additional self-supervised learning on unlabeled domain-specific medical images significantly improves the accuracy of medical image classifiers. We introduce a novel Multi-Instance Contrastive Learning (MICLe) method that uses multiple images of the underlying pathology per patient case, when available, to construct more informative positive pairs for self-supervised learning. Combining our contributions, we achieve an improvement of 6.7% in top-1 accuracy and an improvement of 1.1% in mean AUC on dermatology and chest X-ray classification respectively, outperforming strong supervised baselines pretrained on ImageNet. In addition, we show that big self-supervised models are robust to distribution shift and can learn efficiently with a small number of labeled medical images.

自监督预训练和监督微调在图像识别中取得了成功，特别是在标记样本较少的情况下，但在医学图像分析中受到的关注有限。本文研究了自监督学习作为医学图像分类预训练策略的有效性。我们在两个不同的任务上进行了实验：基于数码相机图像的皮肤病病情分类和多标签胸部X射线分类，并证明了ImageNet上的自监督学习，然后对未标记的特定领域医学图像进行额外的自监督学习，显著提高了医学图像分类器的准确性。我们介绍了一种新的多实例对比学习 (MICLe) 方法，该方法使用每个患者病例的多张基本病理图像（如果可用），为自我监督学习构建更多信息的阳性对。结合我们的贡献，我们在皮肤科和胸部X光分类上的top-1准确度和平均AUC分别提高了6.7%和1.1%，优于ImageNet上预训练的强监督基线。此外，我们还证明了大的自监督模型对分布漂移具有鲁棒性，并且能够有效地学习少量标记的医学图像。

Salient object detection identifies objects in an image that grab visual attention. Although contextual features are considered in recent literature, they often fail in real-world complex scenarios. We observe that this is mainly due to two issues: First, most existing datasets consist of simple foregrounds and backgrounds that hardly represent real-life scenarios. Second, current methods only learn contextual features of salient objects, which are insufficient to model high-level semantics for saliency reasoning in complex scenes. To address these problems, we first construct a new large-scale dataset with complex scenes in this paper. We then propose a context-aware learning approach to explicitly exploit the semantic scene contexts. Specifically, two modules are proposed to achieve the goal: 1) a Semantic Scene Context Refinement module to enhance contextual features learned from salient objects with scene context, and 2) a Contextual Instance Transformer to learn contextual relations between objects and scene context. To our knowledge, such high-level semantic contextual information of image scenes is under-explored for saliency detection in the literature. Extensive experiments demonstrate that the proposed approach outperforms state-of-the-art techniques in complex scenarios for saliency detection, and transfers well to other existing datasets. The code and dataset are available at [https://github.com/SirisAvishek/Scene\\_Context\\_Aware\\_Saliency](https://github.com/SirisAvishek/Scene_Context_Aware_Saliency).

突出物体检测识别图像中吸引视觉注意力的物体。虽然在最近的文献中考虑了上下文特征，但在现实世界的复杂场景中它们往往失败。我们观察到，这主要是由于两个问题：第一，大多数现有数据集由简单的前景和背景组成，几乎不表现现实生活场景。其次，现有的方法只学习显著对象的上下文特征，不足以建立高级语义模型。为了解决这些问题，本文首先构造了一个新的具有复杂场景的大规模数据集。然后，我们提出了一种上下文感知学习方法来显式地利用语义场景上下文。具体来说，提出了两个模块来实现这一目标：1) 语义场景上下文细化模块，用于增强从具有场景上下文的显著对象中学习的上下文特征；2) 上下文实例转换器，用于学习对象与场景上下文之间的上下文关系。据我们所知，文献中对图像场景的这种高级语义上下文信息进行显著性检测的探索不足。大量实验表明，在复杂的显著性检测场景中，该方法的性能优于最新的显著性检测技术，并能很好地传输到其他现有数据集。代码和数据集可在[https://github.com/SirisAvishek/Scene\\_Context\\_Aware\\_Saliency](https://github.com/SirisAvishek/Scene_Context_Aware_Saliency)。

Deep learning based methods, especially convolutional neural networks (CNNs) have been successfully applied in the field of single image super-resolution (SISR). To obtain better fidelity and visual quality, most of existing networks are of heavy design with massive computation. However, the computation resources of modern mobile devices are limited, which cannot easily support the expensive cost. To this end, this paper explores a novel frequency-aware dynamic network for dividing the input into multiple parts according to its coefficients in the discrete cosine transform (DCT) domain. In practice, the high-frequency part will be processed using expensive operations and the lower-frequency part is assigned with cheap operations to relieve the computation burden. Since pixels or image patches belong to low-frequency areas contain relatively few textural details, this dynamic network will not affect the quality of resulting super-resolution images. In addition, we embed predictors into the proposed dynamic network to end-to-end fine-tune the handcrafted frequency-aware masks. Extensive experiments conducted on benchmark SISR models and datasets show that the frequency-aware dynamic network can be employed for various SISR neural architectures to obtain the better tradeoff between visual quality and computational complexity. For instance, we can reduce the FLOPs of SR models by approximate 50% while preserving the state-of-the-art SISR performance.

基于深度学习的方法，特别是卷积神经网络（CNN）已经成功地应用于单幅图像超分辨率（SISR）领域。为了获得更好的逼真度和视觉质量，大多数现有网络都采用了繁重的设计和大量的计算。然而，现代移动设备的计算资源有限，难以承受昂贵的成本。为此，本文探索了一种新的频率感知动态网络，用于根据离散余弦变换（DCT）域中的系数将输入划分为多个部分。在实践中，高频部分将使用昂贵的运

算进行处理，而低频部分将分配廉价的运算以减轻计算负担。由于像素或图像块属于低频区域，包含的纹理细节相对较少，因此该动态网络不会影响生成的超分辨率图像的质量。此外，我们将预测器嵌入到所提出的动态网络中，以端到端微调手工制作的频率感知掩码。在基准SISR模型和数据集上进行的大量实验表明，频率感知动态网络可用于各种SISR神经结构，以在视觉质量和计算复杂度之间获得更好的折衷。例如，我们可以将SR模型的失败率降低约50%，同时保持最先进的SISR性能。

multi-label image recognition is a challenging computer vision task of practical use. Progresses in this area, however, are often characterized by complicated methods, heavy computations, and lack of intuitive explanations. To effectively capture different spatial regions occupied by objects from different categories, we propose an embarrassingly simple module, named class-specific residual attention (CSRA). CSRA generates class-specific features for every category by proposing a simple spatial attention score, and then combines it with the class-agnostic average pooling feature. CSRA achieves state-of-the-art results on multilabel recognition, and at the same time is much simpler than them. Furthermore, with only 4 lines of code, CSRA also leads to consistent improvement across many diverse pretrained models and datasets without any extra training. CSRA is both easy to implement and light in computations, which also enjoys intuitive explanations and visualizations.

多标签图像识别是一项具有挑战性的计算机视觉实际应用任务。然而，这一领域的进展往往以复杂的方法、繁重的计算和缺乏直观的解释为特点。为了有效地捕获不同类别的对象所占据的不同空间区域，我们提出了一个令人尴尬的简单模块，名为类特定剩余注意（CSRA）。CSRA通过提出一个简单的空间注意分数为每个类别生成特定于类的特征，然后将其与不可知类的平均池特征相结合。CSRA在多标签识别方面实现了最先进的结果，同时比它们简单得多。此外，只有4行代码，CSRA还可以在许多不同的预训练模型和数据集上实现一致的改进，而无需任何额外培训。CSRA不仅易于实现，而且计算量小，还具有直观的解释和可视化功能。

while the untargeted black-box transferability of adversarial perturbations has been extensively studied before, changing an unseen model's decisions to a specific 'targeted' class remains a challenging feat. In this paper, we propose a new generative approach for highly transferable targeted perturbations (\textit{ours}). We note that the existing methods are less suitable for this task due to their reliance on class-boundary information that changes from one model to another, thus reducing transferability. In contrast, our approach matches perturbed image 'distribution' with that of the target class, leading to high targeted transferability rates. To this end, we propose a new objective function that not only aligns the global distributions of source and target images, but also matches the local neighbourhood structure between the two domains. Based on the proposed objective, we train a generator function that can adaptively synthesize perturbations specific to a given input. Our generative approach is independent of the source or target domain labels, while consistently performs well against state-of-the-art methods on a wide range of attack settings. As an example, we achieve 32.63\% target transferability from (an adversarially weak) VGG19\\_BN to (a strong) WideResNet on ImageNet val. set, which is 4x higher than the previous best generative attack and 16x better than instance-specific iterative attack.

虽然对抗性干扰的非目标黑盒可转移性之前已经被广泛研究过，但将一个看不见的模型的决定改变为一个特定的“目标”类别仍然是一项具有挑战性的壮举。在本文中，我们提出了一种新的生成方法来处理高度可转移的目标扰动（\textit{ours}）。我们注意到，现有的方法不太适合这个任务，因为它们依赖于从一个模型到另一个模型的类边界信息，从而降低了可转移性。相比之下，我们的方法将受干扰的图像“分布”与目标类的图像“分布”相匹配，从而获得高的目标可转移率。为此，我们提出了一个新的目标函数，不仅使源图像和目标图像的全局分布保持一致，而且使两个域之间的局部邻域结构相匹配。基于所提出的目标，我们训练一个生成函数，该生成函数能够自适应地合成特定于给定输入的扰动。我们的生成方法独

立于源域或目标域标签，同时在广泛的攻击设置下，与最先进的方法相比，始终表现良好。例如，我们在ImageNet val.set上实现了32.63%的目标可从（敌对较弱的）VGG19\_BN转移到（较强的）WideResNet，这比之前的最佳生成攻击高4倍，比特定实例的迭代攻击高16倍。

Feature pyramids have been proven powerful in image understanding tasks that require multi-scale features. State-of-the-art methods for multi-scale feature learning focus on performing feature interactions across space and scales using neural networks with a fixed topology. In this paper, we propose graph feature pyramid networks that are capable of adapting their topological structures to varying intrinsic image structures, and supporting simultaneous feature interactions across all scales. We first define an image specific superpixel hierarchy for each input image to represent its intrinsic image structures. The graph feature pyramid network inherits its structure from this superpixel hierarchy. Contextual and hierarchical layers are designed to achieve feature interactions within the same scale and across different scales, respectively. To make these layers more powerful, we introduce two types of local channel attention for graph neural networks by generalizing global channel attention for convolutional neural networks. The proposed graph feature pyramid network can enhance the multiscale features from a convolutional feature pyramid network. We evaluate our graph feature pyramid network in the object detection task by integrating it into the Faster RCNN algorithm. The modified algorithm not only outperforms previous state-of-the-art feature pyramid based methods with a clear margin but also outperforms other popular detection methods on both MS-COCO 2017 validation and test datasets.

在需要多尺度特征的图像理解任务中，特征金字塔已经被证明是强大的。用于多尺度特征学习的最新方法侧重于使用具有固定拓扑结构的神经网络跨空间和尺度执行特征交互。在本文中，我们提出了一种图特征金字塔网络，它能够使其拓扑结构适应不同的内在图像结构，并支持所有尺度上的同时特征交互。我们首先为每个输入图像定义一个特定于图像的超像素层次，以表示其固有的图像结构。图形特征金字塔网络继承了这种超像素层次结构的结构。上下文层和层次层旨在分别实现同一尺度内和不同尺度之间的特征交互。为了使这些层更强大，我们通过对卷积神经网络的全局通道注意进行推广，为图神经网络引入了两种类型的局部通道注意。所提出的图特征金字塔网络可以增强卷积特征金字塔网络的多尺度特征。我们通过将图形特征金字塔网络集成到更快的RCNN算法中来评估其在目标检测任务中的作用。改进后的算法不仅优于以前基于特征金字塔的最新方法，具有明显的优势，而且在MS-COCO 2017验证和测试数据集上也优于其他流行的检测方法。

Neuromorphic vision sensor is a new bio-inspired imaging paradigm that emerged in recent years, which continuously sensing luminance intensity and firing asynchronous spikes (events) with high temporal resolution. Typically, there are two types of neuromorphic vision sensors, namely dynamic vision sensor (DVS) and spike camera. From the perspective of bio-inspired sampling, DVS only perceives movement by imitating the retinal periphery, while the spike camera was developed to perceive fine textures by simulating the fovea. It is meaningful to explore how to combine two types of neuromorphic cameras to reconstruct high quality image like human vision. In this paper, we propose a NeuSpike-Net to learn both the high dynamic range and high motion sensitivity of DVS and the full texture sampling of spike camera to achieve high-speed and high dynamic image reconstruction. We propose a novel representation to effectively extract the temporal information of spike and event data. By introducing the feature fusion module, the two types of neuromorphic data achieve complementary to each other. The experimental results on the simulated and real datasets demonstrate that the proposed approach is effective to reconstruct high-speed and high dynamic range images via the combination of spike and event data.

神经形态视觉传感器是近年来出现的一种新的仿生成像模式，它以高时间分辨率连续感知亮度强度并发射异步尖峰（事件）。通常，有两种神经形态视觉传感器，即动态视觉传感器（DVS）和spike摄像机。从仿生取样的角度来看，DVS仅通过模仿视网膜周边来感知运动，而spike摄像头则通过模拟中央凹来感知精细纹理。探索如何将两种神经形态摄像机结合起来，重建出类似人类视觉的高质量图像，具有重要意义。在本文中，我们提出了一个Neuspeak网络来学习DVS的高动态范围和高运动灵敏度，以及spike摄像机的全纹理采样，以实现高速和高动态图像重建。我们提出了一种新的表示方法来有效地提取尖峰和事件数据的时间信息。通过引入特征融合模块，两类神经形态数据实现了互补。在模拟数据集和真实数据集上的实验结果表明，该方法能够有效地结合峰值和事件数据重建高速、高动态范围的图像。

For semantic segmentation, label probabilities are often uncalibrated as they are typically only the by-product of a segmentation task. Intersection over Union (IoU) and Dice score are often used as criteria for segmentation success, while metrics related to label probabilities are not often explored. However, probability calibration approaches have been studied, which match probability outputs with experimentally observed errors. These approaches mainly focus on classification tasks, but not on semantic segmentation. Thus, we propose a learning-based calibration method that focuses on multi-label semantic segmentation. Specifically, we adopt a convolutional neural network to predict local temperature values for probability calibration. One advantage of our approach is that it does not change prediction accuracy, hence allowing for calibration as a post-processing step. Experiments on the COCO, CamVid, and LPBA40 datasets demonstrate improved calibration performance for a range of different metrics. We also demonstrate the good performance of our method for multi-atlas brain segmentation from magnetic resonance images.

对于语义分割，标签概率通常是未经校准的，因为它们通常只是分割任务的副产品。联合交集 (IoU) 和 骰子分数通常被用作分割成功的标准，而与标签概率相关的度量通常不被探讨。然而，已经研究了概率校准方法，该方法将概率输出与实验观测误差相匹配。这些方法主要关注分类任务，而不是语义分割。因此，我们提出了一种基于学习的多标签语义分割校正方法。具体而言，我们采用卷积神经网络预测局部温度值，以进行概率校准。我们的方法的一个优点是它不会改变预测精度，因此允许作为后处理步骤进行校准。在COCO、CamVid和LPBA40数据集上的实验表明，对于一系列不同的度量，校准性能得到了改进。我们还展示了我们的方法从磁共振图像多图谱脑分割的良好性能。

This work addresses the problem of discovering, in an unsupervised manner, interpretable paths in the latent space of pretrained GANs, so as to provide an intuitive and easy way of controlling the underlying generative factors. In doing so, it addresses some of the limitations of the state-of-the-art works, namely, a) that they discover directions that are independent of the latent code, i.e., paths that are linear, and b) that their evaluation relies either on visual inspection or on laborious human labeling. More specifically, we propose to learn non-linear warpings on the latent space, each one parametrized by a set of RBF-based latent space warping functions, and where each warping gives rise to a family of non-linear paths via the gradient of the function. Building on the work of Voynov and Babenko, that discovers linear paths, we optimize the trainable parameters of the set of RBFs, so as that images that are generated by codes along different paths, are easily distinguishable by a discriminator network. This leads to easily distinguishable image transformations, such as pose and facial expressions in facial images. We show that linear paths can be derived as a special case of our method, and show experimentally that non-linear paths in the latent space lead to steeper, more disentangled and interpretable changes in the image space than in state-of-the art methods, both qualitatively and quantitatively. We make the code and the pretrained models publicly available at: <https://github.com/chi0tzp/WarpedGANSpace>.

这项工作解决了在无监督的情况下，在预训练的GANs的潜在空间中发现可解释路径的问题，从而提供了一种直观且简单的方法来控制潜在的生成因素。在这样做的过程中，它解决了最先进作品的一些局限性，即，a) 他们发现了独立于潜在代码的方向，即线性路径，以及b) 他们的评估依赖于视觉检查或费力的人类标记。更具体地说，我们建议学习潜在空间上的非线性扭曲，每个扭曲由一组基于RBF的潜在空间扭曲函数参数化，其中每个扭曲通过函数的梯度产生一系列非线性路径。在Voynov和Babenko发现线性路径的工作的基础上，我们优化了RBF集的可训练参数，从而使代码沿不同路径生成的图像易于通过鉴别器网络识别。这将导致易于区分的图像变换，例如面部图像中的姿势和面部表情。我们证明了线性路径可以作为我们方法的一个特例导出，并且在实验上表明，与最先进的方法相比，潜在空间中的非线性路径在定性和定量上会导致图像空间中更陡峭、更混乱和更可解释的变化。我们在以下网站上公开代码和预培训模型：<https://github.com/chi0tzc/WarpedGANSpace>.

Transformers are increasingly dominating multi-modal reasoning tasks, such as visual question answering, achieving state-of-the-art results thanks to their ability to contextualize information using the self-attention and co-attention mechanisms. These attention modules also play a role in other computer vision tasks including object detection and image segmentation. Unlike Transformers that only use self-attention, Transformers with co-attention require to consider multiple attention maps in parallel in order to highlight the information that is relevant to the prediction in the model's input. In this work, we propose the first method to explain prediction by any Transformer-based architecture, including bi-modal Transformers and Transformers with co-attentions. We provide generic solutions and apply these to the three most commonly used of these architectures: (i) pure self-attention, (ii) self-attention combined with co-attention, and (iii) encoder-decoder attention. We show that our method is superior to all existing methods which are adapted from single modality explainability.

变形金刚越来越多地主导着多模态推理任务，如视觉问答，由于它们能够利用自我注意和共同注意机制将信息情境化，从而获得最先进的结果。这些注意模块也在其他计算机视觉任务中发挥作用，包括目标检测和图像分割。与仅使用自我关注的变压器不同，具有共同关注的变压器需要并行考虑多个关注图，以便突出与模型输入中的预测相关的信息。在这项工作中，我们提出了第一种方法来解释任何基于变压器的架构的预测，包括双模变压器和具有共同关注的变压器。我们提供了通用的解决方案，并将其应用于三种最常用的体系结构：(i) 纯自我注意，(ii) 结合共同注意的自我注意，以及(iii) 编码器-解码器注意。我们证明了我们的方法优于所有现有的基于单模态解释的方法。

Recently, deep learning-based image enhancement algorithms achieved state-of-the-art (SOTA) performance on several publicly available datasets. However, most existing methods fail to meet practical requirements either for visual perception or for computation efficiency, especially for high-resolution images. In this paper, we propose a novel real-time image enhancer via learnable spatial-aware 3-dimentional lookup tables(3D LUTs), which well considers global scenario and local spatial information. Specifically, we introduce a light weight two-head weight predictor that has two outputs. One is a 1D weight vector used for image-level scenario adaptation, the other is a 3D weight map aimed for pixel-wise category fusion. We learn the spatial-aware 3D LUTs and fuse them according to the aforementioned weights in an end-to-end manner. The fused LUT is then used to transform the source image into the target tone in an efficient way. Extensive results show that our model outperforms SOTA image enhancement methods on public datasets both subjectively and objectively, and that our model only takes about 4ms to process a 4K resolution image on one NVIDIA V100 GPU.

最近，基于深度学习的图像增强算法在几个公开的数据集上实现了最先进的（SOTA）性能。然而，大多数现有的方法都不能满足视觉感知或计算效率的实际要求，特别是对于高分辨率图像。在本文中，我们提出了一种新的实时图像增强器，该增强器通过可学习的空间感知三维查找表（3D LUT）实现，充分考虑了全局场景和局部空间信息。具体地说，我们介绍了一种具有两个输出的轻型双头权重预测器。一个是用于图像级场景自适应的一维权重向量，另一个是用于像素级类别融合的三维权重映射。我们学习空间感知3D LUT，并根据上述权重以端到端的方式对其进行融合。然后使用融合的LUT以有效的方式将源图像转换为目标音调。大量结果表明，我们的模型在主观和客观上都优于SOTA图像增强方法，并且我们的模型在一个NVIDIA V100 GPU上处理4K分辨率的图像只需要大约4ms。

Image and video enhancement such as color constancy, low light enhancement, and tone mapping on smartphones is challenging because high-quality images should be achieved efficiently with a limited resource budget. Unlike prior works that either used very deep CNNs or large Transformer models, we propose a \underline{s} eman\underline{t} ic-\underline{w} a\underline{r} e lightweight Transformer, termed STAR, for real-time image enhancement. STAR is formulated to capture long-range dependencies between image patches, which naturally and implicitly captures the semantic relationships of different regions in an image. STAR is a general architecture that can be easily adapted to different image enhancement tasks. Extensive experiments show that STAR can effectively boost the quality and efficiency of many tasks such as illumination enhancement, auto white balance, and photo retouching, which are indispensable components for image processing on smartphones. For example, STAR reduces model complexity and improves image quality compared to the recent state-of-the-art [??] on the MIT-Adobe FiveK dataset [??] (i.e., 1.8dB PSNR improvements with 25% parameters and 13% float operations.)

智能手机上的图像和视频增强（如颜色恒定性、弱光增强和色调映射）具有挑战性，因为高质量图像应该在有限的资源预算下高效实现。与以前使用非常深的CNN或大型变压器模型的工作不同，我们提出了一种用于实时图像增强的\underline{s} eman\underline{t} ic-\underline{w} a\underline{r} e轻型变压器，称为STAR。STAR用于捕获图像块之间的长期依赖关系，这自然且隐式地捕获图像中不同区域的语义关系。STAR是一种通用架构，可以轻松地适应不同的图像增强任务。大量实验表明，STAR可以有效提高许多任务的质量和效率，如照明增强、自动白平衡和照片修饰，这些都是智能手机图像处理不可或缺的组件。例如，与最新技术相比，STAR降低了模型复杂性并提高了图像质量[? ? ]在麻省理工学院Adobe FiveK数据集[? ? ]（即，在25%的参数和13%的浮动操作下，1.8dB的峰值信噪比提高。）

Classification and regression are two pillars of object detectors. In most CNN-based detectors, these two pillars are optimized independently. Without direct interactions between them, the classification loss and the regression loss can not be optimized synchronously toward the optimal direction in the training phase. This clearly leads to lots of inconsistent predictions with high classification score but low localization accuracy or low classification score but high localization accuracy in the inference phase, especially for the objects of irregular shape and occlusion, which severely hurts the detection performance of existing detectors after NMS. To reconcile prediction consistency for balanced object detection, we propose a Harmonic loss to harmonize the optimization of classification branch and localization branch. The Harmonic loss enables these two branches to supervise and promote each other during training, thereby producing consistent predictions with high co-occurrence of top classification and localization in the inference phase. Furthermore, in order to prevent the localization loss from being dominated by outliers during training phase, a Harmonic IoU loss is proposed to harmonize the weight of the localization loss of different IoU-level samples. Comprehensive experiments on benchmarks PASCAL VOC and MS COCO demonstrate the generality and effectiveness of our model for facilitating existing object detectors to state-of-the-art accuracy.

分类和回归是目标检测器的两大支柱。在大多数基于CNN的检测器中，这两个支柱是独立优化的。如果没有它们之间的直接交互作用，分类损失和回归损失就无法在训练阶段朝着最优方向同步优化。这显然会导致在推理阶段出现大量分类分数高但定位精度低或分类分数低但定位精度高的不一致预测，尤其是对于形状不规则和遮挡的对象，这严重影响了NMS后现有检测器的检测性能。为了协调平衡目标检测的预测一致性，我们提出了一种谐波损耗来协调分类分支和定位分支的优化。谐波损耗使这两个分支能够在训练期间相互监督和促进，从而产生一致的预测，在推理阶段，顶级分类和本地化高度共存。此外，为了防止在训练阶段局部化损失被异常值控制，提出了一种调和IoU损失来协调不同IoU水平样本局部化损失的权重。在基准PASCAL VOC和MS COCO上进行的综合实验证明了我们的模型的通用性和有效性，可以使现有的目标检测器达到最先进的精度。

In this paper, we introduce a deep multi-view stereo (MVS) system that jointly predicts depths, surface normals and per-view confidence maps. The key to our approach is a novel solver that iteratively solves for per-view depth map and normal map by optimizing an energy potential based upon the local planar assumption. Specifically, the algorithm updates depth map by propagating from neighboring pixels with slanted planes, and updates normal map with local probabilistic plane fitting. Both two steps are monitored by a customized confidence map. This confidence-based solver is not only effective as a post-processing tool for plane based depth refinement and completion, but also differentiable such that it can be efficiently integrated into deep learning pipelines. Our multi-view stereo system employs multiple optimization steps of the solver over the initial prediction of depths and surface normals. The whole system can be trained end-to-end, decoupling the challenging problem of matching pixels within poorly textured regions from the cost volume based neural network. Experimental results on ScanNet and RGB-D Scenes V2 demonstrate state-of-the-art performance of the proposed deep MVS system on multi-view depth estimation, with our proposed solver consistently improving the depth quality over both conventional and deep learning based MVS pipelines.

在本文中，我们介绍了一个深度多视图立体（MVS）系统，该系统可以联合预测深度、表面法线和每视图置信度贴图。我们的方法的关键是一个新的求解器，它通过基于局部平面假设优化能量势来迭代求解逐视图深度贴图和法线贴图。具体来说，该算法通过从倾斜平面的相邻像素传播来更新深度图，并通过局部概率平面拟合来更新法线图。这两个步骤都由自定义的置信度图监控。这种基于置信度的求解器不仅可以作为基于平面的深度细化和完成的后处理工具，而且可以进行微分，从而可以有效地集成到深度学习管道中。我们的多视图立体系统在深度和曲面法线的初始预测上采用了解算器的多个优化步骤。整个系统可以进行端到端的训练，将纹理较差区域内像素匹配的挑战性问题与基于成本-体积的神经网络解耦。在ScanNet和RGB-D Scenes V2上的实验结果表明，所提出的深度MVS系统在多视图深度估计方面具有最先进的性能，与传统的和基于深度学习的MVS管道相比，我们提出的解算器持续提高了深度质量。

Our objective in this work is video-text retrieval - in particular a joint embedding that enables efficient text-to-video retrieval. The challenges in this area include the design of the visual architecture and the nature of the training data, in that the available large scale video-text training datasets, such as HowTo100M, are noisy and hence competitive performance is achieved only at scale through large amounts of compute. We address both these challenges in this paper. We propose an end-to-end trainable model that is designed to take advantage of both large-scale image and video captioning datasets. Our model is an adaptation and extension of the recent ViT and Timesformer architectures, and consists of attention in both space and time. The model is flexible and can be trained on both image and video text datasets, either independently or in conjunction. It is trained with a curriculum learning schedule that begins by treating images as 'frozen' snapshots of video, and then gradually learns to attend to increasing temporal context when trained on video datasets. We also provide a new video-text pretraining dataset WebVid-2M, comprised of over two million videos with weak captions scraped from the internet. Despite training on datasets that are an order of magnitude smaller, we show that this approach yields state-of-the-art results on standard downstream video-retrieval benchmarks including MSR-VTT, DiDeMo and MSVD.

我们在这项工作中的目标是视频文本检索-特别是一个联合嵌入，使高效的文本到视频检索。这一领域的挑战包括视觉体系结构的设计和训练数据的性质，因为可用的大规模视频文本训练数据集（如HowTo100M）是有噪声的，因此只有通过大量计算才能在规模上实现有竞争力的性能。我们在本文中解决了这两个挑战。我们提出了一种端到端的可训练模型，旨在利用大规模图像和视频字幕数据集。我们的模型是对最近的ViT和时间转换器架构的改编和扩展，包括对空间和时间的关注。该模型是灵活的，可以在图像和视频文本数据集上单独或联合进行训练。它通过课程学习计划进行培训，该计划首先将图像视为“冻结”的视频快照，然后在视频数据集上进行培训时，逐渐学习关注不断增加的时间上下文。我们还提供了一个新的视频文本预训练数据集WebVid-2M，它由200多万个从互联网上截取的弱标题视频组成。尽管对较小数量级的数据集进行了培训，但我们表明，这种方法在标准下游视频检索基准（包括MSR-VTT、DiDeMo和MSVD）上产生了最先进的结果。

Quantization is a widely used technique to compress and accelerate deep neural networks. However, conventional quantization methods use the same bit-width for all (or most of) the layers, which often suffer significant accuracy degradation in the ultra-low precision regime and ignore the fact that emergent hardware accelerators begin to support mixed-precision computation. Consequently, we present a novel and principled framework to solve the mixed-precision quantization problem in this paper. Briefly speaking, we first formulate the mixed-precision quantization as a discrete constrained optimization problem. Then, to make the optimization tractable, we approximate the objective function with second-order Taylor expansion and propose an efficient approach to compute its Hessian matrix. Finally, based on the above simplification, we show that the original problem can be reformulated as a Multiple Choice Knapsack Problem (MCKP) and propose a greedy search algorithm to solve it efficiently. Compared with existing mixed-precision quantization works, our method is derived in a principled way and much more computationally efficient. Moreover, extensive experiments conducted on the ImageNet dataset and various kinds of network architectures also demonstrate its superiority over existing uniform and mixed-precision quantization approaches.

量化是一种广泛用于压缩和加速深层神经网络的技术。然而，传统的量化方法对所有（或大部分）层使用相同的比特宽度，这在超低精度区域中通常会遭受显著的精度降低，并且忽略了新兴硬件加速器开始支持混合精度计算的事实。因此，本文提出了一种新的、有原则的框架来解决混合精度量化问题。简单地说，我们首先将混合精度量化描述为一个离散约束优化问题。然后，为了使优化易于处理，我们用二阶泰勒展开逼近目标函数，并提出了一种计算其Hessian矩阵的有效方法。最后，在上述简化的基础上

上，我们证明了原问题可以转化为一个多选择背包问题 (MCKP)，并提出了一种贪婪搜索算法来有效地解决该问题。与现有的混合精度量化方法相比，该方法具有原则性，计算效率更高。此外，在 ImageNet 数据集和各种网络结构上进行的大量实验也证明了其优于现有的均匀和混合精度量化方法。

We present a novel approach to reference-based super-resolution (RefSR) with the focus on dual-camera super-resolution (DCSR), which utilizes reference images for high-quality and high-fidelity results. Our proposed method generalizes the standard patch-based feature matching with spatial alignment operations. We further explore the dual-camera super-resolution that is one promising application of RefSR, and build a dataset that consists of 146 image pairs from the main and telephoto cameras in a smartphone. To bridge the domain gaps between real-world images and the training images, we propose a self-supervised domain adaptation strategy for real-world images. Extensive experiments on our dataset and a public benchmark demonstrate clear improvement achieved by our method over state of the art in both quantitative evaluation and visual comparisons.

我们提出了一种新的基于参考的超分辨率 (RefSR) 方法，重点是双摄像机超分辨率 (DCSR)，它利用参考图像获得高质量和高保真的结果。我们提出的方法推广了标准的基于面片的特征匹配和空间对齐操作。我们进一步探索了RefSR的一个有前途的应用——双摄像机超分辨率，并构建了一个数据集，该数据集由智能手机中主摄像机和长焦摄像机的146个图像对组成。为了弥补真实世界图像和训练图像之间的领域差距，我们提出了一种针对真实世界图像的自监督领域自适应策略。在我们的数据集和一个公共基准上进行的大量实验表明，我们的方法在定量评估和视觉比较方面都比最先进的方法有明显的改进。

Fine-tuning from pre-trained ImageNet models has been a simple, effective, and popular approach for various computer vision tasks. The common practice of fine-tuning is to adopt a default hyperparameter setting with a fixed pre-trained model, while both of them are not optimized for specific tasks and time constraints. Moreover, in cloud computing or GPU clusters where the tasks arrive sequentially in a stream, faster online fine-tuning is a more desired and realistic strategy for saving money, energy consumption, and CO<sub>2</sub> emission. In this paper, we propose a joint Neural Architecture Search and Online Adaption framework named NASOA towards a faster task-oriented fine-tuning upon the request of users. Specifically, NASOA first adopts an offline NAS to identify a group of training-efficient networks to form a pretrained model zoo. We propose a novel joint block and macro level search space to enable a flexible and efficient search. Then, by estimating fine-tuning performance via an adaptive model by accumulating experience from the past tasks, an online schedule generator is proposed to pick up the most suitable model and generate a personalized training regime with respect to each desired task in a one-shot fashion. The resulting model zoo is more training efficient than SOTA NAS models, e.g. 6x faster than RegNetY-16GF, and 1.7x faster than EfficientNetB3. Experiments on multiple datasets also show that NASOA achieves much better fine-tuning results, i.e. improving around 2.1% accuracy than the best performance in RegNet series under various time constraints and tasks; 40x faster compared to the BOHB method.

对于各种计算机视觉任务，从预先训练好的ImageNet模型进行微调是一种简单、有效且流行的方法。微调的常见做法是采用带有固定预训练模型的默认超参数设置，而这两种设置均未针对特定任务和时间约束进行优化。此外，在任务顺序到达的云计算或GPU集群中，更快的在线微调对于节约资金、能源消耗和二氧化碳排放来说是一种更理想、更现实的策略。在本文中，我们提出了一个名为NASOA的联合神经结构搜索和在线自适应框架，以根据用户的要求进行更快的面向任务的微调。具体来说，NASOA首先采用离线NAS来识别一组训练有效的网络，以形成一个预训练的模型动物园。我们提出了一种新的联合块和宏级搜索空间，以实现灵活高效的搜索。然后，通过从过去的任务中积累经验，通过自适应模型来估计微调性能，提出了一种在线调度生成器，用于选择最合适的模型，并以一次性方式针对每个期望任务生成个性化的训练机制。由此产生的模型zoo比SOTA NAS模型训练效率更高，例如比RegNetY-16GF快

6倍，比EfficientNetB3快1.7倍。在多个数据集上的实验还表明，NASOA实现了更好的微调结果，即在各种时间约束和任务下，比RegNet系列中的最佳性能提高了约2.1%的精度；比BOHB方法快40倍。

3D point cloud understanding has made great progress in recent years. However, one major bottleneck is the scarcity of annotated real datasets, especially compared to 2D object detection tasks, since a large amount of labor is involved in annotating the real scans of a scene. A promising solution to this problem is to make better use of the synthetic dataset, which consists of CAD object models, to boost the learning on real datasets. This can be achieved by the pre-training and fine-tuning procedure. However, recent work on 3D pre-training exhibits failure when transfer features learned on synthetic objects to other real-world applications. In this work, we put forward a new method called RandomRooms to accomplish this objective. In particular, we propose to generate random layouts of a scene by making use of the objects in the synthetic CAD dataset and learn the 3D scene representation by applying object-level contrastive learning on two random scenes generated from the same set of synthetic objects. The model pre-trained in this way can serve as a better initialization when later fine-tuning on the 3D object detection task. Empirically, we show consistent improvement in downstream 3D detection tasks on several base models, especially when less training data are used, which strongly demonstrates the effectiveness and generalization of our method. Benefiting from the rich semantic knowledge and diverse objects from synthetic data, our method establishes the new state-of-the-art on widely-used 3D detection benchmarks ScanNetV2 and SUN RGB-D. We expect our attempt to provide a new perspective for bridging object and scene-level 3D understanding.

三维点云理解近年来取得了很大进展。然而，一个主要的瓶颈是缺少带注释的真实数据集，特别是与2D对象检测任务相比，因为注释场景的真实扫描需要大量的劳动力。解决这个问题的一个很有希望的方法是更好地利用由CAD对象模型组成的合成数据集，以促进对真实数据集的学习。这可以通过预培训和微调程序实现。然而，最近关于3D预训练的工作在将合成对象上学习到的特征转移到其他实际应用中时显示出失败。在这项工作中，我们提出了一种称为随机房间的新方法来实现这一目标。特别是，我们建议利用合成CAD数据集中的对象生成场景的随机布局，并通过对同一组合成对象生成的两个随机场景应用对象级对比学习来学习3D场景表示。在以后对3D对象检测任务进行微调时，以这种方式预先训练的模型可以作为更好的初始化。从经验上看，我们在几个基础模型上显示了下游3D检测任务的持续改进，特别是当使用较少的训练数据时，这有力地证明了我们方法的有效性和泛化性。得益于合成数据中丰富的语义知识和多样的对象，我们的方法在广泛使用的3D检测基准ScanNetV2和SUN RGB-D上建立了最新的技术水平。我们期望我们的尝试为桥接对象和场景级3D理解提供新的视角。

Adversarial data examples have drawn significant attention from the machine learning and security communities. A line of work on tackling adversarial examples is certified robustness via randomized smoothing that can provide a theoretical robustness guarantee. However, such a mechanism usually uses floating-point arithmetic for calculations in inference and requires large memory footprints and daunting computational costs. These defensive models cannot run efficiently on edge devices nor be deployed on integer-only logical units such as Turing Tensor Cores or integer-only ARM processors. To overcome these challenges, we propose an integer randomized smoothing approach with quantization to convert any classifier into a new smoothed classifier, which uses integer-only arithmetic for certified robustness against adversarial perturbations. We prove a tight robustness guarantee under L2-norm for the proposed approach. We show our approach can obtain a comparable accuracy and 4x 5x speedup over floating-point arithmetic certified robust methods on general-purpose CPUs and mobile devices on two distinct datasets (CIFAR-10 and Caltech-101).

对抗性数据示例引起了机器学习和安全社区的极大关注。处理对抗性示例的一系列工作通过随机平滑验证了鲁棒性，可提供理论上的鲁棒性保证。然而，这种机制通常使用浮点算法进行推理计算，并且需要大量内存占用和令人望而生畏的计算成本。这些防御模型既不能在边缘设备上高效运行，也不能部署在纯整数逻辑单元（如图灵张量核或纯整数ARM处理器）上。为了克服这些挑战，我们提出了一种带量化的整数随机化平滑方法，将任何分类器转换为一个新的平滑分类器，该方法使用仅整数算法来证明对对抗性扰动的鲁棒性。我们证明了该方法在L2范数下的严格鲁棒性保证。我们表明，我们的方法可以在两个不同数据集（CIFAR-10和Caltech-101）上的通用CPU和移动设备上获得与浮点算术认证稳健方法相当的精度和4x 5x的加速比。

within Convolutional Neural Network (CNN), the convolution operations are good at extracting local features but experience difficulty to capture global representations. Within visual transformer, the cascaded self-attention modules can capture long-distance feature dependencies but unfortunately deteriorate local feature details. In this paper, we propose a hybrid network structure, termed Conformer, to take advantage of convolutional operations and self-attention mechanisms for enhanced representation learning. Conformer roots in the Feature Coupling Unit (FCU), which fuses local features and global representations under different resolutions in an interactive fashion. Conformer adopts a concurrent structure so that local features and global representations are retained to the maximum extent. Experiments show that Conformer, under the comparable parameter complexity, outperforms the visual transformer (DeiT-B) by 2.3% on ImageNet. On MSCOCO, it outperforms ResNet-101 by 3.7% and 3.6% mAPs for object detection and instance segmentation, respectively, demonstrating the great potential to be a general backbone network. Code is available at [github.com/pengzhiliang/Conformer](https://github.com/pengzhiliang/Conformer).

在卷积神经网络 (CNN) 中，卷积运算擅长于提取局部特征，但难以捕获全局表示。在 VisualTransformer中，级联的自我注意模块可以捕获长距离的特征依赖，但不幸的是会恶化局部特征细节。在本文中，我们提出了一种称为Conformer的混合网络结构，以利用卷积运算和自我注意机制来增强表征学习。构象源于特征耦合单元 (FCU) ，它以交互方式融合不同分辨率下的局部特征和全局表示。Conformer采用并行结构，以便最大程度地保留局部特征和全局表示。实验表明，在参数复杂度相当的情况下，Conformer在ImageNet上的性能比VisualTransformer (DeiT-B) 好2.3%。在MSCOCO 上，它在对象检测和实例分割方面的性能分别比ResNet-101高3.7%和3.6%，显示出作为通用主干网络的巨大潜力。代码可以在github上找到。[com/彭志良/合规者](https://github.com/pengzhiliang/Conformer)。

Multi-person total motion capture is extremely challenging when it comes to handle severe occlusions, different reconstruction granularities from body to face and hands, drastically changing observation scales and fast body movements. To overcome these challenges above, we contribute a lightweight total motion capture system for multi-person interactive scenarios using only sparse multi-view cameras. By contributing a novel hand and face bootstrapping algorithm, our method is capable of efficient localization and accurate association of the hands and faces even on severe occluded occasions. We leverage both pose regression and keypoints detection methods and further propose a unified two-stage parametric fitting method for achieving pixel-aligned accuracy. Moreover, for extremely self-occluded poses and close interactions, a novel feedback mechanism is proposed to propagate the pixel-aligned reconstructions into the next frame for more accurate association. Overall, we propose the first light-weight total capture system and achieves fast, robust and accurate multi-person total motion capture performance. The results and experiments show that our method achieves more accurate results than existing methods under sparse-view setups.

当涉及到处理严重的遮挡、从身体到面部和手的不同重建粒度、急剧变化的观察尺度和快速的身体运动时，多人全运动捕捉是极具挑战性的。为了克服上述挑战，我们提供了一个轻量级的全运动捕捉系统，用于仅使用稀疏多视图摄影机的多人交互场景。通过提供一种新的手和脸自举算法，我们的方法能够有效地定位和准确地关联手和脸，即使在严重遮挡的情况下。我们利用姿态回归和关键点检测方法，进一步提出了一种统一的两阶段参数拟合方法，以实现像素对齐精度。此外，对于极度自遮挡的姿势和密切的交互，提出了一种新的反馈机制，将像素对齐的重建传播到下一帧，以实现更精确的关联。总体而言，我们提出了第一个轻量级的全捕获系统，实现了快速、鲁棒和准确的多人全运动捕获性能。结果和实验表明，在稀疏视图设置下，我们的方法比现有方法获得更精确的结果。

while deep neural networks have shown impressive performance in many tasks, they are fragile to carefully designed adversarial attacks. We propose a novel adversarial training-based model by Attention Guided Knowledge Distillation and Bi-directional Metric Learning (AGKD-BML). The attention knowledge is obtained from a weight-fixed model trained on a clean dataset, referred to as a teacher model, and transferred to a model that is under training on adversarial examples (AEs), referred to as a student model. In this way, the student model is able to focus on the correct region, as well as correcting the intermediate features corrupted by AEs to eventually improve the model accuracy. Moreover, to efficiently regularize the representation in feature space, we propose a bidirectional metric learning. Specifically, given a clean image, it is first attacked to its most confusing class to get the forward AE. A clean image in the most confusing class is then randomly picked and attacked back to the original class to get the backward AE. A triplet loss is then used to shorten the representation distance between original image and its AE, while enlarge that between the forward and backward AEs. We conduct extensive adversarial robustness experiments on two widely used datasets with different attacks. Our proposed AGKD-BML model consistently outperforms the state-of-the-art approaches. The code of AGKD-BML will be available at: <https://github.com/hongw579/AGKD-BML>.

虽然深度神经网络在许多任务中表现出令人印象深刻的性能，但它们对精心设计的对抗性攻击很脆弱。提出了一种基于注意引导知识提取和双向度量学习的对抗训练模型（AGKD-BML）。注意力知识从在干净数据集上训练的权重固定模型（称为教师模型）中获得，并转移到正在接受对抗性示例（AEs）训练的模型（称为学生模型）中。通过这种方式，学生模型能够聚焦于正确的区域，并纠正被AEs破坏的中间特征，以最终提高模型精度。此外，为了有效地正则化特征空间中的表示，我们提出了一种双向度量学习方法。具体地说，给定一个干净的图像，它首先被攻击到其最混乱的类别，以获得前向AE。然后在最混乱的类中随机选取一个干净的图像，并攻击回原始类，以获得向后的AE。然后利用三重态损耗来缩短原始图像与其声发射之间的表示距离，同时放大前向和后向声发射之间的表示距离。我们在两个广泛使用的具有不同攻击的数据集上进行了广泛的对抗性鲁棒性实验。我们提出的AGKD-BML模型始终优于最先进的方法。AGKD-BML的代码将在以下位置提供：<https://github.com/hongw579/AGKD-BML>。

Frame sampling is a fundamental problem in video action recognition due to the essential redundancy in time and limited computation resources. The existing sampling strategy often employs a fixed frame selection and lacks the flexibility to deal with complex variations in videos. In this paper, we present a simple, sparse, and explainable frame sampler, termed as Motion-Guided Sampler (MGSampler). Our basic motivation is that motion is an important and universal signal that can drive us to adaptively select frames from videos. Accordingly, we propose two important properties in our MGSampler design: motion sensitive and motion uniform. First, we present two different motion representations to enable us to efficiently distinguish the motion-salient frames from the background. Then, we devise a motion-uniform sampling strategy based on the cumulative motion distribution to ensure the sampled frames evenly cover all the important segments with high motion salience. Our MGSampler yields a new principled and holistic sample scheme, that could be incorporated into any existing video architecture. Experiments on five benchmarks demonstrate the effectiveness of our MGSampler over previous fixed sampling strategies, and its generalization power across different backbones, video models, and datasets.

帧采样是视频动作识别中的一个基本问题，由于其在时间上的本质冗余和有限的计算资源。现有的采样策略通常采用固定的帧选择，并且缺乏处理视频中复杂变化的灵活性。在本文中，我们提出了一种简单、稀疏且可解释的帧取样器，称为运动引导取样器（MGSampler）。我们的基本动机是，运动是一种重要而普遍的信号，可以驱动我们从视频中自适应地选择帧。因此，我们提出了MG采样器设计中的两个重要特性：运动敏感和运动均匀。首先，我们提出了两种不同的运动表示，使我们能够有效地从背景中区分运动显著帧。然后，我们设计了一种基于累积运动分布的运动均匀采样策略，以确保采样帧均匀地覆盖所有具有高运动显著性的重要片段。我们的MGSampler产生了一个新的原则性和整体性的样本方案，可以整合到任何现有的视频架构中。在五个基准测试上的实验证明了我们的MGSampler比以前的固定采样策略更有效，它在不同主干、视频模型和数据集上的泛化能力更强。

Although having achieved great success in medical image segmentation, deep convolutional neural networks usually require a large dataset with manual annotations for training and are difficult to generalize to unseen classes. Few-shot learning has the potential to address these challenges by learning new classes from only a few labeled examples. In this work, we propose a new framework for few-shot medical image segmentation based on prototypical networks. Our innovation lies in the design of two key modules: 1) a context relation encoder (CRE) that uses correlation to capture local relation features between foreground and background regions; and 2) a recurrent mask refinement module that repeatedly uses the CRE and a prototypical network to recapture the change of context relationship and refine the segmentation mask iteratively. Experiments on two abdomen CT datasets and an abdomen MRI dataset show the proposed method obtains substantial improvement over the state-of-the-art methods by an average of 16.32%, 8.45% and 6.24% in terms of DSC, respectively. Code is publicly available.

尽管深卷积神经网络在医学图像分割方面取得了巨大的成功，但它通常需要一个带有人工标注的大数据集进行训练，并且很难推广到不可见的类。通过仅从几个标记的示例学习新课程，很少有机会学习能够解决这些挑战。在这项工作中，我们提出了一种新的基于原型网络的少镜头医学图像分割框架。我们的创新之处在于设计了两个关键模块：1）上下文关系编码器（CRE），该编码器使用相关性捕获前景和背景区域之间的局部关系特征；2）一个递归掩码细化模块，该模块反复使用CRE和原型网络来重新捕获上下文关系的变化，并迭代地细化分段掩码。在两个腹部CT数据集和一个腹部MRI数据集上的实验表明，所提出的方法与最先进的方法相比，在DSC方面的平均改善率分别为16.32%、8.45%和6.24%。代码是公开的。

The deep learning methods in addressing semantic segmentation typically demand vast amount of pixel-wise annotated training samples. In this work, we present zero-shot semantic segmentation, which aims to identify not only the seen classes contained in training but also the novel classes that have never been seen. We adopt a stringent inductive setting in which only the instances of seen classes are accessible during training. We propose an open-aware prototypical matching approach to accomplish the segmentation. The prototypical way extracts the visual representations by a set of prototypes, making it convenient and flexible to add new unseen classes. A prototype projection is trained to map the semantic representations towards prototypes based on seen instances, and will generate prototypes for unseen classes. Moreover, an open-set rejection is utilized to detect the objects that do not belong to any seen classes, which greatly reduces the misclassifications of unseen objects as seen classes caused by the lack of unseen training instances. We apply the framework on two segmentation datasets, Pascal VOC 2012 and Pascal Context, and achieve impressively state-of-the-art performance.

处理语义分割的深度学习方法通常需要大量的像素注释训练样本。在这项工作中，我们提出了零镜头语义分割，其目的不仅是识别训练中包含的已知类，而且还识别从未见过的新类。我们采用严格的归纳设置，在培训期间，只有SEED课程的实例可以访问。我们提出了一种开放感知的原型匹配方法来完成分割。原型方法通过一组原型提取视觉表示，从而方便灵活地添加新的不可见类。原型投影经过训练，可以根据看到的实例将语义表示映射到原型，并为看不见的类生成原型。此外，使用开集拒绝检测不属于任何可见类的对象，这大大减少了由于缺少可见训练实例而导致的不可见对象可见类的错误分类。我们在两个分割数据集Pascal VOC 2012和Pascal Context上应用了该框架，取得了令人印象深刻的最新性能。

This paper tackles the task of category-level pose estimation for garments. With a near infinite degree of freedom, a garment's full configuration (i.e., poses) is often described by the per-vertex 3D locations of its entire 3D surface. However, garments are also commonly subject to extreme cases of self-occlusion, especially when folded or crumpled, making it challenging to perceive their full 3D surface. To address these challenges, we propose GarmentNets, where the key idea is to formulate the deformable object pose estimation problem as a shape completion task in the canonical space. This canonical space is defined across garments instances within a category, therefore, specifies the shared category-level pose. By mapping the observed partial surface to the canonical space and completing it in this space, the output representation describes the garment's full configuration using a complete 3D mesh with the per-vertex canonical coordinate label. To properly handle the thin 3D structure presented on garments, we proposed a novel 3D shape representation using the generalized winding number field. Experiments demonstrate that GarmentNets is able to generalize to unseen garment instances and achieve significantly better performance compared to alternative approaches. Code and data will be available online.

本文研究了服装类别级姿态估计问题。在接近无限自由度的情况下，服装的完整配置（即姿势）通常由其整个3D曲面的逐顶点3D位置来描述。然而，服装通常也会受到极端的自遮挡情况的影响，尤其是在折叠或皱折时，这使得感知其完整的3D表面非常困难。为了解决这些挑战，我们提出了GarmentNets，其中的关键思想是将可变形对象姿势估计问题表述为规范空间中的形状完成任务。此规范空间是在类别内的多个实例之间定义的，因此指定共享类别级别的姿势。通过将观察到的部分曲面映射到规范空间并在空间中完成，输出表示将使用带有逐顶点规范坐标标签的完整三维网格描述服装的完整配置。为了正确处理服装上呈现的薄三维结构，我们提出了一种基于广义卷绕数场的三维形状表示方法。实验表明，GarmentNets能够推广到看不见的服装实例，并且与其他方法相比，其性能显著提高。代码和数据将在线提供。

While multi-step adversarial training is widely popular as an effective defense method against strong adversarial attacks, its computational cost is notoriously expensive, compared to standard training. Several single-step adversarial training methods have been proposed to mitigate the above-mentioned overhead cost; however, their performance is not sufficiently reliable depending on the optimization setting. To overcome such limitations, we deviate from the existing input-space-based adversarial training regime and propose a single-step latent adversarial training method (SLAT), which leverages the gradients of latent representation as the latent adversarial perturbation. We demonstrate that the L1 norm of feature gradients is implicitly regularized through the adopted latent perturbation, thereby recovering local linearity and ensuring reliable performance, compared to the existing single-step adversarial training methods. Because latent perturbation is based on the gradients of the latent representations which can be obtained for free in the process of input gradients computation, the proposed method costs roughly the same time as the fast gradient sign method. Experiment results demonstrate that the proposed method, despite its structural simplicity, outperforms state-of-the-art accelerated adversarial training methods.

虽然多步骤对抗性训练作为一种有效防御强对抗性攻击的方法受到广泛欢迎，但与标准训练相比，其计算成本是出了名的昂贵。提出了几种单步对抗式训练方法，以降低上述开销成本；但是，根据优化设置，它们的性能不够可靠。为了克服这些限制，我们偏离了现有的基于输入空间的对抗性训练机制，提出了一种单步潜在对抗性训练方法（SLAT），该方法利用潜在表示的梯度作为潜在对抗性干扰。我们证明，与现有的单步对抗训练方法相比，特征梯度的L1范数通过所采用的潜在扰动隐式正则化，从而恢复局部线性并确保可靠的性能。由于潜在扰动是基于在输入梯度计算过程中可以免费获得的潜在表示的梯度，因此该方法的成本与快速梯度符号法大致相同。实验结果表明，尽管该方法结构简单，但其性能优于最先进的加速对抗训练方法。

We present a method for optimization-based recovery of eye motion from rolling shutter video of the retina. Our approach formulates eye tracking as an optimization problem that jointly estimates the retina's motion and appearance using convex optimization and a constrained version of gradient descent. By incorporating the rolling shutter imaging model into the formulation of our joint optimization, we achieve state-of-the-art accuracy both offline and in real-time. We apply our method to retina video captured with an adaptive optics scanning laser ophthalmoscope (AOSLO), demonstrating eye tracking at 1 kHz with accuracies below one arcminute -- over an order of magnitude higher than conventional eye tracking systems.

我们提出了一种从视网膜的滚动快门视频中恢复眼球运动的优化方法。我们的方法将眼睛跟踪描述为一个优化问题，该问题使用凸优化和梯度下降的约束形式联合估计视网膜的运动和外观。通过将滚动快门成像模型纳入我们的联合优化公式中，我们实现了最先进的离线和实时精度。我们将我们的方法应用于自适应光学扫描激光检眼镜（AOSLO）捕获的视网膜视频，演示了1kHz下的眼睛跟踪，精度低于1弧分——比传统的眼睛跟踪系统高出一个数量级。

We develop an approach to recover the underlying properties of fluid-dynamical processes from sparse measurements. We are motivated by the task of imaging the stochastically evolving environment surrounding black holes, and demonstrate how flow parameters can be estimated from sparse interferometric measurements used in radio astronomical imaging. To model the stochastic flow we use spatio-temporal Gaussian Random Fields (GRFs). The high dimensionality of the underlying source video makes direct representation via a GRF's full covariance matrix intractable. In contrast, stochastic partial differential equations are able to capture correlations at multiple scales by specifying only local interaction coefficients. Our approach estimates the coefficients of a space-time diffusion equation that dictates the stationary statistics of the dynamical process. We analyze our approach on realistic simulations of black hole evolution and demonstrate its advantage over state-of-the-art dynamic black hole imaging techniques.

我们开发了一种从稀疏测量中恢复流体动力学过程基本性质的方法。我们的任务是对黑洞周围随机演化的环境进行成像，并演示如何通过射电天文成像中使用的稀疏干涉测量来估计流量参数。为了模拟随机流，我们使用时空高斯随机场（GRF）。底层源视频的高维性使得通过GRF的全协方差矩阵直接表示变得难以处理。相比之下，随机偏微分方程能够通过仅指定局部相互作用系数在多个尺度上捕获相关性。我们的方法估计时空扩散方程的系数，该方程决定了动力学过程的平稳统计。我们分析了我们的方法对黑洞演化的真实模拟，并展示了它相对于最先进的动态黑洞成像技术的优势。

We introduce an evaluation methodology for visual question answering (VQA) to better diagnose cases of shortcut learning. These cases happen when a model exploits spurious statistical regularities to produce correct answers but does not actually deploy the desired behavior. There is a need to identify possible shortcuts in a dataset and assess their use before deploying a model in the real world. The research community in VQA has focused exclusively on question-based shortcuts, where a model might, for example, answer "What is the color of the sky" with "blue" by relying mostly on the question-conditional training prior and give little weight to visual evidence. We go a step further and consider multimodal shortcuts that involve both questions and images. We first identify potential shortcuts in the popular VQA v2 training set by mining trivial predictive rules such as co-occurrences of words and visual elements. We then introduce VQA-CounterExamples (VQA-CE), an evaluation protocol based on our subset of CounterExamples i.e. image-question-answer triplets where our rules lead to incorrect answers. We use this new evaluation in a large-scale study of existing approaches for VQA. We demonstrate that even state-of-the-art models perform poorly and that existing techniques to reduce biases are largely ineffective in this context. Our findings suggest that past work on question-based biases in VQA has only addressed one facet of a complex issue. The code for our method is available at \url{https://github.com/cdancette/detect-shortcuts}

我们介绍了一种视觉问答（VQA）评估方法，以更好地诊断快捷学习案例。这些情况发生在模型利用虚假的统计规律来产生正确的答案，但实际上没有部署所需的行为时。在现实世界中部署模型之前，需要确定数据集中可能的快捷方式并评估其使用情况。VQA的研究社区专门关注基于问题的快捷方式，例如，模型可能会通过主要依赖之前的问题条件训练，用“蓝色”回答“天空的颜色是什么”，而很少考虑视觉证据。我们进一步考虑多模态快捷方式，包括问题和图像。我们首先通过挖掘琐碎的预测规则（如单词和视觉元素的共现），识别流行VQA v2训练集中的潜在捷径。然后，我们介绍了VQA反例（VQA-CE），这是一种基于反例子集的评估协议，即图像问答三元组，其中我们的规则会导致错误答案。我们在对现有VQA方法的大规模研究中使用了这种新的评估。我们证明，即使是最先进的模型也表现不佳，现有的减少偏差的技术在这种情况下基本上是无效的。我们的发现表明，过去关于VQA中基于问题的偏见的研究只解决了复杂问题的一个方面。我们的方法的代码可从\url{https://github.com/cdancette/detect-shortcuts}获得

Recent learning approaches that implicitly represent surface geometry using coordinate-based neural representations have shown impressive results in the problem of multi-view 3D reconstruction. The effectiveness of these techniques is, however, subject to the availability of a large number (several tens) of input views of the scene, and computationally demanding optimizations. In this paper, we tackle these limitations for the specific problem of few-shot full 3D head reconstruction, by endowing coordinate-based representations with a probabilistic shape prior that enables faster convergence and better generalization when using few input images (down to three). First, we learn a shape model of 3D heads from thousands of incomplete raw scans using implicit representations. At test time, we jointly overfit two coordinate-based neural networks to the scene, one modeling the geometry and another estimating the surface radiance, using implicit differentiable rendering. We devise a two-stage optimization strategy in which the learned prior is used to initialize and constrain the geometry during an initial optimization phase. Then, the prior is unfrozen and fine-tuned to the scene. By doing this, we achieve high-fidelity head reconstructions, including hair and shoulders, and with a high level of detail that consistently outperforms both state-of-the-art 3D Morphable Models methods in the few-shot scenario, and non-parametric methods when large sets of views are available.

最近使用基于坐标的神经表示隐式表示曲面几何的学习方法在多视图三维重建问题上取得了令人印象深刻的成果。然而，这些技术的有效性取决于大量（几十个）场景输入视图的可用性和计算要求的优化。在本文中，我们针对少镜头全三维头部重建的具体问题解决了这些限制，通过赋予基于坐标的表示以概率形状先验，在使用少量输入图像（最多三幅）时，能够更快地收敛并更好地泛化。首先，我们使用隐式表示从数千次不完整的原始扫描中学习3D头部的形状模型。在测试时，我们使用隐式可微渲染将两个基于坐标的神经网络联合过度拟合到场景中，一个用于建模几何体，另一个用于估计表面辐射。我们设计了一个两阶段优化策略，在初始优化阶段，使用学到的先验知识初始化和约束几何体。然后，先验解冻结并微调到场景。通过这样做，我们实现了高保真的头部重建，包括头发和肩膀，并且具有高水平的细节，在少数镜头场景中始终优于最先进的3D可变形模型方法，在大视图集可用时优于非参数化方法。

Video Question Answering (Video QA) aims to give an answer to the question through semantic reasoning between visual and linguistic information. Recently, handling large amounts of multi-modal video and language information of a video is considered important in the industry. However, the current video QA models use deep features, suffered from significant computational complexity and insufficient representation capability both in training and testing. Existing features are extracted using pre-trained networks after all the frames are decoded, which is not always suitable for video QA tasks. In this paper, we develop a novel deep neural network to provide video QA features obtained from coded video bit-stream to reduce the complexity. The proposed network includes several dedicated deep modules to both the video QA and the video compression system, which is the first attempt at the video QA task. The proposed network is predominantly model-agnostic. It is integrated into the state-of-the-art networks for improved performance without any computationally expensive motion-related deep models. The experimental results demonstrate that the proposed network outperforms the previous studies at lower complexity.

视频问答 (Video-Question-Answering, Video-QA) 旨在通过视觉信息和语言信息之间的语义推理来回答问题。最近，处理大量的多模态视频和视频的语言信息在业界被认为是重要的。然而，目前的视频质量保证模型使用深度特征，在训练和测试中存在计算复杂度高和表示能力不足的问题。现有的特征是在所有帧解码后使用预先训练的网络提取的，这并不总是适用于视频质量保证任务。在本文中，我们开发了一种新的深度神经网络来提供从编码视频比特流获得的视频质量保证特征，以降低复杂性。建议的网络包括几个专用于视频QA和视频压缩系统的深层模块，这是视频QA任务的首次尝试。提出的网络主

要是模型不可知的。它被集成到最先进的网络中，以提高性能，而无需任何计算昂贵的运动相关深度模型。实验结果表明，在较低的复杂度下，该网络的性能优于以往的研究。

Measuring the acoustic characteristics of a space is often done by capturing its impulse response (IR), a representation of how a full-range stimulus sound excites it. This work generates an IR from a single image, which can then be applied to other signals using convolution, simulating the reverberant characteristics of the space shown in the image. Recording these IRs is both time-intensive and expensive, and often infeasible for inaccessible locations. We use an end-to-end neural network architecture to generate plausible audio impulse responses from single images of acoustic environments. We evaluate our method both by comparisons to ground truth data and by human expert evaluation. We demonstrate our approach by generating plausible impulse responses from diverse settings and formats including well known places, musical halls, rooms in paintings, images from animations and computer games, synthetic environments generated from text, panoramic images, and video conference backgrounds.

测量一个空间的声学特性通常是通过捕捉它的脉冲响应 (IR) 来完成的，这是一种表示全范围刺激声音如何激励它的表示。这项工作从一幅图像生成一个IR，然后通过卷积将其应用于其他信号，模拟图像中所示空间的混响特性。记录这些IRs既耗时又昂贵，而且对于无法访问的位置通常不可行。我们使用端到端的神经网络结构，从单个声环境图像生成合理的音频脉冲响应。我们通过与地面真实数据的比较和人类专家的评估来评估我们的方法。我们通过从不同的环境和格式生成合理的脉冲响应来演示我们的方法，这些环境和格式包括知名场所、音乐厅、绘画房间、动画和电脑游戏中的图像、文本生成的合成环境、全景图像和视频会议背景。

Group activity recognition aims to understand the activity performed by a group of people. In order to solve it, modeling complex spatio-temporal interactions is the key. Previous methods are limited in reasoning on a predefined graph, which ignores the inherent person-specific interaction context. Moreover, they adopt inference schemes that are computationally expensive and easily result in the over-smoothing problem. In this paper, we manage to achieve spatio-temporal person-specific inferences by proposing Dynamic Inference Network (DIN), which composes of Dynamic Relation (DR) module and Dynamic Walk (DW) module. We firstly propose to initialize interaction fields on a primary spatio-temporal graph. Within each interaction field, we apply DR to predict the relation matrix and DW to predict the dynamic walk offsets in a joint-processing manner, thus forming a person-specific interaction graph. By updating features on the specific graph, a person can possess a global-level interaction field with a local initialization. Experiments indicate both modules' effectiveness. Moreover, DIN achieves significant improvement compared to previous state-of-the-art methods on two popular datasets under the same setting, while costing much less computation overhead of the reasoning module.

团体活动识别的目的是了解一组人所进行的活动。为了解决这一问题，对复杂的时空相互作用进行建模是关键。以前的方法局限于在预定义的图上进行推理，这忽略了固有的特定于人的交互上下文。此外，它们采用的推理方案计算量大，容易导致过度平滑问题。本文提出了动态推理网络 (DIN)，它由动态关系 (DR) 模块和动态行走 (DW) 模块组成，实现了时空的特定人推理。我们首先提出在主时空图上初始化交互场。在每个交互场中，我们应用DR来预测关系矩阵，DW以联合处理的方式预测动态行走偏移，从而形成特定于人的交互图。通过更新特定图形上的特征，一个人可以拥有一个具有局部初始化的全局级交互字段。实验表明这两个模块都是有效的。此外，在相同设置下，DIN在两个流行数据集上实现了与以前最先进方法相比的显著改进，同时大大降低了推理模块的计算开销。

A high quality disparity remapping method that preserves 2D shapes and 3D structures, and adjusts disparities of important objects in stereo image pairs is proposed. It is formulated as a constrained optimization problem, whose solution is challenging, since we need to meet multiple requirements of disparity remapping simultaneously. The one-stage optimization process either degrades the quality of important objects or introduces serious distortions in background regions. To address this challenge, we propose a two-stage warping process to solve it. In the first stage, we develop a warping model that finds the optimal warping grids for important objects to fulfill multiple requirements of disparity remapping. In the second stage, we derive another warping model to refine warping results in less important regions by eliminating serious distortions in shape, disparity and 3D structure. The superior performance of the proposed method is demonstrated by experimental results

提出了一种高质量视差重映射方法，该方法保留了二维形状和三维结构，并调整了立体图像对中重要对象的视差。它被描述为一个约束优化问题，由于我们需要同时满足视差重映射的多个要求，其解是具有挑战性的。一步优化过程要么降低重要对象的质量，要么在背景区域引入严重的失真。为了应对这一挑战，我们提出了一种两阶段翘曲工艺来解决这一问题。在第一阶段，我们开发了一个扭曲模型，为重要对象找到最佳扭曲网格，以满足视差重映射的多种要求。在第二阶段中，我们推导了另一个扭曲模型，通过消除形状、视差和三维结构中的严重扭曲来细化不太重要区域中的扭曲结果。实验结果表明，该方法具有良好的性能

Automatic augmentation methods have recently become a crucial pillar for strong model performance in vision tasks. While existing automatic augmentation methods need to trade off simplicity, cost and performance, we present a most simple baseline, TrivialAugment, that outperforms previous methods for almost free. TrivialAugment is parameter-free and only applies a single augmentation to each image. Thus, TrivialAugment's effectiveness is very unexpected to us and we performed very thorough experiments to study its performance. First, we compare TrivialAugment to previous state-of-the-art methods in a variety of image classification scenarios. Then, we perform multiple ablation studies with different augmentation spaces, augmentation methods and setups to understand the crucial requirements for its performance. Additionally, we provide a simple interface to facilitate the widespread adoption of automatic augmentation methods, as well as our full code base for reproducibility. Since our work reveals a stagnation in many parts of automatic augmentation research, we end with a short proposal of best practices for sustained future progress in automatic augmentation methods.

最近，自动增强方法已成为视觉任务中强大模型性能的关键支柱。虽然现有的自动增强方法需要在简单性、成本和性能之间进行权衡，但我们提供了一个最简单的基线，即平凡增强，它几乎免费地优于以前的方法。TrivialAugment是无参数的，只对每个图像应用单个增强。因此，琐碎增强的有效性对我们来说是非常意外的，我们进行了非常彻底的实验来研究它的性能。首先，在各种图像分类场景中，我们比较了TrivialAugment与以前最先进的方法。然后，我们使用不同的增强空间、增强方法和设置进行多次消融研究，以了解其性能的关键要求。此外，我们还提供了一个简单的接口，以促进自动增强方法的广泛采用，并提供了完整的代码库，以确保再现性。由于我们的工作揭示了自动增强研究的许多方面停滞不前，因此我们最后提出了一个简短的最佳实践建议，以促进自动增强方法在未来的持续发展。

The task of reflection symmetry detection remains challenging due to significant variations and ambiguities of symmetry patterns in the wild. Furthermore, since the local regions are required to match in reflection for detecting a symmetry pattern, it is hard for standard convolutional networks, which are not equivariant to rotation and reflection, to learn the task. To address the issue, we introduce a new convolutional technique, dubbed the polar matching convolution, which leverages a polar feature pooling, a self-similarity encoding, and a systematic kernel design for axes of different angles. The proposed high-dimensional kernel convolution network effectively learns to discover symmetry patterns from real-world images, overcoming the limitations of standard convolution. In addition, we present a new dataset and introduce a self-supervised learning strategy by augmenting the dataset with synthesizing images. Experiments demonstrate that our method outperforms state-of-the-art methods in terms of accuracy and robustness.

由于野外对称模式的显著变化和模糊性，反射对称性检测的任务仍然具有挑战性。此外，由于检测对称模式需要在反射中匹配局部区域，因此对于与旋转和反射不等价的标准卷积网络来说，很难学习该任务。为了解决这个问题，我们引入了一种新的卷积技术，称为极性匹配卷积，它利用极性特征池、自相似编码和不同角度轴的系统内核设计。所提出的高维核卷积网络克服了标准卷积的局限性，有效地从真实图像中发现对称模式。此外，我们还提出了一种新的数据集，并通过合成图像来增强数据集，从而引入了一种自监督学习策略。实验表明，我们的方法在准确性和鲁棒性方面优于现有的方法。

Modeling temporal visual context across frames is critical for video instance segmentation (VIS) and other video understanding tasks. In this paper, we propose a fast online VIS model termed CrossVIS. For temporal information modeling in VIS, we present a novel crossover learning scheme that uses the instance feature in the current frame to pixel-wisely localize the same instance in other frames. Different from previous schemes, crossover learning does not require any additional network parameters for feature enhancement. By integrating with the instance segmentation loss, crossover learning enables efficient cross-frame instance-to-pixel relation learning and brings cost-free improvement during inference. Besides, a global balanced instance embedding branch is proposed for better and more stable online instance association. We conduct extensive experiments on three challenging VIS benchmarks, i.e., YouTube-VIS-2019, OVIS, and YouTube-VIS-2021 to evaluate our methods. CrossVIS achieves state-of-the-art online VIS performance and shows a decent trade-off between latency and accuracy. Code is available at <https://github.com/hustvl/CrossVIS>.

跨帧建模时间视觉上下文对于视频实例分割 (VIS) 和其他视频理解任务至关重要。在本文中，我们提出了一种称为CrossVIS的快速在线VIS模型。对于VIS中的时间信息建模，我们提出了一种新的交叉学习方案，该方案使用当前帧中的实例特征对其他帧中的相同实例进行像素智能定位。与以前的方案不同，交叉学习不需要任何额外的网络参数来增强特征。交叉学习结合实例分割损失，实现了高效的跨帧实例-像素关系学习，并在推理过程中带来了无代价的改进。此外，为了更好、更稳定的在线实例关联，提出了一种全局平衡的实例嵌入分支。我们在三个具有挑战性的VIS基准上进行了广泛的实验，即YouTube-VIS-2019、OVIS和YouTube-VIS-2021，以评估我们的方法。CrossVIS实现了最先进的在线VIS性能，并在延迟和准确性之间进行了适当的权衡。代码可在<https://github.com/hustvl/CrossVIS>.

We present a multiview pseudo-labeling approach to video learning, a novel framework that uses complementary views in the form of appearance and motion information for semi-supervised learning in video. The complementary views help obtain more reliable "pseudo-labels" on unlabeled video, to learn stronger video representations than from purely supervised data. Though our method capitalizes on multiple views, it nonetheless trains a model that is shared across appearance and motion input and thus, by design, incurs no additional computation overhead at inference time. On multiple video recognition datasets, our method substantially outperforms its supervised counterpart, and compares favorably to previous work on standard benchmarks in self-supervised video representation learning.

我们提出了一种用于视频学习的多视图伪标记方法，这是一种新的框架，使用外观和运动信息形式的互补视图进行视频半监督学习。补充视图有助于获得更可靠的“伪标签”在未标记的视频上，学习比纯监督数据更强的视频表示。虽然我们的方法利用了多个视图，但它训练了一个在外观和运动输入上共享的模型，因此，通过设计，在推断时不会产生额外的计算开销。在多个视频识别数据集上，ou在自监督视频表示学习中，r方法的性能明显优于有监督的方法，并优于以前在标准基准上的工作。

The key challenge in designing a sketch representation lies with handling the abstract and iconic nature of sketches. Existing work predominantly utilizes either, (i) a pixelative format that treats sketches as natural images employing off-the-shelf CNN-based networks, or (ii) an elaborately designed vector format that leverages the structural information of drawing orders using sequential RNN-based methods. While the pixelative format lacks intuitive exploitation of structural cues, sketches in vector format are absent in most cases limiting their practical usage. Hence, in this paper, we propose a lattice structured sketch representation that not only removes the bottleneck of requiring vector data but also preserves the structural cues that vector data provides. Essentially, sketch lattice is a set of points sampled from the pixelative format of the sketch using a lattice graph. We show that our lattice structure is particularly amenable to structural changes that largely benefits sketch abstraction modeling for generation tasks. Our lattice representation could be effectively encoded using a graph model, that uses significantly fewer model parameters (13.5 times lesser) than existing state-of-the-art. Extensive experiments demonstrate the effectiveness of sketch lattice for sketch manipulation, including sketch healing and image-to-sketch synthesis.

设计草图表示的关键挑战在于处理草图的抽象和标志性。现有工作主要利用：(i) 采用现成的基于CNN的网络将草图视为自然图像的像素格式，或 (ii) 精心设计的向量格式，利用基于顺序RNN的方法绘制顺序的结构信息。虽然像素格式缺乏对结构线索的直观利用，但矢量格式的草图在大多数情况下都不存在，限制了它们的实际使用。因此，在本文中，我们提出了一种格结构草图表示法，它不仅消除了需要矢量数据的瓶颈，而且保留了矢量数据提供的结构线索。基本上，草图晶格是使用晶格图从草图的像素格式中采样的一组点。我们表明，我们的晶格结构特别适合于结构变化，这在很大程度上有利生成任务的草图抽象建模。我们的晶格表示可以使用图形模型进行有效编码，该模型使用的模型参数比现有的最新技术少得多（少13.5倍）。大量实验证明了草图格在草图处理中的有效性，包括草图修复和图像到草图合成。

Person search suffers from the conflicting objectives of commonness and uniqueness between the person detection and re-identification tasks that make the end-to-end training of person search networks difficult. In this paper, we propose a trident network for person search that performs detection, re-identification, and part classification together. We also devise a novel end-to-end training method using adaptive gradient weighting that controls the flow of back-propagated gradients through the re-identification and part classification networks according to the quality of the person detection. The proposed method not only prevents the over-fitting but encourages to exploit fine-grained features by incorporating the part classification branch into the person search framework. Experimental results on the CUHK-SYSU and PRW datasets demonstrate that the proposed method achieves the best performance among the state-of-the-art end-to-end person search methods.

人员搜索面临着人员检测和重新识别任务之间的共同性和唯一性目标的冲突，这使得人员搜索网络的端到端训练变得困难。在本文中，我们提出了一种用于人员搜索的三叉戟网络，它同时执行检测、重新识别和零件分类。我们还设计了一种新的端到端训练方法，使用自适应梯度加权，根据人员检测的质量，通过重新识别和部分分类网络控制反向传播的梯度流。该方法不仅防止了过度拟合，而且通过将零件分类分支合并到人员搜索框架中，鼓励利用细粒度特征。在中大SYSU和PRW数据集上的实验结果表明，该方法在最先进的端到端人员搜索方法中取得了最好的性能。

Helmholtz stereopsis (HS) exploits the reciprocity principle of light propagation (i.e., the Helmholtz reciprocity) for 3D reconstruction of surfaces with arbitrary reflectance. In this paper, we present the polarimetric Helmholtz stereopsis (polar-HS), which extends the classical HS by considering the polarization state of light in the reciprocal paths. With the additional phase information from polarization, polar-HS requires only one reciprocal image pair. We formulate new reciprocity and diffuse/specular polarimetric constraints to recover surface depths and normals using an optimization framework. Using a hardware prototype, we show that our approach produces high-quality 3D reconstruction for different types of surfaces, ranging from diffuse to highly specular.

亥姆霍兹立体视觉 (HS) 利用光传播的互易性原理（即亥姆霍兹互易性）对具有任意反射率的表面进行三维重建。在本文中，我们提出了偏振亥姆霍兹立体视觉 (polarimetric Helmholtz stereopsis, polar HS)，它通过考虑光在倒数路径中的偏振状态来扩展经典HS。由于来自偏振的附加相位信息，极性HS只需要一个倒数图像对。我们制定了新的互易性和漫反射/镜面偏振约束，以使用优化框架恢复曲面深度和法线。通过一个硬件原型，我们证明了我们的方法可以为不同类型的曲面（从漫反射到高镜面反射）生成高质量的三维重建。

The domain-adaptive semantic segmentation in aerial images by a deep-learning technique remains a challenge owing to the domain gaps caused by a resolution, image sensors, time-zone, the density of buildings, and even building styles of each city. Currently, convolutional neural network (CNN)-based domain adaptation methodologies have been developed to decrease the domain gaps, but, they have shown still poor performance to utilize multiple aerial images in different domains. In this paper, therefore, the CNN-based network denoted as Self-Mutating Network, which changes the values of parameters of convolutional filters itself according to the domain of input image, is proposed. By adopting Parameter Mutation to change the values of parameters and Parameter Fluctuation to randomly convulse the parameters, the network self-changes and fine-tunes the parameters, then achieves better predictions of a domain-adaptive segmentation. Through the ablation study of the Self-Mutating Network, we concluded that the Self-Mutating Network can be utilized in the domain-adaptive semantic segmentation of aerial images in different domains.

由于分辨率、图像传感器、时区、建筑密度甚至每个城市的建筑风格都会造成领域差异，因此利用深度学习技术在航空图像中进行领域自适应语义分割仍然是一个挑战。目前，基于卷积神经网络（CNN）的领域自适应方法已经被开发出来以减少领域差距，但是，它们在不同领域中利用多幅航空图像的性能仍然很差。因此，本文提出了一种基于CNN的网络，称为自变异网络，它根据输入图像的区域改变卷积滤波器本身的参数值。通过参数变异改变参数值，参数波动随机振荡参数，网络自动改变并微调参数，从而实现更好的域自适应分割预测。通过对自变异网络的研究，我们得出结论：自变异网络可以用于不同领域航空影像的领域自适应语义分割。

Neural volumetric representations such as Neural Radiance Fields (NeRF) have emerged as a compelling technique for learning to represent 3D scenes from images with the goal of rendering photorealistic images of the scene from unobserved viewpoints. However, NeRF's computational requirements are prohibitive for real-time applications: rendering views from a trained NeRF requires querying a multilayer perceptron (MLP) hundreds of times per ray. We present a method to train a NeRF, then precompute and store (i.e. ""bake"" it as a novel representation called a Sparse Neural Radiance Grid (SNeRG) that enables real-time rendering on commodity hardware. To achieve this, we introduce 1) a reformulation of NeRF's architecture, and 2) a sparse voxel grid representation with learned feature vectors. The resulting scene representation retains NeRF's ability to render fine geometric details and view-dependent appearance, is compact (averaging less than 90 MB per scene), and can be rendered in real-time (higher than 30 frames per second on a laptop GPU). Actual screen captures are shown in our video.

神经辐射场（NeRF）等神经体积表示已成为一种引人注目的技术，用于学习从图像中表示三维场景，目的是从未观察到的视点渲染场景的照片级真实感图像。然而，NeRF的计算要求对于实时应用来说是禁止的：从经过训练的NeRF渲染视图需要每射线查询多层次感知器（MLP）数百次。我们提出了一种方法来训练NeRF，然后预算并存储（即“烘焙”）它，作为一种称为稀疏神经辐射网格（SNeRG）的新表示，它可以在商品硬件上进行实时渲染。为了实现这一点，我们引入了1) NeRF架构的重新表述，以及2) 具有学习特征向量的稀疏体素网格表示。生成的场景表示保留了NeRF渲染精细几何细节和视图相关外观的能力，紧凑（每个场景平均小于90 MB），并且可以实时渲染（在笔记本电脑GPU上高于每秒30帧）。实际屏幕截图显示在我们的视频中。

Object detection has been widely used in many safety-critical tasks, such as autonomous driving. However, its vulnerability to adversarial examples has not been sufficiently studied, especially under the practical scenario of black-box attacks, where the attacker can only access the query feedback of predicted bounding-boxes and top-1 scores returned by the attacked model. Compared with black-box attack to image classification, there are two main challenges in black-box attack to detection. Firstly, even if one bounding-box is successfully attacked, another sub-optimal bounding-box may be detected near the attacked bounding-box. Secondly, there are multiple bounding-boxes, leading to very high attack cost. To address these challenges, we propose a Parallel Rectangle Flip Attack (PRFA) via random search. Specifically, we generate perturbations in each rectangle patch to avoid sub-optimal detection near the attacked region. Besides, utilizing the observation that adversarial perturbations mainly locate around objects' contours and critical points under white-box attacks, the search space of attacked rectangles is reduced to improve the attack efficiency. Moreover, we develop a parallel mechanism of attacking multiple rectangles simultaneously to further accelerate the attack process. Extensive experiments demonstrate that our method can effectively and efficiently attack various popular object detectors, including anchor-based and anchor-free, and generate transferable adversarial examples.

目标检测已广泛应用于许多安全关键任务，如自动驾驶。然而，其对敌对示例的脆弱性尚未得到充分研究，特别是在黑盒攻击的实际场景下，攻击者只能访问预测边界框的查询反馈和被攻击模型返回的top-1分数。与针对图像分类的黑盒攻击相比，黑盒攻击在检测方面面临两大挑战。首先，即使一个边界框被成功攻击，也可能在被攻击的边界框附近检测到另一个次优边界框。其次，存在多个边界盒，导致非常高的攻击成本。为了应对这些挑战，我们提出了一种通过随机搜索的并行矩形翻转攻击（PRFA）。具体而言，我们在每个矩形面片中生成扰动，以避免攻击区域附近的次优检测。此外，利用白盒攻击下敌方扰动主要分布在目标轮廓和临界点附近的观察结果，减少了被攻击矩形的搜索空间，提高了攻击效率。此外，我们还开发了一种同时攻击多个矩形的并行机制，以进一步加快攻击过程。大量的实验表明，我们的方法可以有效地攻击各种流行的目标检测器，包括基于锚和无锚，并生成可转移的对抗性示例。

Every time you sit in front of a TV or monitor, your face is actively illuminated by time-varying patterns of light. This paper proposes to use this time-varying illumination for synthetic relighting of your face with any new illumination condition. In doing so, we take inspiration from the light stage work of Debevec et al. [4], who first demonstrated the ability to relight people captured in a controlled lighting environment. Whereas existing light stages require expensive, room-scale spherical capture gantries and exist in only a few labs in the world, we demonstrate how to acquire useful data from a normal TV or desktop monitor. Instead of subjecting the user to uncomfortable rapidly flashing light patterns, we operate on images of the user watching a YouTube video or other standard content. We train a deep network on images plus monitor patterns of a given user and learn to predict images of that user under any target illumination (monitor pattern). Experimental evaluation shows that our method produces realistic relighting results.

每次你坐在电视机或显示器前，你的脸都会被时变的光照活跃地照亮。本文建议在任何新的照明条件下，使用这种时变照明对您的脸进行合成重新照明。在这样做的过程中，我们从Debevec等人[4]的灯光舞台工作中获得了灵感，他们首先展示了在受控灯光环境中重新照亮被抓获的能力。鉴于现有的光台需要昂贵的、房间大小的球形捕获机架，并且世界上只有少数实验室存在，我们将演示如何从普通电视或桌面显示器获取有用的数据。我们对用户观看YouTube视频或其他标准内容的图像进行操作，而不是让用户感到不舒服的快速闪烁的灯光模式。我们在给定用户的图像和监控模式上训练深度网络，并学习在任何目标照明（监控模式）下预测该用户的图像。实验评估表明，我们的方法产生了真实的重新照明结果。

Barrel distortion rectification aims at removing the radial distortion in a distorted image captured by a wide-angle lens. Previous deep learning methods mainly solve this problem by learning the implicit distortion parameters or the nonlinear rectified mapping function in a direct manner. However, this type of manner results in an indistinct learning process of rectification and thus limits the deep perception of distortion. In this paper, inspired by the curriculum learning, we analyze the barrel distortion rectification task in a progressive and meaningful manner. By considering the relationship among different construction levels in an image, we design a multi-level curriculum that disassembles the rectification task into three levels, structure recovery, semantics embedding, and texture rendering. With the guidance of the curriculum that corresponds to the construction of images, the proposed hierarchical architecture enables a progressive rectification and achieves more accurate results. Moreover, we present a novel distortion-aware pre-training strategy to facilitate the initial learning of neural networks, promoting the model to converge faster and better. Experimental results on the synthesized and real-world distorted image datasets show that the proposed approach significantly outperforms other learning methods, both qualitatively and quantitatively.

桶形畸变校正旨在消除广角镜头拍摄的畸变图像中的径向畸变。以往的深度学习方法主要通过直接学习隐式失真参数或非线性校正映射函数来解决这一问题。然而，这种类型的方式导致矫正的模糊学习过程，从而限制了对扭曲的深度感知。本文受课程学习的启发，以渐进和有意义的方式分析了桶形失真校正任务。通过考虑图像中不同结构层次之间的关系，我们设计了一个多层次的课程，将矫正任务分解为三个层次：结构恢复、语义嵌入和纹理渲染。在与图像构建相对应的课程的指导下，所提出的层次结构能够逐步校正，并获得更准确的结果。此外，我们提出了一种新的失真感知预训练策略，以促进神经网络的初始学习，促进模型更快更好地收敛。在合成和真实扭曲图像数据集上的实验结果表明，该方法在定性和定量上都明显优于其他学习方法。

We present DepthInSpace, a self-supervised deep-learning method for depth estimation using a structured-light camera. The design of this method is motivated by the commercial use case of embedded depth sensors in nowadays smartphones. We first propose to use estimated optical flow from ambient information of multiple video frames as a complementary guide for training a single-frame depth estimation network, helping to preserve edges and reduce over-smoothing issues. Utilizing optical flow, we also propose to fuse the data of multiple video frames to get a more accurate depth map. In particular, fused depth maps are more robust in occluded areas and incur less in flying pixels artifacts. We finally demonstrate that these more precise fused depth maps can be used as self-supervision for fine-tuning a single-frame depth estimation network to improve its performance. Our models' effectiveness is evaluated and compared with state-of-the-art models on both synthetic and our newly introduced real datasets. The implementation code, training procedure, and both synthetic and captured real datasets are available at <https://www.idiap.ch/paper/depthinspace>.

我们提出了一种使用结构光相机进行深度估计的自监督深度学习方法DepthInSpace。该方法的设计是基于当今智能手机中嵌入式深度传感器的商业用例。我们首先提出使用多个视频帧的环境信息估计的光流作为训练单帧深度估计网络的补充指南，帮助保留边缘并减少过度平滑问题。利用光流，我们还建议融合多个视频帧的数据，以获得更精确的深度图。特别是，融合的深度贴图在被遮挡的区域更健壮，在飞行像素伪影中产生的伪影更少。最后，我们证明了这些更精确的融合深度图可以用作微调单帧深度估计网络的自我监控，以提高其性能。在合成数据集和新引入的真实数据集上，对我们的模型的有效性进行了评估，并与最先进的模型进行了比较。实现代码、培训程序以及合成和捕获的真实数据集可在<https://www.idiap.ch/paper/depthinspace>。

In this paper, we tackle the problem of dense light field (LF) reconstruction from sparsely-sampled ones with wide baselines and propose a learnable model, namely dynamic interpolation, to replace the commonly-used geometry warping operation. Specifically, with the estimated geometric relation between input views, we first construct a lightweight neural network to dynamically learn weights for interpolating neighbouring pixels from input views to synthesize each pixel of novel views independently. In contrast to the fixed and content-independent weights employed in the geometry warping operation, the learned interpolation weights implicitly incorporate the correspondences between the source and novel views and adapt to different image content information. Then, we recover the spatial correlation between the independently synthesized pixels of each novel view by referring to that of input views using a geometry-based spatial refinement module. We also constrain the angular correlation between the novel views through a disparity-oriented LF structure loss. Experimental results on LF datasets with wide baselines show that the reconstructed LFs achieve much higher PSNR/SSIM and preserve the LF parallax structure better than state-of-the-art methods. The source code is publicly available at <https://github.com/MantangGuo/DI4SLF>.

在本文中，我们解决了从宽基线稀疏采样光场（LF）重建稠密光场的问题，并提出了一种可学习的模型，即动态插值，以取代常用的几何扭曲操作。具体地说，利用输入视图之间的几何关系，我们首先构造一个轻量级的神经网络，动态地学习从输入视图插值相邻像素的权重，从而独立地合成新视图的每个像素。与几何扭曲操作中使用的固定权重和内容无关权重不同，学习的插值权重隐含地包含源视图和新视图之间的对应关系，并适应不同的图像内容信息。然后，通过使用基于几何的空间细化模块参考输入视图的像素，恢复每个新视图的独立合成像素之间的空间相关性。我们还通过面向视差的LF结构损耗来约束新视图之间的角度相关性。在宽基线LF数据集上的实验结果表明，与现有方法相比，重构LFs获得了更高的PSNR/SSIM，并且更好地保持了LF视差结构。源代码可在<https://github.com/MantangGuo/DI4SLF>。

Existing single image high dynamic range (HDR) reconstruction attempt to expand the range of luminance. They are not effective to generate plausible textures and colors in the reconstructed results, especially for high-density pixels in ultra-high-definition (UHD) images. To address these problems, we propose a new HDR reconstruction network for UHD images by collaboratively learning color and texture details. First, we propose a dual-path network to extract content and chromatic features at a reduced resolution of the low dynamic range (LDR) input. These two types features are used to fit bilateral-space affine models for real-time HDR reconstruction. To extract the main data structure of the LDR input, we propose to use 3D Tucker decomposition and reconstruction to prevent false edges and noise amplification in the learned bilateral grid. As a result, the high-quality content and chromatic features can be reconstructed capitalized on guided bilateral upsampling. Finally, we fuse these two full-resolution feature maps into the HDR reconstructed results. Our proposed method can achieve real-time processing for UHD image (about 160 fps). Experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art HDR reconstruction approaches on public benchmarks and real-world UHD images.

现有的单图像高动态范围（HDR）重建试图扩大亮度范围。它们不能有效地在重建结果中生成合理的纹理和颜色，特别是对于超高清（UHD）图像中的高密度像素。为了解决这些问题，我们提出了一种新的超高清图像HDR重建网络，通过协作学习颜色和纹理细节。首先，我们提出了一种双路径网络，以降低低动态范围（LDR）输入的分辨率来提取内容和颜色特征。这两类特征用于拟合双边空间仿射模型，以进行实时HDR重建。为了提取LDR输入的主要数据结构，我们建议使用3D Tucker分解和重建来防止学习的双边网格中的假边缘和噪声放大。因此，可以利用引导双边采样来重构高质量的内容和颜色特征。最后，我们将这两个全分辨率特征映射融合到HDR重建结果中。我们提出的方法可以实现UHD图像

(约160fps) 的实时处理。实验结果表明，该算法在公共基准和真实UHD图像上具有良好的HDR重建性能。

Computer vision applications such as visual relationship detection and human object interaction can be formulated as a composite (structured) set detection problem in which both the parts (subject, object, and predicate) and the sum (triplet as a whole) are to be detected in a hierarchical fashion. In this paper, we present a new approach, denoted Part-and-Sum detection Transformer (PST), to perform end-to-end visual composite set detection. Different from existing Transformers in which queries are at a single level, we simultaneously model the joint part and sum hypotheses/interactions with composite queries and attention modules. We explicitly incorporate sum queries to enable better modeling of the part-and-sum relations that are absent in the standard Transformers. Our approach also uses novel tensor-based part queries and vector-based sum queries, and models their joint interaction. We report experiments on two vision tasks, visual relationship detection and human object interaction and demonstrate that PST achieves state of the art results among single-stage models, while nearly matching the results of custom designed two-stage models.

计算机视觉应用，如视觉关系检测和人机交互，可以表述为一个复合（结构化）集合检测问题，其中部分（主语、宾语和谓语）和总和（三元组作为一个整体）都要以分层的方式进行检测。在本文中，我们提出了一种新的方法，表示部分和检测变压器（PST），以执行端到端的视觉组合集检测。与现有的变换器不同，在变换器中，查询处于单一级别，我们同时使用复合查询和注意模块对联合部分和总和假设/交互进行建模。我们明确地合并了求和查询，以便更好地建模标准转换器中缺少的部分和求和关系。我们的方法还使用新的基于张量的部分查询和基于向量的求和查询，并对它们的联合交互进行建模。我们报告了两个视觉任务的实验，视觉关系检测和人机交互，并证明PST在单阶段模型中实现了最先进的结果，同时几乎与定制设计的两阶段模型的结果相匹配。

Domain generalization (DG) aims to generalize a model trained on multiple source (i.e., training) domains to a distributionally different target (i.e., test) domain. In contrast to the DG setup that strictly requires the availability of multiple source domains, this paper considers a more realistic yet challenging scenario, namely Single Domain Generalization (SDG). In this new setting, there is only one source domain available for training, from which the limited diversity may jeopardize the model generalization on unseen target domains. To tackle this problem, we propose a style-complement module to enhance the generalization power of the model by synthesizing images from diverse distributions that are complementary to the source ones. More specifically, we adopt tractable upper and lower bounds of mutual information (MI) between the generated and source samples and perform the two-step optimization iteratively: (1) by minimizing MI upper bound approximation for each pair, the generated images are forced to diversify from the source samples; (2) subsequently, we maximize the lower bound of MI between the samples from the same semantic category, which assists the network to learn discriminative features from diverse-styled images. Extensive experiments on three benchmark datasets demonstrate the superiority of our approach, which surpasses the state-of-the-art single DG methods by up to 25.14%.

领域泛化 (DG) 旨在将在多个源 (即训练) 领域上训练的模型泛化到分布不同的目标 (即测试) 领域。与严格要求多个源域可用性的DG设置不同，本文考虑了更现实但更具挑战性的场景，即单域泛化 (SDG)。在这个新的设置中，只有一个源域可用于训练，有限的多样性可能会危及模型在看不见的目标域上的泛化。为了解决这个问题，我们提出了一个样式补码模块，通过合成与源分布互补的不同分布的图像来增强模型的泛化能力。更具体地说，我们在生成的样本和源样本之间采用可处理的互信息上下界 (MI)，并迭代执行两步优化：(1) 通过最小化每对互信息上下界近似，生成的图像被迫从源样本多样化；(2) 随后，我们最大化来自同一语义类别的样本之间的MI下限，这有助于网络从不同样式的

图像中学习区分特征。在三个基准数据集上的大量实验证明了我们的方法的优越性，它比最先进的单DG方法高达25.14%。

This paper proposes a framework to guide an optical flow network with external cues to achieve superior accuracy either on known or unseen domains. Given the availability of sparse yet accurate optical flow hints from an external source, these are injected to modulate the correlation scores computed by a state-of-the-art optical flow network and guide it towards more accurate predictions. Although no real sensor can provide sparse flow hints, we show how these can be obtained by combining depth measurements from active sensors with geometry and hand-crafted optical flow algorithms, leading to accurate enough hints for our purpose. Experimental results with a state-of-the-art flow network on standard benchmarks support the effectiveness of our framework, both in simulated and real conditions.

本文提出了一个框架，用外部线索引导光流网络在已知或不可见的域上实现更高的精度。考虑到来自外部源的稀疏但精确的光流提示的可用性，这些提示被注入以调制由最先进的光流网络计算的相关分数，并引导其进行更精确的预测。虽然没有真正的传感器可以提供稀疏的流量提示，但我们展示了如何通过将主动传感器的深度测量与几何和手工制作的光流算法相结合来获得这些提示，从而为我们的目的提供足够准确的提示。在标准基准上使用最先进的流网络进行的实验结果支持我们的框架在模拟和真实条件下的有效性。

Graphs play an important role in cross-modal image-text understanding as they characterize the intrinsic structure which is robust and crucial for the measurement of cross-modal similarity. In this work, we propose a Wasserstein Coupled Graph Learning (WCGL) method to deal with the cross-modal retrieval task. First, graphs are constructed according to two input cross-modal samples separately, and passed through the corresponding graph encoders to extract robust features. Then, a Wasserstein coupled dictionary, containing multiple pairs of counterpart graph keys with each key corresponding to one modality, is constructed for further feature learning. Based on this dictionary, the input graphs can be transformed into the dictionary space to facilitate the similarity measurement through a Wasserstein Graph Embedding (WGE) process. The WGE could capture the graph correlation between the input and each corresponding key through optimal transport, and hence well characterize the inter-graph structural relationship. To further achieve discriminant graph learning, we specifically define a Wasserstein discriminant loss on the coupled graph keys to make the intra-class (counterpart) keys more compact and inter-class (non-counterpart) keys more dispersed, which further promotes the final cross-modal retrieval task. Experimental results demonstrate the effectiveness and state-of-the-art performance.

图形在跨模态图像文本理解中起着重要的作用，因为图形表征了跨模态相似性度量的鲁棒性和关键性的内在结构。在这项工作中，我们提出了一种Wasserstein耦合图学习 (WCGL) 方法来处理跨模态检索任务。首先，根据两个输入交叉模态样本分别构造图，并通过相应的图编码器提取鲁棒特征。然后，构造一个Wasserstein耦合字典，包含多对对应的图键，每个键对应一个模态，用于进一步的特征学习。基于该字典，可以将输入图转换为字典空间，以便通过Wasserstein图嵌入 (WGE) 过程进行相似性度量。WGE可以通过最优传输捕获输入和每个对应键之间的图相关性，从而很好地描述图间的结构关系。为了进一步实现判别图学习，我们在耦合图键上专门定义了一个Wasserstein判别损失，使类内（对应）键更加紧凑，类间（非对应）键更加分散，从而进一步促进了最终的跨模式检索任务。实验结果证明了该方法的有效性和最先进的性能。

The recent progress of CNN has dramatically improved face alignment performance. However, few works have paid attention to the error-bias with respect to error distribution of facial landmarks. In this paper, we investigate the error-bias issue in face alignment, where the distributions of landmark errors tend to spread along the tangent line to landmark curves. This error-bias is not trivial since it is closely connected to the ambiguous landmark labeling task. Inspired by this observation, we seek a way to leverage the error-bias property for better convergence of CNN model. To this end, we propose anisotropic direction loss (ADL) and anisotropic attention module (AAM) for coordinate and heatmap regression, respectively. ADL imposes strong binding force in normal direction for each landmark point on facial boundaries. On the other hand, AAM is an attention module which can get anisotropic attention mask focusing on the region of point and its local edge connected by adjacent points, it has a stronger response in tangent than in normal, which means relaxed constraints in the tangent. These two methods work in a complementary manner to learn both facial structures and texture details. Finally, we integrate them into an optimized end-to-end training pipeline named ADNet. Our ADNet achieves state-of-the-art results on 300W, WFLW and COFW datasets, which demonstrates the effectiveness and robustness.

CNN的最新进展极大地改善了人脸对齐性能。然而，很少有研究关注人脸标志点误差分布的误差偏差。在本文中，我们研究了人脸对齐中的误差偏差问题，其中地标误差的分布趋向于沿着地标曲线的切线分布。这种错误偏差并非微不足道，因为它与模糊的地标标记任务密切相关。受这一观察结果的启发，我们寻求一种利用误差偏差特性更好地收敛CNN模型的方法。为此，我们分别提出了用于坐标回归和热图回归的各向异性方向损失（ADL）和各向异性注意模块（AAM）。ADL对面部边界上的每个标志点在法线方向施加强大的约束力。另一方面，AAM是一个注意模块，它可以将各向异性的注意掩模聚焦于点的区域及其由相邻点连接的局部边缘，它在切线上的响应比在法线上更强，这意味着切线上的约束松弛。这两种方法以互补的方式学习面部结构和纹理细节。最后，我们将它们集成到一个名为ADNet的优化端到端训练管道中。我们的ADNet在300W、WFLW和COFW数据集上实现了最先进的结果，这证明了它的有效性和健壮性。

Finding local features that are repeatable across multiple views is a cornerstone of sparse 3D reconstruction. The classical image matching paradigm detects keypoints per-image once and for all, which can yield poorly-localized features and propagate large errors to the final geometry. In this paper, we refine two key steps of structure-from-motion by a direct alignment of low-level image information from multiple views: we first adjust the initial keypoint locations prior to any geometric estimation, and subsequently refine points and camera poses as a post-processing. This refinement is robust to large detection noise and appearance changes, as it optimizes a featuremetric error based on dense features predicted by a neural network. This significantly improves the accuracy of camera poses and scene geometry for a wide range of keypoint detectors, challenging viewing conditions, and off-the-shelf deep features. Our system easily scales to large image collections, enabling pixel-perfect crowd-sourced localization at scale. Our code is publicly available at <https://github.com/cvg/pixel-perfect-sfm> as an add-on to the popular SfM software COLMAP.

寻找可在多个视图中重复的局部特征是稀疏三维重建的基础。经典的图像匹配范式一次性地检测每个图像的关键点，这可能会产生局部性差的特征，并将较大的误差传播到最终的几何体。在本文中，我们通过直接对齐来自多个视图的低级图像信息，从运动中细化结构的两个关键步骤：首先在任何几何估计之前调整初始关键点位置，然后作为后处理细化点和相机姿势。这种改进对大的检测噪声和外观变化具有鲁棒性，因为它优化了基于神经网络预测的密集特征的特征度量误差。这显著提高了各种关键点探测器、具有挑战性的观察条件和现成深度特征的摄影机姿势和场景几何体的准确性。我们的系统可以轻松

扩展到大型图像集，实现大规模像素完美的众包定位。我们的代码在<https://github.com/cvg/pixel-perf-ect-sfm>作为流行的SfM软件COLMAP的附加组件。

We address the problem of network quantization, that is, reducing bit-widths of weights and/or activations to lighten network architectures. Quantization methods use a rounding function to map full-precision values to the nearest quantized ones, but this operation is not differentiable. There are mainly two approaches to training quantized networks with gradient-based optimizers. First, a straight-through estimator (STE) replaces the zero derivative of the rounding with that of an identity function, which causes a gradient mismatch problem. Second, soft quantizers approximate the rounding with continuous functions at training time, and exploit the rounding for quantization at test time. This alleviates the gradient mismatch, but causes a quantizer gap problem. We alleviate both problems in a unified framework. To this end, we introduce a novel quantizer, dubbed a distance-aware quantizer (DAQ), that mainly consists of a distance-aware soft rounding (DASR) and a temperature controller. To alleviate the gradient mismatch problem, DASR approximates the discrete rounding with the kernel soft argmax, which is based on our insight that the quantization can be formulated as a distance-based assignment problem between full-precision values and quantized ones. The controller adjusts the temperature parameter in DASR adaptively according to the input, addressing the quantizer gap problem. Experimental results on standard benchmarks show that DAQ outperforms the state of the art significantly for various bit-widths without bells and whistles.

我们解决了网络量化问题，即减少权重和/或激活的比特宽度以减轻网络架构。量化方法使用舍入函数将全精度值映射到最近的量化值，但此操作是不可微的。使用基于梯度的优化器训练量化网络主要有两种方法。首先，直通估计器（STE）将舍入的零导数替换为单位函数的零导数，这会导致梯度失配问题。其次，软量化器在训练时用连续函数近似取整，并在测试时利用取整进行量化。这会减轻梯度失配，但会导致量化器间隙问题。我们在一个统一的框架内缓解了这两个问题。为此，我们介绍了一种新型量化器，称为距离感知量化器（DAQ），它主要由距离感知软舍入（DASR）和温度控制器组成。为了缓解梯度失配问题，DASR使用内核软件argmax近似离散舍入，这是基于我们的见解，即量化可以表示为全精度值和量化值之间基于距离的分配问题。控制器根据输入自适应调整DASR中的温度参数，解决了量化器间隙问题。在标准基准上的实验结果表明，在不同的比特宽度下，DAQ的性能明显优于现有技术。

There is an emerging sense that the vulnerability of Image Convolutional Neural Networks (CNN), i.e., sensitivity to image corruptions, perturbations, and adversarial attacks, is connected with Texture Bias. This relative lack of shape Bias is also responsible for poor performance in Domain Generalization (DG). The inclusion of a role of shape alleviates these vulnerabilities and some approaches have achieved this by training on negative images, images endowed with edge maps, or images with conflicting shape and texture information. This paper advocates an explicit and complete representation of shape using a classical computer vision approach, namely, representing the shape content of an image with the shock graph of its contour map. The resulting graph and its descriptor is a complete representation of contour content and is classified using recent Graph Neural Network (GNN) methods. The experimental results on three domain shift datasets, Colored MNIST, PACS, and VLCS demonstrate that even without using appearance the shape-based approach exceeds classical Image CNN based methods in domain generalization.

人们逐渐意识到，图像卷积神经网络（CNN）的脆弱性，即对图像损坏、干扰和敌对攻击的敏感性，与纹理偏差有关。这种相对缺乏形状偏差的情况也是导致领域泛化（DG）性能差的原因。包含形状角色可以缓解这些漏洞，一些方法通过对负面图像、具有边缘贴图的图像或具有冲突形状和纹理信息的图像进行训练来实现这一点。本文提倡使用经典的计算机视觉方法对形状进行显式和完整的表示，即用轮廓图的冲击图表示图像的形状内容。生成的图形及其描述符是轮廓内容的完整表示，并使用最新的图形神经

网络 (GNN) 方法进行分类。在三个域移位数据集、彩色MNIST、PACS和VLCS上的实验结果表明，即使不使用外观，基于形状的方法在域泛化方面也超过了经典的基于图像CNN的方法。

Video content creation keeps growing at an incredible pace; yet, creating engaging stories remains challenging and requires non-trivial video editing expertise. Many video editing components are astonishingly hard to automate primarily due to the lack of raw video materials. This paper focuses on a new task for computational video editing, namely the task of raking cut plausibility. Our key idea is to leverage content that has already been edited to learn fine-grained audiovisual patterns that trigger cuts. To do this, we first collected a data source of more than 10K videos, from which we extract more than 260K cuts. We devise a model that learns to discriminate between real and artificial cuts via contrastive learning. We set up a new task and a set of baselines to benchmark video cut generation. We observe that our proposed model outperforms the baselines by large margins. To demonstrate our model in real-world applications, we conduct human studies in a collection of unedited videos. The results show that our model does a better job at cutting than random and alternative baselines.

视频内容创作以惊人的速度持续增长；然而，创造引人入胜的故事仍然具有挑战性，需要非同寻常的视频编辑专业知识。由于缺乏原始视频材料，许多视频编辑组件很难实现自动化。本文重点研究了计算视频编辑的一项新任务，即倾斜切割合理性任务。我们的关键思想是利用已经编辑过的内容来学习触发剪切的细粒度视听模式。为此，我们首先收集了超过10万个视频的数据源，从中提取了超过260万个剪辑。我们设计了一个模型，通过对比学习来区分真实切割和人工切割。我们建立了一个新的任务和一组基准来测试视频剪辑生成。我们观察到，我们提出的模型大大优于基线。为了在实际应用中演示我们的模型，我们在一组未经编辑的视频中进行人体研究。结果表明，我们的模型在切割方面比随机基线和替代基线做得更好。

This paper presents a novel task together with a new benchmark for detecting generic, taxonomy-free event boundaries that segment a whole video into chunks. Conventional work in temporal video segmentation and action detection focuses on localizing pre-defined action categories and thus does not scale to generic videos. Cognitive Science has known since last century that humans consistently segment videos into meaningful temporal chunks. This segmentation happens naturally, without pre-defined event categories and without being explicitly asked to do so. Here, we repeat these cognitive experiments on mainstream CV datasets; with our novel annotation guideline which addresses the complexities of taxonomy-free event boundary annotation, we introduce the task of Generic Event Boundary Detection (GEBD) and the new benchmark Kinetics-GEBD. We view GEBD as an important stepping stone towards understanding the video as a whole, and believe it has been previously neglected due to a lack of proper task definition and annotations. Through experiment and human study we demonstrate the value of the annotations. Further, we benchmark supervised and un-supervised GEBD approaches on the TAPOS dataset and our Kinetics-GEBD. We release our annotations and baseline codes at CVPR'21 LOVEU Challenge:  
<https://sites.google.com/view/loveucvpr21>.

本文提出了一个新的任务和一个新的基准，用于检测将整个视频分割成块的通用、无分类的事件边界。时间视频分割和动作检测的传统工作侧重于定位预定义的动作类别，因此无法扩展到通用视频。自上个世纪以来，认知科学已经知道，人类总是将视频分割成有意义的时间块。这种分割是自然发生的，没有预定义的事件类别，也没有明确要求这样做。在这里，我们在主流CV数据集上重复这些认知实验；我们的新注释指南解决了无分类事件边界注释的复杂性，我们介绍了通用事件边界检测（GEBD）的任务和新的基准动力学GEBD。我们将GEBD视为理解整个视频的一个重要垫脚石，并且认为由于缺乏适当的任务定义和注释，它以前被忽略了。通过实验和人体研究，我们证明了注释的价值。此外，我们在TAPOS

数据集和我们的动力学GEBD上对有监督和无监督的GEBD方法进行基准测试。我们在CVPR的21 LOVEU 挑战赛上发布了注释和基线代码：<https://sites.google.com/view/loveucvpr21>.

Metric learning has received conflicting assessments concerning its suitability for solving instance segmentation tasks. It has been dismissed as theoretically flawed due to the shift equivariance of the employed CNNs and their respective inability to distinguish same-looking objects. Yet it has been shown to yield state of the art results for a variety of tasks, and practical issues have mainly been reported in the context of tile-and-stitch approaches, where discontinuities at tile boundaries have been observed. To date, neither of the reported issues have undergone thorough formal analysis. In our work, we contribute a comprehensive formal analysis of the shift equivariance properties of encoder-decoder-style CNNs, which yields a clear picture of what can and cannot be achieved with metric learning in the face of same-looking objects. In particular, we prove that a standard encoder-decoder network that takes  $d$ -dimensional images as input, with  $l$  pooling layers and pooling factor  $f$ , has the capacity to distinguish at most  $f^{(d)} l$  same-looking objects, and we show that this upper limit can be reached. Furthermore, we show that to avoid discontinuities in a tile-and-stitch approach, assuming standard batch size 1, it is necessary to employ valid convolutions in combination with a training output window size strictly greater than  $f^l$ , while at test-time it is necessary to crop tiles to size  $n * f^l$  before stitching, with  $n \geq 1$ . We complement these theoretical findings by discussing a number of insightful special cases for which we show empirical results on synthetic and real data.

关于度量学习是否适合解决实例分割任务，人们对其进行了相互矛盾的评估。由于所使用的CNN的位移等效性以及它们各自无法区分相同外观的对象，因此它在理论上存在缺陷，因此被驳回。然而，它已被证明能在各种任务中产生最先进的结果，实际问题主要是在瓷砖和缝合方法的背景下报告的，其中观察到瓷砖边界处的不连续性。迄今为止，所报告的两个问题都没有经过彻底的正式分析。在我们的工作中，我们对编码器-解码器风格的CNN的移位等变特性进行了全面的形式化分析，这清楚地说明了在面对相同外观的对象时，度量学习可以实现什么和不能实现什么。特别是，我们证明了一个以 $d$ 维图像为输入，具有 $l$ 个池层和池因子 $f$ 的标准编码器-解码器网络具有最多区分 $f^{(d)}$ 相同外观对象的能力，并且我们证明了可以达到这个上限。此外，我们还表明，为了避免平铺和缝合方法中的不连续性，假设标准批量大小为1，则有必要结合严格大于 $f^l$ 的训练输出窗口大小使用有效卷积，而在测试时，有必要在缝合之前将平铺裁剪为大小为 $n*f^l$ ，且 $n \geq 1$ 。我们通过讨论一些有见地的特殊案例来补充这些理论发现，我们在合成和真实数据上展示了这些案例的实证结果。

In semantic segmentation tasks, input images can often have more than one plausible interpretation, thus allowing for multiple valid labels. To capture such ambiguities, recent work has explored the use of probabilistic networks that can learn a distribution over predictions. However, these do not necessarily represent the empirical distribution accurately. In this work, we present a strategy for learning a calibrated predictive distribution over semantic maps, where the probability associated with each prediction reflects its ground truth correctness likelihood. To this end, we propose a novel two-stage, cascaded approach for calibrated adversarial refinement: (i) a standard segmentation network is trained with categorical cross-entropy to predict a pixelwise probability distribution over semantic classes and (ii) an adversarially trained stochastic network is used to model the inter-pixel correlations to refine the output of the first network into coherent samples. Importantly, to calibrate the refinement network and prevent mode collapse, the expectation of the samples in the second stage is matched to the probabilities predicted in the first. We demonstrate the versatility and robustness of the approach by achieving state-of-the-art results on the multigrader LIDC dataset and on a modified Cityscapes dataset with injected ambiguities. In addition, we show that the core design can be adapted to other tasks requiring learning a calibrated predictive distribution by experimenting on a toy regression dataset. We provide an open source implementation of our method at <https://github.com/EliasKassapis/CARSSS>.

在语义分割任务中，输入图像通常可以有多个合理的解释，从而允许多个有效标签。为了捕捉这种模糊性，最近的工作探索了概率网络的使用，这种网络可以通过预测了解分布。然而，这些并不一定准确地代表经验分布。在这项工作中，我们提出了一种学习语义映射上校准预测分布的策略，其中与每个预测相关联的概率反映了其基本真理正确性可能性。为此，我们提出了一种新颖的两阶段，校准敌对细化的级联方法：(i) 使用分类交叉熵训练标准分割网络，以预测语义类上的像素概率分布；(ii) 使用敌对训练的随机网络对像素间相关性建模，以细化第一个网络的输出转换为相干样本。重要的是，为了校准细化网络并防止模式崩溃，第二阶段中样本的期望值与第一阶段中预测的概率相匹配。我们通过在多级LIDC数据集和具有注入模糊性的改进城市景观数据集上获得最先进的结果，证明了该方法的多功能性和鲁棒性。此外，我们通过在玩具回归数据集上的实验表明，核心设计可以适应需要学习校准预测分布的其他任务。我们在<https://github.com/EliasKassapis/CARSSS>.

LiDAR point clouds collected from a moving vehicle are functions of its trajectories, because the sensor motion needs to be compensated to avoid distortions. When autonomous vehicles are sending LiDAR point clouds to deep networks for perception and planning, could the motion compensation consequently become a wide-open backdoor in those networks, due to both the adversarial vulnerability of deep learning and GPS-based vehicle trajectory estimation that is susceptible to wireless spoofing? We demonstrate such possibilities for the first time: instead of directly attacking point cloud coordinates which requires tampering with the raw LiDAR readings, only adversarial spoofing of a self-driving car's trajectory with small perturbations is enough to make safety-critical objects undetectable or detected with incorrect positions. Moreover, polynomial trajectory perturbation is developed to achieve a temporally-smooth and highly-imperceptible attack. Extensive experiments on 3D object detection have shown that such attacks not only lower the performance of the state-of-the-art detectors effectively, but also transfer to other detectors, raising a red flag for the community. The code is available on <https://ai4ce.github.io/FLAT/>.

从移动车辆收集的激光雷达点云是其轨迹的函数，因为需要对传感器运动进行补偿以避免失真。当自动驾驶车辆将激光雷达点云发送到深度网络进行感知和规划时，由于深度学习和基于GPS的车辆轨迹估计的对抗性漏洞易受无线欺骗的影响，运动补偿是否会因此成为这些网络中的一个大开后门？我们首次展示了这种可能性：与其直接攻击需要篡改原始激光雷达读数的点云坐标，只需在小扰动下对自动驾驶汽车的轨迹进行对抗性欺骗，就足以使安全关键对象无法检测或位置不正确。此外，多项式轨迹摄动的发

展，以实现时间平滑和高度不可察觉的攻击。大量的3D目标检测实验表明，此类攻击不仅有效地降低了最先进探测器的性能，而且还会转移到其他探测器，给社区带来危险。该代码可在<https://ai4ce.github.io/FLAT/>.

This paper proposes an extreme structure from motion (SfM) algorithm for residential indoor panoramas that have little to no visual overlaps. Only a single panorama is present in a room for many cases, making the task infeasible for existing SfM algorithms. Our idea is to learn to evaluate the realism of room/door/window arrangements in the top-down semantic space. After using heuristics to enumerate possible arrangements based on door detections, we evaluate their realism scores, pick the most realistic arrangement, and return the corresponding camera poses. We evaluate the proposed approach on a dataset of 1029 panorama images with 286 houses. Our qualitative and quantitative evaluations show that an existing SfM approach completely fails for most of the houses. The proposed approach achieves the mean positional error of less than 1.0 meter for 47% of the houses and even 78% when considering the top five reconstructions. We will share the code and data in <https://github.com/aminshabani/extreme-indoor-sfm>.

本文提出了一种基于运动的极限结构 (SfM) 算法，该算法适用于几乎没有视觉重叠的住宅室内全景图。在许多情况下，房间中只有一张全景图，这使得现有的SfM算法无法执行该任务。我们的想法是学习评估自上而下语义空间中房间/门/窗安排的真实性。在使用启发式算法枚举基于门检测的可能排列之后，我们评估它们的真实性分数，选择最真实的排列，并返回相应的相机姿势。我们在一个包含1029幅全景图像和286栋房屋的数据集上评估了所提出的方法。我们的定性和定量评估表明，现有的SfM方法在大多数房屋中完全失败。所提出的方法在47%的房屋中实现了小于1.0米的平均位置误差，甚至在考虑前五位重建时达到了78%。我们将在中共享代码和数据<https://github.com/aminshabani/extreme-indoor-sfm>。

The ability to capture inter-frame dynamics has been critical to the development of video salient object detection (VSOD). While many works have achieved great success in this field, a deeper insight into its dynamic nature should be developed. In this work, we aim to answer the following questions: How can a model adjust itself to dynamic variations as well as perceive fine differences in the real-world environment; How are the temporal dynamics well introduced into spatial information over time? To this end, we propose a dynamic context-sensitive filtering network (DCFNet) equipped with a dynamic context-sensitive filtering module (DCFM) and an effective bidirectional dynamic fusion strategy. The proposed DCFM sheds new light on dynamic filter generation by extracting location-related affinities between consecutive frames. Our bidirectional dynamic fusion strategy encourages the interaction of spatial and temporal information in a dynamic manner. Experimental results demonstrate that our proposed method can achieve state-of-the-art performance on most VSOD datasets while ensuring a real-time speed of 28 fps. The source code is publicly available at <https://github.com/OIPLab-DUT/DCFNet>.

捕获帧间动态的能力对于视频显著目标检测 (VSOD) 的发展至关重要。虽然许多作品在这一领域取得了巨大的成功，但对其动态性质的深入了解仍有待发展。在这项工作中，我们的目标是回答以下问题：模型如何调整自身以适应动态变化，以及如何感知真实世界环境中的细微差异；随着时间的推移，如何将时间动态很好地引入空间信息？为此，我们提出了一种动态上下文敏感过滤网络 (DCFNet)，该网络配备了动态上下文敏感过滤模块 (DCFM) 和有效的双向动态融合策略。提出的DCFM通过提取连续帧之间的位置相关亲和力，为动态滤波器的生成提供了新的思路。我们的双向动态融合策略鼓励以动态方式进行空间和时间信息的交互。实验结果表明，我们提出的方法可以在大多数VSOD数据集上实现最先进的性能，同时确保28 fps的实时速度。源代码可在<https://github.com/OIPLab-DUT/DCFNet>。

We demonstrate that it is possible to perform face-related computer vision in the wild using synthetic data alone. The community has long enjoyed the benefits of synthesizing training data with graphics, but the domain gap between real and synthetic data has remained a problem, especially for human faces. Researchers have tried to bridge this gap with data mixing, domain adaptation, and domain-adversarial training, but we show that it is possible to synthesize data with minimal domain gap, so that models trained on synthetic data generalize to real in-the-wild datasets. We describe how to combine a procedurally-generated parametric 3D face model with a comprehensive library of hand-crafted assets to render training images with unprecedented realism and diversity. We train machine learning systems for face-related tasks such as landmark localization and face parsing, showing that synthetic data can both match real data in accuracy, as well as open up new approaches where manual labeling would be impossible.

我们证明，仅使用合成数据就可以在野外执行与人脸相关的计算机视觉。社区长期以来享受着用图形合成训练数据的好处，但真实数据和合成数据之间的领域差距仍然是一个问题，特别是对于人脸而言。研究人员试图通过数据混合、领域适应和领域对抗性训练来弥合这一差距，但我们表明，以最小的领域差距合成数据是可能的，因此，在合成数据上训练的模型可以在野外数据集中推广到实际。我们描述了如何将程序生成的参数化三维人脸模型与手工制作的综合资源库相结合，以前所未有的真实感和多样性呈现训练图像。我们训练机器学习系统来完成与人脸相关的任务，如地标定位和人脸解析，结果表明，合成数据既可以精确地匹配真实数据，也可以在无法手动标记的情况下开辟新的方法。

Future segmentation prediction aims to predict the segmentation masks for unobserved future frames. Most existing works addressed it by directly predicting the intermediate features extracted by existing segmentation models. However, these segmentation features are learned to be local discriminative (with rich details) and are always of high resolution/dimension. Hence, the complicated spatio-temporal variations of these features are difficult to predict, which motivates us to learn a more predictive representation. In this work, we develop a novel framework called Predictive Feature Autoencoder. In the proposed framework, we construct an autoencoder which serves as a bridge between the segmentation features and the predictor. In the latent feature learned by the autoencoder, global structures are enhanced and local details are suppressed so that it is more predictive. In order to reduce the risk of vanishing the suppressed details during recurrent feature prediction, we further introduce a reconstruction constraint in the prediction module. Extensive experiments show the effectiveness of the proposed approach and our method outperforms state-of-the-arts by a considerable margin.

未来分割预测的目的是预测未观测到的未来帧的分割模板。大多数现有的工作通过直接预测现有分割模型提取的中间特征来解决这一问题。然而，这些分割特征被学习为具有局部辨别性（具有丰富的细节），并且总是具有高分辨率/维度。因此，这些特征的复杂时空变化很难预测，这促使我们学习更具预测性的表示。在这项工作中，我们开发了一个新的框架称为预测特征自动编码器。在该框架中，我们构造了一个自动编码器，作为分割特征和预测器之间的桥梁。在由自动编码器学习的潜在特征中，全局结构被增强，局部细节被抑制，从而更具预测性。为了降低在重复特征预测过程中被抑制细节消失的风险，我们在预测模块中进一步引入了重构约束。大量的实验表明了该方法的有效性，我们的方法比现有的方法有很大的优势。

There exists many powerful architectures for object detection and semantic segmentation of both biomedical and natural images. However, a difficulty arises in the ability to create training datasets that are large and well-varied. The importance of this subject is nested in the amount of training data that artificial neural networks need to accurately identify and segment objects in images and the infeasibility of acquiring a sufficient dataset within the biomedical field. This paper introduces a new data augmentation method that generates artificial cell nuclei microscopical images along with their correct semantic segmentation labels. Data augmentation provides a step toward accessing higher generalization capabilities of artificial neural networks. An initial set of segmentation objects is used with Greedy AutoAugment to find the strongest performing augmentation policies. The found policies and the initial set of segmentation objects are then used in the creation of the final artificial images. When comparing the state-of-the-art data augmentation methods with the proposed method, the proposed method is shown to consistently outperform current solutions in the generation of nuclei microscopical images.

在生物医学图像和自然图像的目标检测和语义分割方面存在着许多强大的体系结构。然而，创建大型且变化很大的训练数据集的能力存在困难。本课题的重要性在于人工神经网络需要大量的训练数据来准确识别和分割图像中的对象，以及在生物医学领域获取足够数据集的不可行性。本文介绍了一种新的数据增强方法，该方法生成人工细胞核显微图像及其正确的语义分割标签。数据扩充为获得人工神经网络更高的泛化能力提供了一个步骤。初始分割对象集与贪婪自动增强一起使用，以找到性能最强的增强策略。找到的策略和分割对象的初始集合然后用于创建最终的人工图像。当将最新的数据增强方法与所提出的方法进行比较时，所提出的方法在生成核显微图像方面始终优于当前的解决方案。

In this work, we present a new multi-view depth estimation method that utilizes both conventional SfM reconstruction and learning-based priors over the recently proposed neural radiance fields (NeRF). Unlike existing neural network based optimization method that relies on estimated correspondences, our method directly optimizes over implicit volumes, eliminating the challenging step of matching pixels in indoor scenes. The key to our approach is to utilize the learning-based priors to guide the optimization process of NeRF. Our system firstly adapts a monocular depth network over the target scene by finetuning on its sparse SfM reconstruction. Then, we show that the shape-radiance ambiguity of NeRF still exists in indoor environments and propose to address the issue by employing the adapted depth priors to monitor the sampling process of volume rendering. Finally, a per-pixel confidence map acquired by error computation on the rendered image can be used to further improve the depth quality. Experiments show that our proposed framework significantly outperforms state-of-the-art methods on indoor scenes, with surprising findings presented on the effectiveness of correspondence-based optimization and NeRF-based optimization over the adapted depth priors. In addition, we show that the guided optimization scheme does not sacrifice the original synthesis capability of neural radiance fields, improving the rendering quality on both seen and novel views. Code is available at <https://github.com/weiyithu/NerfingMVS>.

在这项工作中，我们提出了一种新的多视角深度估计方法，该方法在最近提出的神经辐射场（NeRF）上利用传统的SfM重建和基于学习的先验知识。与现有的基于神经网络的优化方法依赖于估计的对应关系不同，我们的方法直接在隐式体积上进行优化，消除了室内场景中像素匹配的挑战性步骤。该方法的关键是利用基于学习的先验知识指导神经网络的优化过程。我们的系统首先通过对稀疏SfM重建进行微调，在目标场景上采用单目深度网络。然后，我们证明了NeRF的形状辐射模糊性在室内环境中仍然存在，并建议通过使用自适应深度先验来监控体绘制的采样过程来解决该问题。最后，通过对渲染图像进行误差计算获得的每像素置信度图可用于进一步改善深度质量。实验表明，我们提出的框架在室内场景中的性能明显优于最先进的方法，基于通信的优化和基于NeRF的优化在自适应深度先验上的有效性令人

惊讶。此外，我们还证明了引导优化方案不会牺牲神经辐射场的原始合成能力，从而提高了可视和新视图的渲染质量。代码可在<https://github.com/weiyithu/NerfingMVS>.

Context is of fundamental importance to both human and machine vision; e.g., an object in the air is more likely to be an airplane than a pig. The rich notion of context incorporates several aspects including physics rules, statistical co-occurrences, and relative object sizes, among others. While previous work has focused on crowd-sourced out-of-context photographs from the web to study scene context, controlling the nature and extent of contextual violations has been a daunting task. Here we introduce a diverse, synthetic out-of-Context Dataset (OCD) with fine-grained control over scene context. By leveraging a 3D simulation engine, we systematically control the gravity, object co-occurrences and relative sizes across 36 object categories in a virtual household environment. We conducted a series of experiments to gain insights into the impact of contextual cues on both human and machine vision using OCD. We conducted psychophysics experiments to establish a human benchmark for out-of-context recognition and then compared it with state-of-the-art computer vision models to quantify the gap between the two. We propose a context-aware recognition transformer model, fusing object and contextual information via multi-head attention. Our model captures useful information for contextual reasoning, enabling human-level performance and better robustness in out-of-context conditions compared to baseline models across OCD and other out-of-context datasets. All source code and data are publicly available at <https://github.com/kreimanlab/WhenPigsFlyContext>

上下文对人类和机器视觉都至关重要；e. 例如，空中的物体更可能是飞机而不是猪。丰富的上下文概念包含几个方面，包括物理规则、统计共现和相对对象大小等。虽然之前的工作重点是从网络上众包背景外的照片来研究场景背景，但控制背景违反的性质和程度一直是一项艰巨的任务。这里，我们介绍一个多样化的、合成的上下文外数据集（OCD），它对场景上下文进行细粒度控制。通过利用3D模拟引擎，我们系统地控制虚拟家庭环境中36个对象类别的重力、对象共现和相对大小。我们使用强迫症进行了一系列实验，以深入了解上下文线索对人类和机器视觉的影响。我们进行了心理物理学实验，以建立一个人类上下文外识别的基准，然后将其与最先进的计算机视觉模型进行比较，以量化两者之间的差距。我们提出了一个上下文感知识别转换模型，通过多头注意融合对象和上下文信息。我们的模型捕获了用于上下文推理的有用信息，与OCD和其他上下文外数据集的基线模型相比，在上下文外条件下实现了人的水平性能和更好的鲁棒性。所有源代码和数据均可在<https://github.com/kreimanlab/WhenPigsFlyContext>

Humans are able to continuously detect and track surrounding objects by constructing a spatial-temporal memory of the objects when looking around. In contrast, 3D object detectors in existing tracking-by-detection systems often search for objects in every new video frame from scratch, without fully leveraging memory from previous detection results. In this work, we propose a novel system for integrated 3D object detection and tracking, which uses a dynamic object occupancy map and previous object states as spatial-temporal memory to assist object detection in future frames. This memory, together with the ego-motion from back-end odometry, guides the detector to achieve more efficient object proposal generation and more accurate object state estimation. The experiments demonstrate the effectiveness of the proposed system and its performance on the ScanNet and KITTI datasets. Moreover, the proposed system produces stable bounding boxes and pose trajectories over time, while being able to handle occluded and truncated objects. Code is available at the project page: <https://zju3dv.github.io/UDOLo>.

人类在环顾四周时，通过构建物体的时空记忆，能够连续地检测和跟踪周围的物体。相比之下，现有跟踪检测系统中的3D目标检测器通常从头开始搜索每个新视频帧中的目标，而没有充分利用以前检测结果的内存。在这项工作中，我们提出了一种新的集成三维目标检测和跟踪系统，该系统使用动态对象占用图和以前的对象状态作为时空记忆，以帮助在未来的帧中进行对象检测。这种记忆，加上来自后端里程计的自我运动，引导检测器实现更有效的对象建议生成和更准确的对象状态估计。在ScanNet和KITTI数据集上的实验证明了该系统的有效性和性能。此外，该系统可以生成稳定的边界框和随时间变化的姿态轨迹，同时能够处理遮挡和截断的对象。代码位于项目页面：<https://zju3dv.github.io/UDOLo>.

For unsupervised image-to-image translation, we propose a discriminator architecture which focuses on the statistical features instead of individual patches. The network is stabilized by distribution matching of key statistical features at multiple scales. Unlike the existing methods which impose more and more constraints on the generator, our method facilitates the shape deformation and enhances the fine details with a greatly simplified framework. We show that the proposed method outperforms the existing state-of-the-art models in various challenging applications including selfie-to-anime, male-to-female and glasses removal.

对于无监督的图像到图像的转换，我们提出了一种鉴别器结构，它关注于统计特征而不是单个面片。通过在多尺度上对关键统计特征进行分布匹配，使网络稳定。与现有的对生成器施加越来越多约束的方法不同，我们的方法简化了框架，有利于形状变形并增强了细节。我们表明，在各种具有挑战性的应用中，所提出的方法优于现有的最先进的模型，包括从自拍到动画、从男性到女性以及摘下眼镜。

In the low-bit quantization field, training Binarized Neural Networks (BNNs) is the extreme solution to ease the deployment of deep models on resource-constrained devices, having the lowest storage cost and significantly cheaper bit-wise operations compared to 32-bit floating-point counterparts. In this paper, we introduce Sub-bit Neural Networks (SNNs), a new type of binary quantization design tailored to compress and accelerate BNNs. SNNs are inspired by an empirical observation, showing that binary kernels learnt at convolutional layers of a BNN model are likely to be distributed over kernel subsets. As a result, unlike existing methods that binarize weights one by one, SNNs are trained with a kernel-aware optimization framework, which exploits binary quantization in the fine-grained convolutional kernel space. Specifically, our method includes a random sampling step generating layer-specific subsets of the kernel space, and a refinement step learning to adjust these subsets of binary kernels via optimization. Experiments on visual recognition benchmarks and the hardware deployment on FPGA validate the great potentials of SNNs. For instance, on ImageNet, SNNs of ResNet-18/ResNet-34 with 0.56-bit weights achieve 3.13/3.33 times runtime speed-up and 1.8 times compression over conventional BNNs with moderate drops in recognition accuracy. Promising results are also obtained when applying SNNs to binarize both weights and activations. Our code is available at <https://github.com/yikaiw/SNN>.

在低位量化领域，训练二值化神经网络（BNN）是简化资源受限设备上深度模型部署的极端解决方案，与32位浮点型相比，具有最低的存储成本和显著更低的位操作成本。在本文中，我们介绍了子比特神经网络（SNN），一种新的二进制量化设计，专门用于压缩和加速BNN。SNN的灵感来自于一项经验观察，表明在BNN模型的卷积层学习到的二进制核可能分布在核子集上。因此，与现有的逐个对权重进行二值化的方法不同，SNN是使用核感知优化框架进行训练的，该框架利用细粒度卷积核空间中的二值量化。具体地说，我们的方法包括一个随机采样步骤，生成特定于层的内核空间子集，以及一个细化步骤，学习通过优化调整这些二进制内核子集。视觉识别基准测试和FPGA硬件部署实验证了SNN的巨大潜力。例如，在ImageNet上，与传统BNN相比，具有0.56位权重的ResNet-18/ResNet-34的SNN实现了3.13/3.33倍的运行时加速和1.8倍的压缩，识别精度略有下降。当应用SNN对权重和激活进行二值化时，也获得了有希望的结果。我们的代码可在<https://github.com/yikaiw/SNN>.

Overconfident predictions on out-of-distribution (OOD) samples is a thorny issue for deep neural networks. The key to resolve the OOD overconfidence issue inherently is to build a subset of OOD samples and then suppress predictions on them. This paper proposes the Chamfer OOD examples (CODEs), whose distribution is close to that of in-distribution samples, and thus could be utilized to alleviate the OOD overconfidence issue effectively by suppressing predictions on them. To obtain CODEs, we first generate seed OOD examples via slicing&splicing operations on in-distribution samples from different categories, and then feed them to the Chamfer generative adversarial network for distribution transformation, without accessing to any extra data. Training with suppressing predictions on CODEs is validated to alleviate the OOD overconfidence issue largely without hurting classification accuracy, and outperform the state-of-the-art methods. Besides, we demonstrate CODEs are useful for improving OOD detection and classification.

对于深度神经网络来说，对分布外（OOD）样本的过度自信预测是一个棘手的问题。从本质上解决OOD过度自信问题的关键是建立OOD样本子集，然后抑制对它们的预测。本文提出了分布接近于分布内样本分布的样本（代码），通过抑制对样本的预测，可以有效地缓解OOD过度自信问题。为了获得代码，我们首先通过对来自不同类别的分布内样本进行切片和拼接操作生成种子OOD样本，然后将它们馈送到CHARM生成对抗网络进行分布转换，而无需访问任何额外数据。通过对代码进行抑制预测的训练，在不影响分类精度的情况下，大大缓解了OOD过度自信问题，并且优于最先进的方法。此外，我们还证明了代码对于改进OOD检测和分类是有用的。

We address the problem of scene layout generation for diverse domains such as images, mobile applications, documents, and 3D objects. Most complex scenes, natural or human-designed, can be expressed as a meaningful arrangement of simpler compositional graphical primitives. Generating a new layout or extending an existing layout requires understanding the relationships between these primitives. To do this, we propose LayoutTransformer, a novel framework that leverages self-attention to learn contextual relationships between layout elements and generate novel layouts in a given domain. Our framework allows us to generate a new layout either from an empty set or from an initial seed set of primitives, and can easily scale to support an arbitrary of primitives per layout. Furthermore, our analyses show that the model is able to automatically capture the semantic properties of the primitives. We propose simple improvements in both representation of layout primitives, as well as training methods to demonstrate competitive performance in very diverse data domains such as object bounding boxes in natural images (COCO bounding box), documents (PubLayoutNet), mobile applications (RICO dataset) as well as 3D shapes (Part-Net). Code and other materials will be made available at <https://kampta.github.io/layout>.

我们解决了图像、移动应用程序、文档和三维对象等不同领域的场景布局生成问题。大多数复杂的场景，无论是自然的还是人为设计的，都可以表示为简单的合成图形原语的有意义的排列。生成新布局或扩展现有布局需要了解这些基本体之间的关系。为此，我们提出了LayoutTransformer，这是一个新的框架，它利用自我关注来学习布局元素之间的上下文关系，并在给定域中生成新的布局。我们的框架允许我们从一个空集或一个初始的原语种子集生成一个新的布局，并且可以轻松地扩展以支持每个布局的任意原语。此外，我们的分析表明，该模型能够自动捕获原语的语义属性。我们建议对布局原语的表示和训练方法进行简单的改进，以展示在各种数据领域的竞争力，如自然图像中的对象边界框（COCO边界框）、文档（PubLayoutNet）、移动应用程序（RICO数据集）以及三维形状（零件网）。守则及其他资料将于<https://kampta.github.io/layout>。

This paper introduces a new method of data-driven microscope design for virtual fluorescence microscopy. We use a deep neural network (DNN) to effectively design optical patterns for specimen illumination that substantially improve upon the ability to infer fluorescence image information from unstained microscope images. To achieve this design, we include an illumination model within the DNN's first layers that is jointly optimized during network training. We validated our method on two different experimental setups, with different magnifications and sample types, to show a consistent improvement in performance as compared to conventional microscope imaging methods. Additionally, to understand the importance of learned illumination on the inference task, we varied the number of illumination patterns being optimized (and thus the number of unique images captured) and analyzed how the structure of the patterns changed as their number increased. This work demonstrates the power of programmable optical elements at enabling better machine learning algorithm performance and at providing physical insight into next generation of machine-controlled imaging systems.

介绍了一种用于虚拟荧光显微镜的数据驱动显微镜设计的新方法。我们使用深度神经网络 (DNN) 有效地设计样品照明的光学图案，大大提高了从未染色显微镜图像推断荧光图像信息的能力。为了实现此设计，我们在DNN的第一层中包含了一个照明模型，该模型在网络训练期间进行了联合优化。我们在两种不同的实验装置上验证了我们的方法，它们具有不同的放大率和样品类型，与传统的显微镜成像方法相比，性能得到了一致的改善。此外，为了理解学习照明对推理任务的重要性，我们改变了被优化的照明模式的数量（以及捕获的唯一图像的数量），并分析了模式的结构如何随着其数量的增加而变化。这项工作证明了可编程光学元件在实现更好的机器学习算法性能和提供对下一代机器控制成像系统的物理洞察力方面的威力。

Non-local self-similarity in natural images has been verified to be an effective prior for image restoration. However, most existing deep non-local methods assign a fixed number of neighbors for each query item, neglecting the dynamics of non-local correlations. Moreover, the non-local correlations are usually based on pixels, prone to be biased due to image degradation. To rectify these weaknesses, in this paper, we propose a dynamic attentive graph learning model (DAGL) to explore the dynamic non-local property on patch level for image restoration. Specifically, we propose an improved graph model to perform patch-wise graph convolution with a dynamic and adaptive number of neighbors for each node. In this way, image content can adaptively balance over-smooth and over-sharp artifacts through the number of its connected neighbors, and the patch-wise non-local correlations can enhance the message passing process. Experimental results on various image restoration tasks: synthetic image denoising, real image denoising, image demosaicing, and compression artifact reduction show that our DAGL can produce state-of-the-art results with superior accuracy and visual quality. The source code is available at <https://github.com/jianzhangcs/DAGL>.

自然图像中的非局部自相似性被证明是一种有效的图像恢复先验。然而，大多数现有的深度非局部方法为每个查询项分配固定数量的邻居，而忽略了非局部关联的动态性。此外，非局部相关性通常基于像素，容易因图像退化而产生偏差。为了纠正这些缺点，本文提出了一种动态注意图学习模型 (DAGL)，用于探索用于图像恢复的面片级动态非局部特性。具体地说，我们提出了一种改进的图模型来执行分片图卷积，每个节点具有动态和自适应数量的邻居。通过这种方式，图像内容可以通过其连接邻居的数量自适应地平衡平滑伪影和锐化伪影，并且分片非局部相关性可以增强消息传递过程。在各种图像恢复任务上的实验结果：合成图像去噪、真实图像去噪、图像去噪和压缩伪影减少表明，我们的DAGL能够以优异的精度和视觉质量产生最先进的结果。源代码可在<https://github.com/jianzhangcs/DAGL>。

Current 3D object detection paradigms highly rely on extensive annotation efforts, which makes them not practical in many real-world industrial applications. Inspired by that a human driver can keep accumulating experiences from self-exploring the roads without any tutor's guidance, we first step forwards to explore a simple yet effective self-supervised learning framework tailored for LiDAR-based 3D object detection. Although the self-supervised pipeline has achieved great success in 2D domain, the characteristic challenges (e.g., complex geometry structure and various 3D object views) encountered in the 3D domain hinder the direct adoption of existing techniques that often contrast the 2D augmented data or cluster single-view features. Here we present a novel self-supervised 3D Object detection framework that seamlessly integrates the geometry-aware contrast and clustering harmonization to lift the unsupervised 3D representation learning, named GCC-3D. First, GCC-3D introduces a Geometric-Aware Contrastive objective to learn spatial-sensitive local structure representation. This objective enforces the spatially-closed voxels to have high feature similarity. Second, a Pseudo-Instance Clustering harmonization mechanism is proposed to encourage that different views of pseudo-instances should have consistent similarities to clustering prototype centers. This module endows our model semantic discriminative capacity. Extensive experiments demonstrate our GCC-3D achieves significant performance improvement on data-efficient 3D object detection benchmarks (nuScenes and Waymo). Moreover, our GCC-3D framework can achieve state-of-the art performances on all popular 3D object detection benchmarks.

当前的三维目标检测模式高度依赖于大量的注释工作，这使得它们在许多实际工业应用中不实用。受人类驾驶员可以在没有任何导师指导下不断积累自我探索道路的经验的启发，我们首先探索了一个简单而有效的自我监督学习框架，该框架专为基于激光雷达的三维目标检测而设计。尽管自监督管道在2D领域取得了巨大成功，但在3D领域遇到的特性挑战（例如，复杂的几何结构和各种3D对象视图）阻碍了直接采用通常对比2D增强数据或群集单视图特征的现有技术。在这里，我们提出了一种新的自监督三维对象检测框架，该框架无缝集成了几何感知对比度和聚类协调，以提升无监督三维表示学习，称为GCC-3D。首先，GCC-3D引入了一个几何感知的对比目标来学习空间敏感的局部结构表示。该目标强制空间闭合体素具有较高的特征相似性。其次，提出了一种伪实例聚类协调机制，以鼓励不同的伪实例视图与聚类原型中心具有一致的相似性。该模块赋予我们的模型语义辨别能力。大量实验表明，我们的GCC-3D在数据高效的3D对象检测基准（nuScenes和Waymo）上实现了显著的性能改进。此外，我们的GCC-3D框架可以在所有流行的3D对象检测基准上实现最先进的性能。

Currently, the state-of-the-art methods treat few-shot semantic segmentation task as a conditional foreground-background segmentation problem, assuming each class is independent. In this paper, we introduce the concept of meta-class, which is the meta information (e.g. certain middle-level features) shareable among all classes. To explicitly learn meta-class representations in few-shot segmentation task, we propose a novel Meta-class Memory based few-shot segmentation method (MM-Net), where we introduce a set of learnable memory embeddings to memorize the meta-class information during the base class training and transfer to novel classes during the inference stage. Moreover, for the k-shot scenario, we propose a novel image quality measurement module to select images from the set of support images. A high-quality class prototype could be obtained with the weighted sum of support image features based on the quality measure. Experiments on both PASCAL-5<sup>i</sup> and COCO datasets show that our proposed method is able to achieve state-of-the-art results in both 1-shot and 5-shot settings. Particularly, our proposed MM-Net achieves 37.5% mIoU on the COCO dataset in 1-shot setting, which is 5.1% higher than the previous state-of-the-art.

目前，最先进的方法将少镜头语义分割任务视为一个条件前景背景分割问题，假设每个类是独立的。在本文中，我们引入了元类的概念，它是所有类之间可共享的元信息（例如，某些中层特征）。为了在少数镜头分割任务中明确学习元类表示，我们提出了一种新的基于元类记忆的少数镜头分割方法（MM Net），在基类训练过程中引入一组可学习的记忆嵌入来记忆元类信息，并在推理阶段转移到新类。此外，对于k-shot场景，我们提出了一种新的图像质量度量模块来从支持图像集中选择图像。基于质量度量的支持图像特征加权和可以得到高质量的类原型。在PASCAL-5^i和COCO数据集上的实验表明，我们提出的方法能够在单镜头和五镜头设置下实现最先进的结果。特别是，我们提出的MM网络在COCO数据集上实现了37.5%的单镜头mIoU，比以前的最先进水平高5.1%。

while deep learning succeeds in a wide range of tasks, it highly depends on the massive collection of annotated data which is expensive and time-consuming. To lower the cost of data annotation, active learning has been proposed to interactively query an oracle to annotate a small proportion of informative samples in an unlabeled dataset. Inspired by the fact that the samples with higher loss are usually more informative to the model than the samples with lower loss, in this paper we present a novel deep active learning approach that queries the oracle for data annotation when the unlabeled sample is believed to incorporate high loss. The core of our approach is a measurement Temporal Output Discrepancy (TOD) that estimates the sample loss by evaluating the discrepancy of outputs given by models at different optimization steps. Our theoretical investigation shows that TOD lower-bounds the accumulated sample loss thus it can be used to select informative unlabeled samples. On basis of TOD, we further develop an effective unlabeled data sampling strategy as well as an unsupervised learning criterion that enhances model performance by incorporating the unlabeled data. Due to the simplicity of TOD, our active learning approach is efficient, flexible, and task-agnostic. Extensive experimental results demonstrate that our approach achieves superior performances than the state-of-the-art active learning methods on image classification and semantic segmentation tasks.

虽然深度学习在广泛的任务中取得了成功，但它高度依赖于大量的注释数据收集，这既昂贵又耗时。为了降低数据标注的成本，提出了一种主动学习方法，即交互式地查询oracle以标注未标记数据集中的一小部分信息样本。受损失较高的样本通常比损失较低的样本对模型的信息量更大这一事实的启发，本文提出了一种新的深度主动学习方法，当未标记的样本被认为包含高损失时，该方法向oracle查询数据注释。我们方法的核心是测量时间输出差异（TOD），通过评估不同优化步骤下模型给出的输出差异来估计样本损失。我们的理论研究表明，TOD的下限是累积样本损失，因此它可以用来选择信息丰富的未标记样本。在TOD的基础上，我们进一步开发了一种有效的无标记数据采样策略以及一种无监督学习准则，该准则通过合并无标记数据来提高模型性能。由于TOD的简单性，我们的主动学习方法是高效、灵活和任务无关的。大量的实验结果表明，我们的方法在图像分类和语义分割任务上的性能优于目前最先进的主动学习方法。

Deep AUC Maximization (DAM) is a new paradigm for learning a deep neural network by maximizing the AUC score of the model on a dataset. Most previous works of AUC maximization focus on the perspective of optimization by designing efficient stochastic algorithms, and studies on generalization performance of large-scale DAM on difficult tasks are missing. In this work, we aim to make DAM more practical for interesting real-world applications (e.g., medical image classification). First, we propose a new margin-based min-max surrogate loss function for the AUC score (named as the AUC min-max-margin loss or simply AUC margin loss for short). It is more robust than the commonly used AUC square loss, while enjoying the same advantage in terms of large-scale stochastic optimization. Second, we conduct extensive empirical studies of our DAM method on four difficult medical image classification tasks, namely (i) classification of chest x-ray images for identifying many threatening diseases, (ii) classification of images of skin lesions for identifying melanoma, (iii) classification of mammogram for breast cancer screening, and (iv) classification of microscopic images for identifying tumor tissue. Our studies demonstrate that the proposed DAM method improves the performance of optimizing cross-entropy loss by a large margin, and also achieves better performance than optimizing the existing AUC square loss on these medical image classification tasks. Specifically, our DAM method has achieved the 1st place on Stanford CheXpert competition on Aug. 31, 2020. To the best of our knowledge, this is the first work that makes DAM succeed on large-scale medical image datasets. We also conduct extensive ablation studies to demonstrate the advantages of the new AUC margin loss over the AUC square loss on benchmark datasets. The proposed method is implemented in our open-sourced library LibAUC ([www.libauc.org](http://www.libauc.org)) whose github address is <https://github.com/Optimization-AI/LibAUC>.

深度AUC最大化 (DAM) 是一种通过最大化数据集上模型的AUC分数来学习深度神经网络的新范式。以往关于AUC最大化的研究大多集中在通过设计有效的随机算法进行优化的角度，而对于大型大坝在困难任务下的泛化性能的研究则较少。在这项工作中，我们的目标是使DAM在有趣的现实世界应用（例如，医学图像分类）中更加实用。首先，我们为AUC分数提出了一个新的基于保证金的最小-最大代理损失函数（简称AUC最小-最大保证金损失或简称AUC保证金损失）。它比常用的AUC平方损失更稳健，同时在大规模随机优化方面也具有相同的优势。其次，我们在四项困难的医学图像分类任务中对DAM方法进行了广泛的实证研究，即 (i) 胸部x射线图像分类，用于识别多种威胁性疾病，(ii) 皮肤病变图像分类，用于识别黑色素瘤，(iii) 乳腺癌筛查的乳房x光片分类，(iv) 显微图像分类，用于识别肿瘤组织。我们的研究表明，在这些医学图像分类任务中，所提出的DAM方法大大提高了交叉熵损失的优化性能，并且比优化现有AUC平方损失的性能更好。具体而言，我们的DAM方法已于2020年8月31日在斯坦福CheXpert竞赛中获得第一名。据我们所知，这是使DAM在大规模医学图像数据集上成功的第一项工作。我们还进行了广泛的消融研究，以证明在基准数据集上，新的AUC边缘损失优于AUC平方损失。该方法在我们的开源库LibAUC ([www.LibAUC.org](http://www.LibAUC.org)) 中实现，其github地址为<https://github.com/Optimization-AI/LibAUC>。

Many vision tasks use secondary information at inference time---a seed---to assist a computer vision model in solving a problem. For example, an initial bounding box is needed to initialize visual object tracking. To date, all such work makes the assumption that the seed is a good one. However, in practice, from crowdsourcing to noisy automated seeds, this is often not the case. We hence propose the problem of seed rejection---determining whether to reject a seed based on the expected performance degradation when it is provided in place of a gold-standard seed. We provide a formal definition to this problem, and focus on two meaningful subgoals: understanding causes of error and understanding the model's response to noisy seeds conditioned on the primary input. With these goals in mind, we propose a novel training method and evaluation metrics for the seed rejection problem. We then use seeded versions of the viewpoint estimation and fine-grained classification tasks to evaluate these contributions. In these experiments, we show our method can reduce the number of seeds that need to be reviewed for a target performance by over 23% compared to strong baselines.

许多视觉任务在推理时使用辅助信息——种子——来帮助计算机视觉模型解决问题。例如，初始化视觉对象跟踪需要初始边界框。迄今为止，所有这些工作都假定种子是好的。然而，在实践中，从众包到嘈杂的自动化种子，情况往往并非如此。因此，我们提出了种子拒绝问题——当提供种子代替金标准种子时，根据预期的性能下降来确定是否拒绝种子。我们为这个问题提供了一个正式的定义，并着重于两个有意义的子目标：理解错误的原因和理解模型对以主要输入为条件的噪声种子的响应。基于这些目标，我们针对种子拒绝问题提出了一种新的训练方法和评估指标。然后，我们使用视点估计和细粒度分类任务的种子版本来评估这些贡献。在这些实验中，我们表明，与强基线相比，我们的方法可以将需要检查目标性能的种子数量减少23%以上。

Dense video captioning aims to generate multiple associated captions with their temporal locations from the video. Previous methods follow a sophisticated "localize-then-describe" scheme, which heavily relies on numerous hand-crafted components. In this paper, we proposed a simple yet effective framework for end-to-end dense video captioning with parallel decoding (PDVC), by formulating the dense caption generation as a set prediction task. In practice, through stacking a newly proposed event counter on the top of a transformer decoder, the PDVC precisely segments the video into a number of event pieces under the holistic understanding of the video content, which effectively increases the coherence and readability of predicted captions. Compared with prior arts, the PDVC has several appealing advantages: (1) Without relying on heuristic non-maximum suppression or a recurrent event sequence selection network to remove redundancy, PDVC directly produces an event set with an appropriate size; (2) In contrast to adopting the two-stage scheme, we feed the enhanced representations of event queries into the localization head and caption head in parallel, making these two sub-tasks deeply interrelated and mutually promoted through the optimization; (3) Without bells and whistles, extensive experiments on ActivityNet Captions and YouCook2 show that PDVC is capable of producing high-quality captioning results, surpassing the state-of-the-art two-stage methods when its localization accuracy is on par with them. Code is available at <https://github.com/ttengwang/PDVC>.

密集视频字幕旨在从视频中生成多个与其时间位置相关联的字幕。以前的方法遵循复杂的“本地化然后描述”方案，严重依赖于大量手工制作的组件。在本文中，我们提出了一个简单而有效的端到端密集视频字幕并行解码（PDVC）框架，将密集字幕生成作为一个集合预测任务。实际上，通过在transformer解码器顶部堆叠新提出的事件计数器，PDVC在对视频内容的整体理解下，将视频精确地分割为多个事件片段，这有效地提高了预测字幕的连贯性和可读性。与现有技术相比，PDVC具有几个吸引人的优点：

(1) PDVC不依赖启发式非最大抑制或循环事件序列选择网络来消除冗余，直接产生具有适当大小的事件集；(2) 与采用两阶段方案相比，我们将事件查询的增强表示并行地馈送到本地化头部和标题头部，通过优化使这两个子任务相互关联并相互促进；(3) 在无铃声和口哨的情况下，ActuvyNET-字幕

和YouCu2的大量实验表明，PDVC能够产生高质量的字幕结果，超越了现有的两阶段方法，当定位精度与之相当时。代码可在<https://github.com/ttengwang/PDVC>.

Image enhancement is a subjective process whose targets vary with user preferences. In this paper, we propose a deep learning-based image enhancement method covering multiple tonal styles using only a single model dubbed StarEnhancer. It can transform an image from one tonal style to another, even if that style is unseen. With a simple one-time setting, users can customize the model to make the enhanced images more in line with their aesthetics. To make the method more practical, we propose a well-designed enhancer that can process a 4K-resolution image over 200 FPS but surpasses the contemporaneous single style image enhancement methods in terms of PSNR, SSIM, and LPIPS. Finally, our proposed enhancement method has good interactability, which allows the user to fine-tune the enhanced image using intuitive options.

图像增强是一个主观过程，其目标随用户偏好而变化。在本文中，我们提出了一种基于深度学习的图像增强方法，该方法仅使用一个称为StarEnhancer的模型就可以覆盖多个色调样式。它可以将图像从一种色调样式转换为另一种，即使该样式不可见。通过简单的一次性设置，用户可以自定义模型，使增强图像更符合其美学。为了使该方法更加实用，我们提出了一种设计良好的增强器，该增强器可以处理超过200fps的4K分辨率图像，但在PSNR、SSIM和LPIPS方面优于同期的单一风格图像增强方法。最后，我们提出的增强方法具有良好的交互性，允许用户使用直观的选项微调增强图像。

Recent studies show that convolutional neural networks (CNNs) are vulnerable under various settings, including adversarial attacks, common corruptions, and backdoor attacks. Motivated by the findings that human visual system pays more attention to global structure (e.g., shapes) for recognition while CNNs are biased towards local texture features in images, in this work we aim to analyze whether "edge features" could improve the recognition robustness in these scenarios, and if so, to what extent? To answer these questions and systematically evaluate the global structure features, we focus on shape features and propose two edge-enabled pipelines EdgeNetRob and Edge-GANRob, forcing the CNNs to rely more on edge features. Specifically, EdgeNetRob and EdgeGANRob first explicitly extract shape structure features from a given image via an edge detection algorithm. Then EdgeNetRob trains down-stream learning tasks directly on the extracted edge features, while EdgeGANRob reconstructs a new image by re-filling the texture information with a trained generative adversarial network (GANs). To reduce the sensitivity of edge detection algorithms to perturbations, we additionally propose a robust edge detection approach Robust Canny based on vanilla Canny. Based on our evaluation, we find that EdgeNetRob can help boost model robustness under different attack scenarios at the cost of the clean model accuracy. EdgeGANRob, on the other hand, is able to improve the clean model accuracy compared to EdgeNetRob while preserving robustness. This shows that given such edge features, how to leverage them matters for robustness, and it also depends on data properties. Our systematic studies on edge structure features under different settings will shed light on future robust feature exploration and optimization.

最近的研究表明，卷积神经网络（CNN）在各种环境下都很脆弱，包括对抗性攻击、常见的腐败和后门攻击。受人类视觉系统更加关注全局结构（如形状）进行识别，而CNN偏向于图像中的局部纹理特征这一发现的启发，在这项工作中，我们旨在分析“边缘特征”是否可以提高这些场景中的识别鲁棒性，如果是，在何种程度上？为了回答这些问题并系统地评估全局结构特征，我们将重点放在形状特征上，并提出了两个支持边缘的管道EdgeNetRob和edge GANRob，迫使CNN更多地依赖边缘特征。具体地说，EdgeNetRob和EdgeGANRob首先通过边缘检测算法从给定图像中明确提取形状结构特征。然后Edgenerob直接在提取的边缘特征上训练下游学习任务，而Edgenerob则通过使用经过训练的生成对抗网络（GANs）重新填充纹理信息来重建新图像。为了降低边缘检测算法对扰动的敏感性，我们在

vanilla-Canny的基础上提出了一种鲁棒边缘检测方法robust-Canny。基于我们的评估，我们发现EdgeNetRob可以帮助提高模型在不同攻击场景下的鲁棒性，但代价是干净的模型精度。另一方面，与EdgeNetRob相比，EdgeGANRob能够提高清洁模型的精度，同时保持鲁棒性。这表明，给定这些边缘特征，如何利用它们对健壮性很重要，而且还取决于数据属性。我们对不同背景下的边缘结构特征进行了系统的研究，这将为未来稳健的特征探索和优化提供参考。

In digital pathology, both detection and classification of cells are important for automatic diagnostic and prognostic tasks. Classifying cells into subtypes, such as tumor cells, lymphocytes or stromal cells is particularly challenging. Existing methods focus on morphological appearance of individual cells, whereas in practice pathologists often infer cell classes through their spatial context. In this paper, we propose a novel method for both detection and classification that explicitly incorporates spatial contextual information. We use the spatial statistical function to describe local density in both a multi-class and a multi-scale manner. Through representation learning and deep clustering techniques, we learn advanced cell representation with both appearance and spatial context. On various benchmarks, our method achieves better performance than state-of-the-arts, especially on the classification task.

在数字病理学中，细胞的检测和分类对于自动诊断和预后任务都很重要。将细胞分为亚型，如肿瘤细胞、淋巴细胞或基质细胞尤其具有挑战性。现有的方法侧重于单个细胞的形态学外观，而在实践中，病理学家通常通过其空间背景推断细胞类别。在本文中，我们提出了一种新的检测和分类方法，明确地结合了空间上下文信息。我们使用空间统计函数以多类和多尺度的方式描述局部密度。通过表征学习和深度聚类技术，我们学习了具有外观和空间背景的高级细胞表征。在各种基准测试中，我们的方法取得了比现有技术更好的性能，尤其是在分类任务上。

Videos with binaural audios provide an immersive viewing experience by enabling 3D sound sensation. Recent works attempt to generate binaural audio in a multimodal learning framework using large quantities of videos with accompanying binaural audio. In contrast, we attempt a more challenging problem -- synthesizing binaural audios for a video with monaural audio in a weakly supervised setting and weakly semi-supervised setting. Our key idea is that any down-stream task that can be solved only using binaural audios can be used to provide proxy supervision for binaural audio generation, thereby reducing the reliance on explicit supervision. In this work, as a proxy-task for weak supervision, we use Sound Source Localization with only audio. We design a two-stage architecture called Localize-to-Binauralize Network (L2BNet). The first stage of L2BNet is a Stereo Generation (SG) network employed to generate two-stream audio from monaural audio using visual frame information as guidance. In the second stage, an Audio Localization (AL) network is designed to use the synthesized two-stream audio to localize sound sources in visual frames. The entire network is trained end-to-end so that the AL network provides necessary supervision for the SG network. We experimentally show that our weakly-supervised framework generates two-stream audio containing binaural cues. Through user study, we further validate that our proposed approach generates binaural-quality audio using as little as 10% of explicit binaural supervision data for the SG network.

具有双耳音频的视频通过实现3D音感提供身临其境的观看体验。最近的工作试图在多模态学习框架中使用大量视频和伴随的双耳音频生成双耳音频。相比之下，我们尝试了一个更具挑战性的问题——在弱监督设置和弱半监督设置下为具有单耳音频的视频合成双耳音频。我们的关键思想是，任何只能使用双耳音频解决的下游任务都可以用于为双耳音频生成提供代理监控，从而减少对显式监控的依赖。在这项工作中，作为弱监督的代理任务，我们只使用音频进行声源定位。我们设计了一个称为本地化到二进制化网络（L2BNet）的两阶段体系结构。L2BNet的第一阶段是一个立体声生成（SG）网络，用于使用可视帧信息作为指导从单声道音频生成双流音频。在第二阶段，设计了一个音频定位（AL）网络，利用合成

的双流音频对视频帧中的声源进行定位。整个网络经过端到端的培训，以便AL网络为SG网络提供必要的监督。我们的实验表明，我们的弱监督框架生成包含双耳线索的双流音频。通过用户研究，我们进一步验证了我们提出的方法仅使用SG网络10%的显式双耳监控数据生成双耳质量音频。

Snow is a highly complicated atmospheric phenomenon that usually contains snowflake, snow streak, and veiling effect (similar to the haze or the mist). In this literature, we propose a single image desnowing algorithm to address the diversity of snow particles in shape and size. First, to better represent the complex snow shape, we apply the dual-tree wavelet transform and propose a complex wavelet loss in the network. Second, we propose a hierarchical decomposition paradigm in our network for better understanding the different sizes of snow particles. Last, we propose a novel feature called the contradict channel (CC) for the snow scenes. we find that the regions containing the snow particles tend to have higher intensity in the CC than that in the snow-free regions. We leverage this discriminative feature to construct the contradict channel loss for improving the performance of snow removal. Moreover, due to the limitation of existing snow datasets, to simulate the snow scenarios comprehensively, we propose a large-scale dataset called Comprehensive Snow Dataset (CSD). Experimental results show that the proposed method can favorably outperform existing methods in three synthetic datasets and real-world datasets. The code and dataset are released in <https://github.com/weitingchen83/ICCV2021-Single-Image-Desnowing-HDCWNet>.

雪是一种高度复杂的大气现象，通常包含雪花、雪纹和遮掩效应（类似于薄雾或薄雾）。在这篇文献中，我们提出了一个单一的图像去噪算法来解决雪粒子在形状和大小上的多样性。首先，为了更好地表示复杂的雪形，我们应用了双树小波变换，提出了一种网络中的复小波损耗。其次，为了更好地理解雪粒子的不同大小，我们在我们的网络中提出了一种分层分解范式。最后，我们提出了一个新的特征，称为矛盾通道（CC）的雪景。我们发现，含有雪粒子的区域在CC中的强度往往高于无雪区域。我们利用这一鉴别特征来构造矛盾信道损耗，以提高除雪性能。此外，由于现有雪数据集的局限性，为了全面模拟雪场景，我们提出了一个大规模数据集，称为综合雪数据集（CSD）。实验结果表明，在三个合成数据集和真实数据集上，该方法均优于现有方法。代码和数据集在中发布<https://github.com/weitingchen83/ICCV2021-Single-Image-Desnowing-HDCWNet>。

This paper tackles the problem of learning a finer representation than the one provided by training labels. This enables fine-grained category retrieval of images in a collection annotated with coarse labels only. Our network is learned with a nearest-neighbor classifier objective, and an instance loss inspired by self-supervised learning. By jointly leveraging the coarse labels and the underlying fine-grained latent space, it significantly improves the accuracy of category-level retrieval methods. Our strategy outperforms all competing methods for retrieving or classifying images at a finer granularity than that available at train time. It also improves the accuracy for transfer learning tasks to fine-grained datasets.

本文解决的问题是学习一个比训练标签提供的更好的表示。这使得仅使用粗标签注释的集合中的图像能够进行细粒度的类别检索。我们的网络是以最近邻分类器为目标进行学习的，实例丢失是由自监督学习启发的。通过联合利用粗标签和底层细粒度潜在空间，它显著提高了类别级检索方法的准确性。在检索或分类图像时，我们的策略优于所有竞争方法，其粒度比训练时可用的粒度更细。它还提高了将学习任务转移到细粒度数据集的准确性。

In various imaging problems, we only have access to compressed measurements of the underlying signals, hindering most learning-based strategies which usually require pairs of signals and associated measurements for training. Learning only from compressed measurements is impossible in general, as the compressed observations do not contain information outside the range of the forward sensing operator. We propose a new end-to-end self-supervised framework that overcomes this limitation by exploiting the equivariances present in natural signals. Our proposed learning strategy performs as well as fully supervised methods. Experiments demonstrate the potential of this framework on inverse problems including sparse-view X-ray computed tomography on real clinical data and image inpainting on natural images. Code has been made available at: <https://github.com/edongdongchen/EI>.

在各种成像问题中，我们只能获得基本信号的压缩测量值，这妨碍了大多数基于学习的策略，这些策略通常需要成对的信号和相关的测量值进行训练。一般来说，仅从压缩测量中学习是不可能的，因为压缩观测不包含前向传感操作员范围之外的信息。我们提出了一种新的端到端自监督框架，通过利用自然信号中存在的等价性克服了这一局限性。我们提出的学习策略执行以及完全监督的方法。实验证明了该框架在反问题上的潜力，包括对真实临床数据的稀疏视图X射线计算机断层扫描和对自然图像的图像修复。代码已在以下网址提供：<https://github.com/edongdongchen/EI>。

While CNNs achieved remarkable progress in shadow detection, they tend to make mistakes in dark non-shadow regions and relatively bright shadow regions. They are also susceptible to brightness change. These two phenomena reveal that deep shadow detectors heavily depend on the intensity cue, which we refer to as intensity bias. In this paper, we propose a novel feature decomposition and reweighting scheme to mitigate this intensity bias, in which multi-level integrated features are decomposed into intensity-variant and intensity-invariant components through self-supervision. By reweighting these two types of features, our method can reallocate the attention to the corresponding latent semantics and achieves balanced exploitation of them. Extensive experiments on three popular datasets show that the proposed method outperforms state-of-the-art shadow detectors.

尽管CNN在阴影检测方面取得了显著的进步，但它们往往会在黑暗的非阴影区域和相对明亮的阴影区域出错。它们也容易受到亮度变化的影响。这两种现象表明，深阴影探测器严重依赖于强度线索，我们称之为强度偏差。在本文中，我们提出了一种新的特征分解和重新加权方案来缓解这种强度偏差，该方案通过自我监督将多级集成特征分解为强度变化和强度不变的分量。通过对这两类特征重新加权，我们的方法可以将注意力重新分配到相应的潜在语义上，并实现对它们的均衡利用。在三个流行数据集上的大量实验表明，该方法优于最先进的阴影检测器。

This paper studies the problem of novel category discovery on single- and multi-modal data with labels from different but relevant categories. We present a generic, end-to-end framework to jointly learn a reliable representation and assign clusters to unlabelled data. To avoid over-fitting the learnt embedding to labelled data, we take inspiration from self-supervised representation learning by noise-contrastive estimation and extend it to jointly handle labelled and unlabelled data. In particular, we propose using category discrimination on labelled data and cross-modal discrimination on multi-modal data to augment instance discrimination used in conventional contrastive learning approaches. We further employ Winner-Take-All (WTA) hashing algorithm on the shared representation space to generate pairwise pseudo labels for unlabelled data to better predict cluster assignments. We thoroughly evaluate our framework on large-scale multi-modal video benchmarks Kinetics-400 and VGG-Sound, and image benchmarks CIFAR10, CIFAR100 and ImageNet, obtaining state-of-the-art results.

本文研究了具有不同但相关类别标签的单模态和多模态数据的新类别发现问题。我们提出了一个通用的端到端框架，以共同学习可靠的表示，并将集群分配给未标记的数据。为了避免学习到的嵌入过度拟合标签数据，我们从噪声对比估计的自监督表示学习中得到启发，并将其扩展到联合处理标签和未标签数据。特别是，我们建议在标记数据上使用类别歧视，在多模态数据上使用跨模态歧视，以增强传统对比如学习方法中使用的实例歧视。我们进一步在共享表示空间上使用赢家通吃（WTA）散列算法为未标记数据生成成对伪标签，以更好地预测集群分配。我们在大规模多模式视频基准Kinetics-400和VGG Sound以及图像基准CIFAR10、CIFAR100和ImageNet上全面评估了我们的框架，获得了最先进的结果。

In this paper, we propose a novel two-stage context-aware network named CANet for shadow removal, in which the contextual information from non-shadow regions is transferred to shadow regions at the embedded feature spaces. At Stage-I, we propose a contextual patch matching module to generate a set of potential matching pairs of shadow and non-shadow patches. Combined with the potential contextual relationships between shadow and non-shadow regions, our well-designed contextual feature transfer (CFT) mechanism can transfer contextual information from non-shadow to shadow regions at different scales. With the reconstructed feature maps, we remove shadows at L and A/B channels separately. At Stage-II, we use an encoder-decoder to refine current results and generate the final shadow removal results. We evaluate our proposed CANet on two benchmark datasets and some real-world shadow images with complex scenes. Extensive experiment results strongly demonstrate the efficacy of our proposed CANet and exhibit superior performance to state-of-the-arts.

在本文中，我们提出了一种新的用于阴影去除的两阶段上下文感知网络CANet，该网络将来自非阴影区域的上下文信息转移到嵌入特征空间中的阴影区域。在第一阶段，我们提出了一个上下文补丁匹配模块来生成一组阴影和非阴影补丁的潜在匹配对。结合阴影和非阴影区域之间潜在的上下文关系，我们精心设计的上下文特征转移（CFT）机制可以在不同的尺度上将上下文信息从非阴影区域转移到阴影区域。利用重构后的特征图，我们分别去除L和A/B通道上的阴影。在第二阶段，我们使用编码器-解码器优化当前结果，并生成最终阴影消除结果。我们在两个基准数据集和一些具有复杂场景的真实阴影图像上评估了我们提出的CANet。大量的实验结果有力地证明了我们提出的CANet的有效性，并显示出优于现有技术的性能。

Data augmentation is a widely adopted technique for avoiding overfitting when training deep neural networks. However, this approach requires domain-specific knowledge and is often limited to a fixed set of hard-coded transformations. Recently, several works proposed to use generative models for generating semantically meaningful perturbations to train a classifier. However, because accurate encoding and decoding is critical, these methods, which use architectures that approximate the latent-variable inference, remained limited to pilot studies on small datasets. Exploiting the exactly reversible encoder-decoder structure of normalizing flows, we perform on-manifold perturbations in the latent space to define fully unsupervised data augmentations. We demonstrate that such perturbations match the performance of advanced data augmentation techniques---reaching 96.6% test accuracy for CIFAR-10 using ResNet-18 and outperform existing methods, particularly in low data regimes---yielding 10--25% relative improvement of test accuracy from classical training. We find that our latent adversarial perturbations adaptive to the classifier throughout its training are most effective, yielding the first test accuracy improvement results on real-world datasets---CIFAR-10/100---via latent-space perturbations.

在训练深度神经网络时，数据增强是一种广泛采用的避免过度拟合的技术。然而，这种方法需要特定领域的知识，并且通常仅限于一组固定的硬编码转换。最近，有几项工作提出使用生成模型生成语义上有意义的扰动来训练分类器。然而，由于准确的编码和解码至关重要，这些使用近似潜在变量推断的架构的方法仍然局限于小型数据集的初步研究。利用规范化流的完全可逆编码器-解码器结构，我们在潜在空

间中执行流形扰动，以定义完全无监督的数据增强。我们证明，这种扰动与先进的数据增强技术的性能相匹配——使用ResNet-18，CIFAR-10的测试精度达到96.6%，并且优于现有方法，尤其是在低数据区域——从经典训练中获得10-25%的测试精度相对提高。我们发现，在分类器的整个训练过程中，我们对其进行自适应的潜在对抗性扰动是最有效的，通过潜在空间扰动，在真实数据集（CIFAR-10/100）上产生了第一个测试精度改进结果。

Given only a few glimpses of an environment, how much can we infer about its entire floorplan? Existing methods can map only what is visible or immediately apparent from context, and thus require substantial movements through a space to fully map it. We explore how both audio and visual sensing together can provide rapid floorplan reconstruction from limited viewpoints. Audio not only helps sense geometry outside the camera's field of view, but it also reveals the existence of distant freespace (e.g., a dog barking in another room) and suggests the presence of rooms not visible to the camera (e.g., a dishwasher humming in what must be the kitchen to the left). We introduce AV-Map, a novel multi-modal encoder-decoder framework that reasons jointly about audio and vision to reconstruct a floorplan from a short input video sequence. We train our model to predict both the interior structure of the environment and the associated rooms' semantic labels. Our results on 85 large real-world environments show the impact: with just a few glimpses spanning 26% of an area, we can estimate the whole area with 66% accuracy---substantially better than the state of the art approach for extrapolating visual maps.

如果只看几眼环境，我们能推断出它的整个平面图是多少？现有方法只能映射上下文中可见或立即可见的内容，因此需要在空间中进行大量移动才能完全映射。我们将探讨如何将音频和视觉感知结合起来，从有限的视角提供快速的平面图重建。音频不仅有助于感知摄像机视野外的几何图形，还可以揭示远处自由空间的存在（例如，狗在另一个房间里吠叫），并提示摄像机看不到的房间（例如，洗碗机在厨房的左边嗡嗡响）。我们介绍了AV Map，一种新颖的多模式编解码框架，它将音频和视觉结合起来，从一个短的输入视频序列中重建一个平面图。我们训练我们的模型来预测环境的内部结构和相关房间的语义标签。我们在85个大型真实环境中的研究结果表明了这一影响：只需对26%的区域进行几次瞥见，我们就可以以66%的准确率估算整个区域——大大优于最先进的外推视觉地图的方法。

One-stage object detection is commonly implemented by optimizing two sub-tasks: object classification and localization, using heads with two parallel branches, which might lead to a certain level of spatial misalignment in predictions between the two tasks. In this work, we propose a Task-aligned One-stage Object Detection (TOOD) that explicitly aligns the two tasks in a learning-based manner. First, we design a novel Task-aligned Head (T-Head) which offers a better balance between learning task-interactive and task-specific features, as well as a greater flexibility to learn the alignment via a task-aligned predictor. Second, we propose Task Alignment Learning (TAL) to explicitly pull closer (or even unify) the optimal anchors for the two tasks during training via a designed sample assignment scheme and a task-aligned loss. Extensive experiments are conducted on MS-COCO, where TOOD achieves a 51.1 AP at single-model single-scale testing. This surpasses the recent one-stage detectors by a large margin, such as ATSS (47.7 AP), GFL (48.2 AP), and PAA (49.0 AP), with fewer parameters and FLOPs. Qualitative results also demonstrate the effectiveness of TOOD for better aligning the tasks of object classification and localization. Code is available at <https://github.com/fcjian/TOOD>.

单阶段目标检测通常通过优化两个子任务来实现：目标分类和定位，使用具有两个平行分支的头部，这可能导致两个任务之间的预测出现一定程度的空间错位。在这项工作中，我们提出了一种任务对齐的一阶段对象检测（TOOD），它以基于学习的方式显式地对齐两个任务。首先，我们设计了一种新的任务对齐头（T-Head），它在学习任务交互和任务特定功能之间提供了更好的平衡，并且通过任务对齐预测器提供了更大的灵活性来学习对齐。其次，我们提出任务对齐学习（TAL），通过设计的样本分配方案

和任务对齐损失，明确拉近（甚至统一）训练期间两个任务的最佳锚。在MS-COCO上进行了大量实验，其中TOOD在单模型单尺度测试中达到51.1 AP。这大大超过了最近的单级探测器，如 ATSS (47.7AP)、GFL (48.2AP) 和PAA (49.0AP)，参数和触发器更少。定性结果还证明了TOOD在更好地协调目标分类和定位任务方面的有效性。代码可在<https://github.com/fcjian/TOOD>.

In this paper, we propose a generalizable mixed-precision quantization (GMPQ) method for efficient inference. Conventional methods require the consistency of datasets for bitwidth search and model deployment to guarantee the policy optimality, leading to heavy search cost on challenging largescale datasets in realistic applications. On the contrary, our GMPQ searches the mixed-quantization policy that can be generalized to largescale datasets with only a small amount of data, so that the search cost is significantly reduced without performance degradation. Specifically, we observe that locating network attribution correctly is general ability for accurate visual analysis across different data distribution. Therefore, despite of pursuing higher model accuracy and complexity, we preserve attribution rank consistency between the quantized models and their full-precision counterparts via efficient capacity-aware attribution imitation for generalizable mixed-precision quantization strategy search. Extensive experiments show that our method obtains competitive accuracy-complexity trade-off compared with the state-of-the-art mixed-precision networks in significantly reduced search cost. The code is available at <https://github.com/ZiweiWangTHU/GMPQ.git>.

在本文中，我们提出了一种通用的混合精度量化 (GMPQ) 方法来进行有效的推理。传统的方法要求比特宽度搜索和模型部署的数据集的一致性，以保证策略的最优性，从而导致在实际应用中对具有挑战性的大规模数据集产生沉重的搜索成本。相反，我们的GMPQ搜索混合量化策略，该策略可以推广到只有少量数据的大规模数据集，因此在不降低性能的情况下显著降低了搜索成本。具体而言，我们观察到，正确定位网络属性是跨不同数据分布进行准确视觉分析的一般能力。因此，尽管我们追求更高的模型精度和复杂度，我们仍然通过有效的容量感知属性模拟来保持量化模型与全精度模型之间的属性等级一致性，以实现广义混合精度量化策略搜索。大量实验表明，与现有的混合精度网络相比，我们的方法在显著降低搜索成本的情况下获得了具有竞争力的精度复杂度权衡。该守则可于<https://github.com/ZiweiWangTHU/GMPQ.git>.

Recent studies have demonstrated the vulnerability of deep neural networks against adversarial examples. Inspired by the observation that adversarial examples often lie outside the natural image data manifold and the intrinsic dimension of image data is much smaller than its pixel space dimension, we propose to embed high-dimensional input images into a low-dimensional space and apply regularization on the embedding space to push the adversarial examples back to the manifold. The proposed framework is called Embedding Regularized Classifier (ER-Classifier), which improves the adversarial robustness of the classifier through embedding regularization. Besides improving classification accuracy against adversarial examples, the framework can be combined with detection methods to detect adversarial examples. Experimental results on several benchmark datasets show that, our proposed framework achieves good performance against strong adversarial attack methods.

最近的研究证明了深层神经网络在对抗性示例中的脆弱性。受到以下观察结果的启发：敌对示例通常位于自然图像数据流形之外，图像数据的固有维度远小于其像素空间维度，我们建议将高维输入图像嵌入到低维空间中，并在嵌入空间上应用正则化将对抗性示例推回到流形中。该框架称为嵌入正则化分类器 (ER分类器)，通过嵌入正则化提高了分类器的对抗性鲁棒性。除了提高对抗性示例的分类精度外，该框架还可以与检测方法相结合来检测对抗性示例。在多个基准数据集上的实验结果表明，我们提出的框架对强对抗攻击方法具有良好的性能。

We propose IntraTomo, a powerful framework that combines the benefits of learning-based and model-based approaches for solving highly ill-posed inverse problems in the Computed Tomography (CT) context. IntraTomo is composed of two core modules: a novel sinogram prediction module, and a geometry refinement module, which are applied iteratively. In the first module, the unknown density field is represented as a continuous and differentiable function, parameterized by a deep neural network. This network is learned, in a self-supervised fashion, from the incomplete or/and degraded input sinogram. After getting estimated through the sinogram prediction module, the density field is consistently refined in the second module using local and non-local geometrical priors. With these two core modules, we show that IntraTomo significantly outperforms existing approaches on several ill-posed inverse problems, such as limited angle tomography with a range of 45 degrees, sparse view tomographic reconstruction with as few as eight views, or super-resolution tomography with eight times increased resolution. The experiments on simulated and real data show that our approach can achieve results of unprecedented quality.

我们提出了IntraTomo，这是一个强大的框架，它结合了基于学习和基于模型的方法的优点，用于解决计算机断层扫描（CT）环境中的高度不适定反问题。IntraTomo由两个核心模块组成：一个新的正弦图预测模块和一个几何精化模块，这两个模块是迭代应用的。在第一个模块中，未知密度场被表示为一个连续的可微函数，由一个深度神经网络参数化。该网络以自我监督的方式从不完整或/或退化的输入正弦图中学习。在通过正弦图预测模块得到估计值后，第二个模块使用局部和非局部几何先验对密度场进行一致的细化。有了这两个核心模块，我们发现IntraTomo在几个不适定逆问题上的性能明显优于现有方法，例如45度范围的有限角度层析成像、只有8个视图的稀疏视图层析成像重建，或分辨率提高8倍的超分辨率层析成像。仿真和实际数据的实验表明，该方法可以获得前所未有的结果。

We present a method for differentiable rendering of 3D surfaces that supports both explicit and implicit representations, provides derivatives at occlusion boundaries, and is fast and simple to implement. The method first samples the surface using non-differentiable rasterization, then applies differentiable, depth-aware point splatting to produce the final image. Our approach requires no differentiable meshing or rasterization steps, making it efficient for large 3D models and applicable to isosurfaces extracted from implicit surface definitions. We demonstrate the effectiveness of our method for implicit-, mesh-, and parametric-surface-based inverse rendering and neural-network training applications. In particular, we show for the first time efficient, differentiable rendering of an isosurface extracted from a neural radiance field (NeRF), and demonstrate surface-based, rather than volume-based, rendering of a NeRF.

我们提出了一种三维曲面的可微绘制方法，该方法支持显式和隐式表示，在遮挡边界处提供导数，并且实现快速简单。该方法首先使用不可微光栅化对曲面进行采样，然后应用可微、深度感知的点飞溅生成最终图像。我们的方法不需要可微的网格划分或光栅化步骤，这使得它对于大型3D模型非常有效，并且适用于从隐式曲面定义中提取的等值面。我们证明了我们的方法对于隐式、网格和参数化基于曲面的逆绘制和神经网络训练应用的有效性。特别是，我们首次展示了从神经辐射场（NeRF）提取的等值面的高效可微绘制，并展示了基于表面而非基于体积的NeRF绘制。

Image inpainting methods have shown significant improvements by using deep neural networks recently. However, many of these techniques often create distorted structures or blurry inconsistent textures. The problem is rooted in the encoder layers' ineffectiveness in building a complete and faithful embedding of the missing regions from scratch. Existing solutions like coarse-to-fine, progressive refinement, structural guidance, etc., suffer from huge computational overheads owing to multiple generator networks, limited ability of handcrafted features, and sub-optimal utilization of the information present in the ground truth. We propose a distillation-based approach for inpainting, where we provide direct feature-level supervision while training. We deploy cross and self-distillation techniques and design a dedicated completion-block in encoder to produce more accurate encoding of the holes. Next, we demonstrate how an inpainting network's attention module can improve by leveraging a distillation-based attention transfer technique and enhancing coherence by using a pixel-adaptive global-local feature fusion. We conduct extensive evaluations on multiple datasets to validate our method. Along with achieving significant improvements over previous SOTA methods, the proposed approach's effectiveness is also demonstrated through its ability to improve existing inpainting works.

近年来，利用深度神经网络对图像修复方法进行了显著改进。然而，这些技术中的许多常常会产生扭曲的结构或模糊的不一致纹理。问题的根源在于编码器层在从头开始构建完整且忠实的缺失区域嵌入方面的无效性。现有的解决方案，如课程细化、渐进细化、结构指导等，由于多个发电机网络、手工特征的能力有限以及对地面真相中存在的信息的次优利用，遭受着巨大的计算开销。我们提出了一种基于蒸馏的修复方法，在训练时提供直接的特征级监控。我们采用交叉和自蒸馏技术，并在编码器中设计一个专用的完成块，以产生更精确的孔编码。接下来，我们将演示修复网络的注意模块如何通过利用基于蒸馏的注意转移技术和使用像素自适应全局局部特征融合增强一致性来改进。我们对多个数据集进行了广泛的评估，以验证我们的方法。与以前的SOTA方法相比，该方法不仅取得了显著的改进，还通过改进现有修复工程的能力证明了该方法的有效性。

Although instance segmentation has made considerable advancement over recent years, it's still a challenge to design high accuracy algorithms with real-time performance. In this paper, we propose a real-time instance segmentation framework termed OrienMask. Upon the one-stage object detector YOLOv3, a mask head is added to predict some discriminative orientation maps, which are explicitly defined as spatial offset vectors for both foreground and background pixels. Thanks to the discrimination ability of orientation maps, masks can be recovered without the need for extra foreground segmentation. All instances that match with the same anchor size share a common orientation map. This special sharing strategy reduces the amortized memory utilization for mask predictions but without loss of mask granularity. Given the surviving box predictions after NMS, instance masks can be concurrently constructed from the corresponding orientation maps with low complexity. Owing to the concise design for mask representation and its effective integration with the anchor-based object detector, our method is qualified under real-time conditions while maintaining competitive accuracy. Experiments on COCO benchmark show that OrienMask achieves 34.8 mask AP at the speed of 42.7 fps evaluated with a single RTX 2080 Ti. Code is available at [github.com/duwt/OrienMask](https://github.com/duwt/OrienMask).

尽管近年来实例分割取得了长足的进步，但设计具有实时性能的高精度算法仍然是一个挑战。在本文中，我们提出了一个实时实例分割框架，称为OrienMask。在一級目标检测器YOLOv3上，添加一个遮罩头来预测一些辨别性方向图，这些方向图被明确定义为前景和背景像素的空间偏移向量。由于方向图的辨别能力，可以在不需要额外前景分割的情况下恢复遮罩。与相同锚定大小匹配的所有实例共享一个公共方向贴图。这种特殊的共享策略降低了掩码预测的摊销内存利用率，但不会损失掩码粒度。给定NMS后的生存框预测，实例掩码可以从相应的方向图以低复杂度同时构造。由于掩模表示的简洁设计及其与基于锚的目标检测器的有效集成，我们的方法在保持竞争精度的同时，在实时条件下是合格的。在

COCO基准上的实验表明，OrienMask以42.7fps的速度实现了34.8mask AP，使用单个RTX 2080Ti进行评估。代码可以在github上找到。[com/duwt/OrienMask](https://github.com/duwt/OrienMask)。

Image segmentation is often ambiguous at the level of individual image patches and requires contextual information to reach label consensus. In this paper we introduce Segmenter, a transformer model for semantic segmentation. In contrast to convolution-based methods, our approach allows to model global context already at the first layer and throughout the network. We build on the recent Vision Transformer (ViT) and extend it to semantic segmentation. To do so, we rely on the output embeddings corresponding to image patches and obtain class labels from these embeddings with a point-wise linear decoder or a mask transformer decoder. We leverage models pre-trained for image classification and show that we can fine-tune them on moderate sized datasets available for semantic segmentation. The linear decoder allows to obtain excellent results already, but the performance can be further improved by a mask transformer generating class masks. We conduct an extensive ablation study to show the impact of the different parameters, in particular the performance is better for large models and small patch sizes. Segmenter attains excellent results for semantic segmentation. It outperforms the state of the art on both ADE20K and Pascal Context datasets and is competitive on Cityscapes.

图像分割在单个图像块的层次上通常是模糊的，需要上下文信息才能达成一致。本文介绍了切分器，一种用于语义切分的变换器模型。与基于卷积的方法相比，我们的方法允许在第一层和整个网络中对全局上下文进行建模。我们建立在最近的视觉转换器（ViT）的基础上，并将其扩展到语义分割。为此，我们依赖于与图像块对应的输出嵌入，并使用逐点线性解码器或掩码转换器解码器从这些嵌入中获取类标签。我们利用预先训练的图像分类模型，并表明我们可以在中等大小的数据集上对其进行微调，以进行语义分割。线性解码器已经允许获得优异的结果，但是通过生成类掩码的掩码转换器可以进一步提高性能。我们进行了广泛的烧蚀研究，以显示不同参数的影响，特别是对于大型模型和小面积贴片，性能更好。Segmenter在语义分割方面取得了很好的效果。它在ADE20K和Pascal上下文数据集上都优于最先进的技术，并且在城市景观上具有竞争力。

Differentiable Architecture Search (DARTS) improves the efficiency of architecture search by learning the architecture and network parameters end-to-end. However, the intrinsic relationship between the architecture's parameters is neglected, leading to a sub-optimal optimization process. The reason lies in the fact that the gradient descent method used in DARTS ignores the coupling relationship of the parameters and therefore degrades the optimization. In this paper, we address this issue by formulating DARTS as a bilinear optimization problem and introducing an Interactive Differentiable Architecture Search (IDARTS). We first develop a backtracking backpropagation process, which can decouple the relationships of different kinds of parameters and train them in the same framework. The backtracking method coordinates the training of different parameters that fully explore their interaction and optimize training. We present experiments on the CIFAR10 and ImageNet datasets that demonstrate the efficacy of the IDARTS approach by achieving a top-1 accuracy of 76.52% on ImageNet without additional search cost vs. 75.8% with the state-of-the-art PC-DARTS.

差分体系结构搜索（DARTS）通过端到端学习体系结构和网络参数，提高了体系结构搜索的效率。然而，架构参数之间的内在关系被忽略，导致次优优化过程。这是因为在省道中使用的梯度下降法忽略了参数之间的耦合关系，从而降低了优化效果。在本文中，我们通过将DARTS描述为一个双线性优化问题并引入交互式可微结构搜索（IDARTS）来解决这个问题。我们首先开发了一个回溯反向传播过程，该过程可以解耦不同类型参数之间的关系，并在同一框架中对它们进行训练。回溯法协调不同参数的训练，充分挖掘它们之间的相互作用，优化训练。我们在CIFAR10和ImageNet数据集上进行了实验，证明了

IDARTS方法的有效性，在ImageNet上实现了76.52%的顶级精度，而无需额外的搜索成本，而在最先进的PC-DART上实现了75.8%。

Current neural architecture search (NAS) algorithms still require expert knowledge and effort to design a search space for network construction. In this paper, we consider automating the search space design to minimize human interference, which however faces two challenges: the explosive complexity of the exploration space and the expensive computation cost to evaluate the quality of different search spaces. To solve them, we propose a novel differentiable evolutionary framework named AutoSpace, which evolves the search space to an optimal one with following novel techniques: a differentiable fitness scoring function to efficiently evaluate the performance of cells and a reference architecture to speedup the evolution procedure and avoid falling into sub-optimal solutions. The framework is generic and compatible with additional computational constraints, making it feasible to learn specialized search spaces that fit different computational budgets. With the learned search space, the performance of recent NAS algorithms can be improved significantly compared with using manually designed spaces. Remarkably, the models generated from the new search space achieve 77.8% top-1 accuracy on ImageNet under the mobile setting (MAdds $\leq$ 500M), outperforming previous SOTA EfficientNet-B0 by 0.7%.

<https://github.com/zhoudaquan/AutoSpace.git>

当前的神经架构搜索 (NAS) 算法仍然需要专家知识和努力来设计用于网络构建的搜索空间。在本文中，我们考虑自动搜索空间设计，以尽量减少人为干扰，然而，面临两个挑战：爆炸复杂性的探索空间和昂贵的计算成本，以评估不同的搜索空间的质量。为了解决这些问题，我们提出了一种新的可微进化框架AutoSpace，它通过以下新技术将搜索空间演化为最优搜索空间：一个可微适应度评分函数，用于有效评估单元的性能；一个参考体系结构，用于加速演化过程并避免陷入次优解。该框架是通用的，并且与其他计算约束兼容，因此可以学习适合不同计算预算的专门搜索空间。通过学习搜索空间，与使用手动设计的空间相比，最近的NAS算法的性能可以显著提高。值得注意的是，新搜索空间生成的模型在移动设置 (MAdds $\leq$ 500M) 下的ImageNet上达到77.8%的top-1精度，比之前的SOTA效率网-B0高出0.7%。<https://github.com/zhoudaquan/AutoSpace.git>

Automation of neural architecture design has been a coveted alternative to human experts. Various search methods have been proposed aiming to find the optimal architecture in the search space. One would expect the search results to improve when the search space grows larger since it would potentially contain more performant candidates. Surprisingly, we observe that enlarging search space is unbeneficial or even detrimental to existing NAS methods such as DARTS, ProxylessNAS, and SPOS. This counterintuitive phenomenon suggests that enabling existing methods to large search space regimes is non-trivial. However, this problem is less discussed in the literature. We present a Neural Search-space Evolution (NSE) scheme, the first neural architecture search scheme designed especially for large space neural architecture search problems. The necessity of a well-designed search space with constrained size is a tacit consent in existing methods, and our NSE aims at minimizing such necessity. Specifically, the NSE starts with a search space subset, then evolves the search space by repeating two steps: 1) search an optimized space from the search space subset, 2) refill this subset from a large pool of operations that are not traversed. We further extend the flexibility of obtainable architectures by introducing a learnable multi-branch setting. With the proposed method, we achieve 77.3% top-1 retrain accuracy on ImageNet with 333M FLOPs, which yielded a state-of-the-art performance among previous auto-generated architectures that do not involve knowledge distillation or weight pruning. When the latency constraint is adopted, our result also performs better than the previous best-performing mobile models with a 77.9% Top-1 retrain accuracy. Code is available at [https://github.com/orashi/NSE\\_NAS](https://github.com/orashi/NSE_NAS).

神经结构设计的自动化已经成为人类专家梦寐以求的替代方案。为了在搜索空间中找到最优结构，人们提出了各种搜索方法。当搜索空间变大时，人们会期望搜索结果会有所改善，因为它可能包含更多性能更好的候选者。令人惊讶的是，我们发现扩大搜索空间对现有的NAS方法（如DART、ProxylessNAS和SPO）不利甚至有害。这种违反直觉的现象表明，使现有方法能够在较大的搜索空间范围内进行搜索并非易事。然而，文献中很少讨论这个问题。我们提出了一个神经搜索空间进化（NSE）方案，这是第一个专门针对大空间神经结构搜索问题设计的神经结构搜索方案。在现有的方法中，一个设计良好、大小受限的搜索空间的必要性是默认的，我们的NSE旨在最小化这种必要性。具体地说，NSE从搜索空间子集开始，然后通过重复两个步骤进化搜索空间：1) 从搜索空间子集搜索优化空间，2) 从未遍历的大型操作池中重新填充该子集。通过引入可学习的多分支设置，我们进一步扩展了可获得体系结构的灵活性。利用所提出的方法，我们在使用333M触发器的ImageNet上实现了77.3%的top-1再训练精度，这在以前不涉及知识提取或权重修剪的自动生成体系结构中产生了最先进的性能。当采用延迟约束时，我们的结果也比以前性能最好的移动模型具有77.9%的Top-1再训练精度。代码可在[https://github.com/orashi/NSE\\_NAS](https://github.com/orashi/NSE_NAS).

With the development of deep convolutional neural networks, image matting has ushered in a new phase. Regarding the nature of image matting, most researches have focused on solutions for transition regions. However, we argue that many existing approaches are excessively focused on transition-dominant local fields and ignored the inherent coordination between global information and transition optimisation. In this paper, we propose the Tripartite Information Mining and Integration Network (TIMI-Net) to harmonize the coordination between global and local attributes formally. Specifically, we resort to a novel 3-branch encoder to accomplish comprehensive mining of the input information, which can supplement the neglected coordination between global and local fields. In order to achieve effective and complete interaction between such multi-branches information, we develop the Tripartite Information Integration ( $TI^2$ ) Module to transform and integrate the interconnections between the different branches. In addition, we built a large-scale human matting dataset (Human-2K) to advance human image matting, which consists of 2100 high-precision human images (2000 images for training and 100 images for test). Finally, we conduct extensive experiments to prove the performance of our proposed TIMI-Net, which demonstrates that our method performs favourably against the SOTA approaches on the alphamatting.com (Rank First), Composition-1K (MSE-0.006, Grad-11.5), Distinctions-646 and our Human-2K. Also, we have developed an online evaluation website to perform natural image matting. Project page: <https://wukaoliu.github.io/TIMI-Net>.

随着深度卷积神经网络的发展，图像抠图技术进入了一个新的阶段。关于图像抠图的性质，大多数研究都集中在过渡区域的解决方案上。然而，我们认为，许多现有的方法过于关注过渡主导的局部场，而忽略了全局信息和过渡优化之间的内在协调。在本文中，我们提出了三方信息挖掘和集成网络（TIMI-Net），以正式协调全局和局部属性之间的协调。具体来说，我们采用了一种新颖的三分支编码器来完成输入信息的综合挖掘，这可以补充被忽略的全局和局部字段之间的协调。为了实现此类多分支信息之间的有效和完整交互，我们开发了三方信息集成（ $TI^2$ ）模块，以转换和集成不同分支之间的互连。此外，我们还构建了一个大规模的人体matting数据集（human-2K）来推进人体图像matting，该数据集由2100张高精度人体图像（2000张用于训练的图像和100张用于测试的图像）组成。最后，我们进行了大量的实验来证明我们提出的TIMI网络的性能，这表明我们的方法在alphamatting上优于SOTA方法。com（排名第一）、Composition-1K（MSE-0.006、Grad-11.5）、dictionaries-646和Human-2K。此外，我们还开发了一个在线评估网站，以执行自然图像抠图。项目页面：<https://wukaoliu.github.io/TIMI-Net>。

Learning mid-level representation for fine-grained recognition is easily dominated by a limited number of highly discriminative patterns, degrading its robustness and generalization capability. To this end, we propose a novel Stochastic Partial Swap (SPS) scheme to address this issue. Our method performs element-wise swapping for partial features between samples to inject noise during training. It equips a regularization effect similar to Dropout, which promotes more neurons to represent the concepts. Furthermore, it also exhibits other advantages: 1) suppressing over-activation to some part patterns to improve feature representativeness, and 2) enriching pattern combination and simulating noisy cases to enhance classifier generalization. We verify the effectiveness of our approach through comprehensive experiments across four network backbones and three fine-grained datasets. Moreover, we demonstrate its ability to complement high-level representations, allowing a simple model to achieve performance comparable to the top-performing technologies in fine-grained recognition, indoor scene recognition, and material recognition while improving model interpretability.

学习用于细粒度识别的中级表示很容易被数量有限的高分辨模式所控制，从而降低其鲁棒性和泛化能力。为此，我们提出了一种新的随机部分交换（SPS）方案来解决这个问题。我们的方法对样本之间的部分特征进行元素交换，以在训练期间注入噪声。它装备了一种类似于辍学的正则化效应，促进了更多的神经元来代表概念。此外，它还显示出其他优点：1）抑制对某些零件模式的过度激活，以提高特征的代表性；2）丰富模式组合并模拟噪声情况，以增强分类器的泛化能力。我们通过四个网络主干和三个细粒度数据集的综合实验验证了我们方法的有效性。此外，我们还展示了其补充高级表示的能力，允许一个简单的模型实现与细粒度识别、室内场景识别和材料识别中的顶级技术相当的性能，同时提高模型的可解释性。

Social distancing, an essential public health measure to limit the spread of contagious diseases, has gained significant attention since the outbreak of the COVID-19 pandemic. In this work, the problem of visual social distancing compliance assessment in busy public areas, with wide field-of-view cameras, is considered. A dataset of crowd scenes with people annotations under a bird's eye view (BEV) and ground truth for metric distances is introduced, and several measures for the evaluation of social distance detection systems are proposed. A multi-branch network, BEV-Net, is proposed to localize individuals in world coordinates and identify high-risk regions where social distancing is violated. BEV-Net combines detection of head and feet locations, camera pose estimation, a differentiable homography module to map image into BEV coordinates, and geometric reasoning to produce a BEV map of the people locations in the scene. Experiments on complex crowded scenes demonstrate the power of the approach and show superior performance over baselines derived from methods in the literature. Applications of interest for public health decision makers are finally discussed. Datasets, code and pretrained models are publicly available at GitHub.

Oracle Distancing是限制传染性疾病传播的一项重要公共卫生措施，自从COVID-19流行病爆发以来得到了广泛关注。在这项工作中，视觉社会距离合规性评估的问题，在繁忙的公共领域，与大视场摄像机，是考虑。介绍了一个在鸟瞰视图（BEV）和地面真实度下具有人物注释的人群场景数据集，并提出了几种评价社会距离检测系统的方法。提出了一种多分支网络，即BEV网络，用于在世界坐标系中定位个体，并识别违反社会距离的高风险区域。BEV网络结合了头和脚位置检测、摄像机姿势估计、将图像映射到BEV坐标的可微单应性模块以及生成场景中人物位置的BEV地图的几何推理。在复杂拥挤场景上的实验证明了该方法的有效性，并显示出优于文献中方法得出的基线的性能。最后讨论了公共卫生决策者感兴趣的应用。数据集、代码和预训练模型可在GitHub上公开获取。

Multi-scale and multi-patch deep models have been shown effective in removing blurs of dynamic scenes. However, these methods still have one major obstacle: manually designing a lightweight and high-efficiency network is challenging and time-consuming. To tackle this problem, we propose a novel deblurring method, dubbed PyNAS (pyramid neural architecture search network), towards automatically designing hyper-parameters including the scales, patches, and standard cell operators. The proposed PyNAS adopts gradient-based search strategies and innovatively searches the hierarchy patch and scale scheme not limited to the cell searching. Specifically, we introduce a hierarchical search strategy tailored for the multi-scale and multi-patch deblurring task. The strategy follows the principle that the first distinguishes between the top-level (pyramid-scales and pyramid-patches) and bottom-level variables (cell operators) and then searches multi-scale variables using the top-to-bottom principle. During the search stage, PyNAS employs an early stopping strategy to avoid the collapse and computational issue. Furthermore, we use a path-level binarization mechanism for multi-scale cell searching to save memory consumption. Our model is a real-time deblurring algorithm (around 58 fps) for 720p images while achieves state-of-the-art deblurring performance on the GoPro and Video Deblurring dataset.

多尺度和多面片深度模型已被证明能有效地去除动态场景中的模糊。然而，这些方法仍然有一个主要的障碍：手工设计一个轻量级和高效的网络是一个挑战和耗时的过程。为了解决这个问题，我们提出了一种新的去模糊方法，称为PyNAS（金字塔神经结构搜索网络），用于自动设计超参数，包括尺度、面片和标准单元算子。所提出的PyNAS采用基于梯度的搜索策略，并创新性地搜索不限于单元搜索的层次补丁和缩放方案。具体来说，我们介绍了一种为多尺度和多块去模糊任务定制的分层搜索策略。该策略遵循以下原则：首先区分顶层（金字塔比例和金字塔面片）和底层变量（单元运算符），然后使用自上而下原则搜索多尺度变量。在搜索阶段，PyNAS采用早期停止策略以避免崩溃和计算问题。此外，我们使用路径级二值化机制进行多尺度单元搜索，以节省内存消耗。我们的模型是针对720p图像的实时去模糊算法（约58 fps），同时在GoPro和视频去模糊数据集上实现了最先进的去模糊性能。

We recover high-frequency information encoded in the shadows cast by an object to estimate a hemispherical photograph from the viewpoint of the object, effectively turning objects into cameras. Estimating environment maps is useful for advanced image editing tasks such as relighting, object insertion or removal, and material parameter estimation. Because the problem is ill-posed, recent works in illumination recovery have tackled the problem of low-frequency lighting for object insertion, rely upon specular surface materials, or make use of data-driven methods that are susceptible to hallucination without physically plausible constraints. We incorporate an optimization scheme to update scene parameters that could enable practical capture of real-world scenes. Furthermore, we develop a methodology for evaluating expected recovery performance for different types and shapes of objects.

我们恢复编码在物体投射阴影中的高频信息，从物体的角度估计半球形照片，有效地将物体转化为相机。估算环境贴图对于高级图像编辑任务（如重新照明、对象插入或删除以及材质参数估算）非常有用。因为这个问题是不稳定的，最近在照明恢复方面的工作已经解决了低频照明用于物体插入的问题，依赖镜面反射表面材料，或者使用易产生幻觉的数据驱动方法，而没有物理上合理的约束。我们采用了一种优化方案来更新场景参数，从而能够实际捕获真实场景。此外，我们还开发了一种评估不同类型和形状对象的预期恢复性能的方法。

We study the problem of aligning two sets of 3D geometric primitives given known correspondences. Our first contribution is to show that this primitive alignment framework unifies five perception problems including point cloud registration, primitive (mesh) registration, category-level 3D registration, absolute pose estimation (APE), and category-level APE. Our second contribution is to propose DynAMical Pose estimation (DAMP), the first general and practical algorithm to solve primitive alignment problem by simulating rigid body dynamics arising from virtual springs and damping, where the springs span the shortest distances between corresponding primitives. We evaluate DAMP in simulated and real datasets across all five problems, and demonstrate (i) DAMP always converges to the globally optimal solution in the first three problems with 3D-3D correspondences; (ii) although DAMP sometimes converges to suboptimal solutions in the last two problems with 2D-3D correspondences, using a scheme for escaping local minima, DAMP always succeeds. Our third contribution is to demystify the surprising empirical performance of DAMP and formally prove a global convergence result in the case of point cloud registration by characterizing local stability of the equilibrium points of the underlying dynamical system.

我们研究了给定已知对应关系的两组三维几何图元的对齐问题。我们的第一个贡献是展示这个基本对齐框架统一了五个感知问题，包括点云注册、基本体（网格）注册、类别级3D注册、绝对姿势估计（APE）和类别级APE。我们的第二个贡献是提出了动态姿态估计（DAMP），这是第一个通过模拟由虚拟弹簧和阻尼产生的刚体动力学来解决原始对准问题的通用实用算法，其中弹簧跨越对应原始体之间的最短距离。我们在所有五个问题的模拟和真实数据集中评估了DAMP，并证明（i）在具有3D-3D对应的前三个问题中，DAMP始终收敛于全局最优解；（ii）尽管在最后两个2D-3D对应问题中，DAMP有时会收敛到次优解，但使用逃避局部极小值的方案，DAMP总是成功的。我们的第三个贡献是通过描述潜在动力系统平衡点的局部稳定性，揭开DAMP令人惊讶的经验性能的神秘面纱，并正式证明点云注册情况下的全局收敛结果。

Training sample re-weighting is an effective approach for tackling data biases such as imbalanced and corrupted labels. Recent methods develop learning-based algorithms to learn sample re-weighting strategies jointly with model training based on the frameworks of reinforcement learning and meta learning. However, depending on additional unbiased reward data is limiting their general applicability. Furthermore, existing learning-based sample re-weighting methods require nested optimizations of models and weighting parameters, which requires expensive second-order computation. This paper addresses these two problems and presents a novel learning-based fast sample re-weighting (FSR) method that does not require additional reward data. The method is based on two key ideas: learning from history to build proxy reward data and feature sharing to reduce the optimization cost. Our experiments show the proposed method achieves competitive results compared to state of the arts on label noise robustness and long-tailed recognition, and does so while achieving significantly improved training efficiency. The source code is publicly available at <https://github.com/google-research/google-research/tree/master/ieg>.

训练样本重新加权是解决数据偏差（如标签不平衡和损坏）的有效方法。最近的方法基于强化学习和元学习的框架，开发了基于学习的算法，结合模型训练来学习样本重新加权策略。然而，依赖于额外的无偏奖励数据限制了它们的普遍适用性。此外，现有的基于学习的样本重加权方法需要对模型和加权参数进行嵌套优化，这需要昂贵的二阶计算。本文针对这两个问题，提出了一种新的基于学习的快速样本重加权（FSR）方法，该方法不需要额外的奖励数据。该方法基于两个关键思想：从历史中学习建立代理奖励数据和特征共享以降低优化成本。我们的实验表明，与现有的标签噪声鲁棒性和长尾识别技术相比，该方法取得了具有竞争力的结果，同时显著提高了训练效率。源代码可在<https://github.com/google-research/google-research/tree/master/ieg>。

The semi-supervised semantic segmentation methods utilize the unlabeled data to increase the feature discriminative ability to alleviate the burden of the annotated data. However, the dominant consistency learning diagram is limited by a) the misalignment between features from labeled and unlabeled data; b) treating each image and region separately without considering crucial semantic dependencies among classes. In this work, we introduce a novel C<sup>3</sup>-Semiseg to improve consistency-based semi-supervised learning by exploiting better feature alignment under perturbations and enhancing discriminative of the inter-class features cross images. Specifically, we first introduce a cross-set region-level data augmentation strategy to reduce the feature discrepancy between labeled data and unlabeled data. Cross-set pixel-wise contrastive learning is further integrated into the pipeline to facilitate discriminative and consistent intra-class features in a 'compared to learn' way. To stabilize training from the noisy label, we propose a dynamic confidence region selection strategy to focus on the high confidence region for loss calculation. We validate the proposed approach on Cityscapes and BDD100K dataset, which significantly outperforms other state-of-the-art semi-supervised semantic segmentation methods.

半监督语义分割方法利用未标记数据提高特征识别能力，减轻标注数据的负担。然而，主要的一致性学习图受到以下限制：a) 标记和未标记数据的特征之间的不对齐；b) 分别处理每个图像和区域，而不考虑类之间的关键语义依赖关系。在这项工作中，我们引入了一种新的C<sup>3</sup>-Seg，通过在扰动下利用更好的特征对齐和增强跨图像类间特征的区分性来改进基于一致性的半监督学习。具体地说，我们首先引入了一种交叉集区域级数据扩充策略，以减少标记数据和未标记数据之间的特征差异。交叉集像素级对比学习进一步整合到管道中，以“对比学习”的方式促进区分性和一致的类内特征。为了稳定噪声标签下的训练，我们提出了一种动态置信域选择策略，将重点放在高置信域上进行损失计算。我们在Cityscapes和BDD100K数据集上验证了所提出的方法，该方法明显优于其他先进的半监督语义分割方法。

There has been a recent surge of interest in cross-modal pre-training. However, existed approaches pre-train a one-stream model to learn joint vision-language representation, which suffers from calculation explosion when conducting cross-modal retrieval. In this work, we propose the Contrastive Cross-Modal Knowledge Sharing Pre-training (COOKIE) method to learn universal text-image representations. There are two key designs in it, one is the weight-sharing transformer on top of the visual and textual encoders to align text and image semantically, the other is three kinds of contrastive learning designed for sharing knowledge between different modalities. Cross-modal knowledge sharing greatly promotes the learning of unimodal representation. Experiments on multi-modal matching tasks including cross-modal retrieval, text matching, and image retrieval show the effectiveness and efficiency of our pre-training framework. Our COOKIE fine-tuned on cross-modal datasets MSCOCO, Flickr30K, and MSRVTT achieves new state-of-the-art results while using only 3/1000 inference time comparing to one-stream models. There are also 5.7 and 3.9 improvements in the task of image retrieval and text matching. Source code will be made public.

最近，人们对跨模式预培训的兴趣激增。然而，现有的方法预先训练一个单流模型来学习联合视觉语言表示，这在进行跨模态检索时会受到计算爆炸的影响。在这项工作中，我们提出了对比跨模态知识共享预训练（COOKIE）方法来学习通用文本图像表示。其中有两个关键设计，一个是视觉和文本编码器上的权重共享变压器，用于对齐文本和图像语义，另一个是三种对比学习，用于在不同模式之间共享知识。跨模态知识共享极大地促进了单峰表示的学习。在多模态匹配任务（包括跨模态检索、文本匹配和图像检索）上的实验表明了我们的预训练框架的有效性和效率。我们的COOKIE在跨模式数据集MSCOCO、Flickr30K和MSRVTT上进行了微调，获得了最新的结果，同时与单流模型相比，仅使用了3/1000的推断时间。图像检索和文本匹配的任务也有5.7和3.9的改进。源代码将公开。

Factorization methods are frequently used for structure from motion problems (SfM). In the presence of noise they are able to jointly estimate camera matrices and scene points in overdetermined settings, without the need for accurate initial solutions. While the early formulations were restricted to affine models, recent approaches have been shown to work with pinhole cameras by minimizing object space errors. In this paper we propose a factorization approach using the so-called radial camera, which is invariant to radial distortion and changes in focal length. Assuming a known principal point our approach can reconstruct the 3D scene in settings with unknown and varying radial distortion and focal length. We show on both real and synthetic data that our approach outperforms state-of-the-art factorization methods under these conditions.

因式分解方法常用于结构自运动问题 (SfM)。在存在噪声的情况下，它们能够在过度确定的设置中联合估计摄影机矩阵和场景点，而不需要精确的初始解。虽然早期的公式仅限于仿射模型，但最近的方法通过最小化对象空间误差来使用针孔相机。在本文中，我们提出了一种使用所谓的径向相机的因式分解方法，该方法对径向畸变和焦距变化保持不变。假设一个已知的主点，我们的方法可以在具有未知和变化的径向畸变和焦距的环境中重建3D场景。我们在真实数据和合成数据上都表明，在这些条件下，我们的方法优于最先进的因子分解方法。

The widespread use of always-connected digital cameras in our everyday life has led to increasing concerns about the users' privacy and security. How to develop privacy-preserving computer vision systems? In particular, we want to prevent the camera from obtaining detailed visual data that may contain private information. However, we also want the camera to capture useful information to perform computer vision tasks. Inspired by the trend of jointly designing optics and algorithms, we tackle the problem of privacy-preserving human pose estimation by optimizing an optical encoder (hardware-level protection) with a software decoder (convolutional neural network) in an end-to-end framework. We introduce a visual privacy protection layer in our optical encoder that, parametrized appropriately, enables the optimization of the camera lens's point spread function (PSF). We validate our approach with extensive simulations and a prototype camera. We show that our privacy-preserving deep optics approach successfully degrades or inhibits private attributes while maintaining important features to perform human pose estimation.

在我们的日常生活中，经常连接的数码相机的广泛使用已经导致人们越来越关注用户的隐私和安全。如何开发保护隐私的计算机视觉系统？特别是，我们希望防止摄像头获取可能包含私人信息的详细视觉数据。然而，我们也希望相机捕捉有用的信息来执行计算机视觉任务。受光学和算法联合设计趋势的启发，我们通过在端到端框架中优化光学编码器（硬件级保护）和软件解码器（卷积神经网络）来解决保护隐私的人体姿势估计问题。我们在我们的光学编码器中引入了一个视觉隐私保护层，经过适当的参数化，可以优化相机镜头的点扩散函数（PSF）。我们通过大量的仿真和原型摄像机验证了我们的方法。我们证明了我们的隐私保护深光学方法成功地降低或抑制了私有属性，同时保持了重要的特征来执行人体姿势估计。

In this paper, we proposed EPP-MVSNet, a novel deep learning network for 3D reconstruction from multi-view stereo (MVS). EPP-MVSNet can accurately aggregate features at high resolution to a limited cost volume with an optimal depth range, thus, leads to effective and efficient 3D construction. Distinct from existing works which measure feature cost at discrete positions which affects the 3D reconstruction accuracy, EPP-MVSNet introduces an epipolar assembling-based kernel that operates on adaptive intervals along epipolar lines for making full use of the image resolution. Further, we introduce an entropy-based refining strategy where the cost volume describes the space geometry with the little redundancy. Moreover, we design a light-weighted network with Pseudo-3D convolutions integrated to achieve high accuracy and efficiency. We have conducted extensive experiments on challenging datasets Tanks & Temples(TNT), ETH3D and DTU. As a result, we achieve promising results on all datasets and the highest F-Score on the online TNT intermediate benchmark. Code is available at [https://gitee.com/mindspore/mindspore/tree/master/model\\_zoo/research/cv/eppmvsnet](https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/eppmvsnet).

在本文中，我们提出了EPP MVSNet，一种新的深度学习网络，用于从多视图立体（MVS）进行三维重建。EPP MVSNet能够以高分辨率精确地聚合特征，以有限的成本体积和最佳的深度范围，从而实现高效的3D构建。与现有的在离散位置测量影响三维重建精度的特征成本的工作不同，EPP MVSNet引入了基于极线组装的内核，该内核沿极线以自适应间隔操作，以充分利用图像分辨率。此外，我们引入了一种基于熵的细化策略，其中成本量描述了空间几何结构，冗余度很小。此外，我们还设计了一种结合伪三维卷积的轻量级网络，以实现高精度和高效率。我们对具有挑战性的数据集 Tanks&Temple (TNT)、ETH3D和DTU进行了广泛的实验。因此，我们在所有数据集上都取得了令人满意的结果，并且在在线TNT中间基准上获得了最高的F分数。代码可在[https://gitee.com/mindspore/mindspore/tree/master/model\\_zoo/research/cv/eppmvsnet](https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/eppmvsnet).

This work studies the Tensor Robust Principal Component Analysis (TRPCA) problem, which aims to exactly recover the low-rank and sparse components from their sum. Our model is motivated by the recently proposed linear transforms based tensor-tensor product and tensor SVD. We define a new transforms depended tensor rank and the corresponding tensor nuclear norm. Then we solve the TRPCA problem by convex optimization whose objective is a weighted combination of the new tensor nuclear norm and  $\ell_1$ -norm. In theory, we prove that under some incoherence conditions, the convex program exactly recovers the underlying low-rank and sparse components with high probability. Our new TRPCA is much more general since it allows to use any invertible linear transforms. Thus, we have more choices in practice for different tasks and different type of data. Numerical experiments verify our results and the application on image recovery demonstrates the superiority of our method.

本文研究了张量稳健主成分分析 (TRPCA) 问题，其目的是从低秩和稀疏分量的和中准确地恢复出低秩和稀疏分量。我们的模型是由最近提出的基于张量积和张量奇异值分解的线性变换驱动的。我们定义了一个新的变换依赖张量秩和相应的张量核范数。然后，我们通过凸优化来解决TRPCA问题，其目标是新的张量核范数和 $\ell_1$ 范数的加权组合。在理论上，我们证明了在某些非相干条件下，凸规划能够以高概率精确地恢复底层的低秩稀疏分量。我们新的TRPCA更加通用，因为它允许使用任何可逆线性变换。因此，在实践中，对于不同的任务和不同类型的数据，我们有更多的选择。数值实验验证了我们的结果，在图像恢复中的应用表明了我们方法的优越性。

The rendering procedure used by neural radiance fields (NeRF) samples a scene with a single ray per pixel and may therefore produce renderings that are excessively blurred or aliased when training or testing images observe scene content at different resolutions. The straightforward solution of supersampling by rendering with multiple rays per pixel is impractical for NeRF, because rendering each ray requires querying a multilayer perceptron hundreds of times. Our solution, which we call "mip-NeRF" (a la "mipmap"), extends NeRF to represent the scene at a continuously-valued scale. By efficiently rendering anti-aliased conical frustums instead of rays, mip-NeRF reduces objectionable aliasing artifacts and significantly improves NeRF's ability to represent fine details, while also being 7% faster than NeRF and half the size. Compared to NeRF, mip-NeRF reduces average error rates by 17% on the dataset presented with NeRF and by 60% on a challenging multiscale variant of that dataset that we present. Mip-NeRF is also able to match the accuracy of a brute-force supersampled NeRF on our multiscale dataset while being 22x faster.

神经辐射场 (NeRF) 使用的渲染过程对每像素一条光线的场景进行采样，因此在训练或测试图像时，可能会产生过度模糊或锯齿的渲染，以不同的分辨率观察场景内容。对于NeRF来说，通过每像素渲染多条光线进行超级采样的简单解决方案是不切实际的，因为渲染每条光线需要查询多层次感知器数百次。我们的解决方案，我们称之为“mip NeRF”（一种“mipmap”）扩展了NeRF，以连续值的比例表示场景。通过高效地渲染消除混叠的圆锥形截锥体而不是光线，mip NeRF减少了令人不快的混叠瑕疵，并显著提高了NeRF表示精细细节的能力，同时比NeRF快7%，大小为NeRF的一半。与NeRF相比，mip NeRF在使用NeRF呈现的数据集上降低了17%的平均错误率，在我们呈现的数据集的具有挑战性的多尺度变体上降低了60%的平均错误率。Mip NeRF还能够在我们的多尺度数据集上与强力超采样NeRF的精度相匹配，同时速度快22倍。

Recently, the problem of inaccurate learning targets in crowd counting draws increasing attention. Inspired by a few pioneering work, we solve this problem by trying to predict the indices of pre-defined interval bins of counts instead of the count values themselves. However, an inappropriate interval setting might make the count error contributions from different intervals extremely imbalanced, leading to inferior counting performance. Therefore, we propose a novel count interval partition criterion called Uniform Error Partition (UEP), which always keeps the expected counting error contributions equal for all intervals to minimize the prediction risk. Then to mitigate the inevitably introduced discretization errors in the count quantization process, we propose another criterion called Mean Count Proxies (MCP). The MCP criterion selects the best count proxy for each interval to represent its count value during inference, making the overall expected discretization error of an image nearly negligible. As far as we are aware, this work is the first to delve into such a classification task and ends up with a promising solution for count interval partition. Following the above two theoretically demonstrated criterions, we propose a simple yet effective model termed Uniform Error Partition Network (UEPNet), which achieves state-of-the-art performance on several challenging datasets. The codes will be available at:

<https://github.com/TencentYoutuResearch/CrowdCounting-UEPNet>.

近年来，人群计数中学习目标不准确的问题越来越引起人们的关注。受一些开创性工作的启发，我们通过尝试预测预定义的计数区间箱的指数而不是计数值本身来解决这个问题。但是，不适当的间隔设置可能会使来自不同间隔的计数误差贡献极不平衡，从而导致较差的计数性能。因此，我们提出了一种新的计数区间划分准则，称为均匀误差划分 (UEP)，它始终保持所有区间的预期计数误差贡献相等，以最小化预测风险。然后，为了缓解计数量化过程中不可避免地引入的离散化错误，我们提出了另一个标准，称为平均计数代理 (MCP)。MCP准则为每个间隔选择最佳计数代理，以表示推断期间的计数值，使得图像的总体预期离散化误差几乎可以忽略不计。据我们所知，这项工作是第一次深入研究这样的分

类任务，并最终为计数区间划分提供了一个有希望的解决方案。根据上述两个理论证明的标准，我们提出了一个简单而有效的模型，称为均匀误差划分网络（UEPNet），它在几个具有挑战性的数据集上实现了最先进的性能。代码将在以下位置提供：<https://github.com/TencentYoutuResearch/CrowdCounting-UEPNet>。

High dynamic range (HDR) video reconstruction from sequences captured with alternating exposures is a very challenging problem. Existing methods often align low dynamic range (LDR) input sequence in the image space using optical flow, and then merge the aligned images to produce HDR output. However, accurate alignment and fusion in the image space are difficult due to the missing details in the over-exposed regions and noise in the under-exposed regions, resulting in unpleasing ghosting artifacts. To enable more accurate alignment and HDR fusion, we introduce a coarse-to-fine deep learning framework for HDR video reconstruction. Firstly, we perform coarse alignment and pixel blending in the image space to estimate the coarse HDR video. Secondly, we conduct more sophisticated alignment and temporal fusion in the feature space of the coarse HDR video to produce better reconstruction. Considering the fact that there is no publicly available dataset for quantitative and comprehensive evaluation of HDR video reconstruction methods, we collect such a benchmark dataset, which contains 97 sequences of static scenes and 184 testing pairs of dynamic scenes. Extensive experiments show that our method outperforms previous state-of-the-art methods. Our dataset, code and model will be made publicly available.

从交替曝光捕获的序列重建高动态范围 (HDR) 视频是一个非常具有挑战性的问题。现有的方法通常使用光流在图像空间中对齐低动态范围 (LDR) 输入序列，然后合并对齐的图像以产生HDR输出。然而，由于过度曝光区域中的细节缺失和欠曝光区域中的噪声，图像空间中的精确对齐和融合非常困难，从而导致不令人满意的重影伪影。为了实现更精确的对齐和HDR融合，我们引入了一个从粗到精的HDR视频重建深度学习框架。首先，我们在图像空间中进行粗略对齐和像素混合，以估计粗略的HDR视频。其次，我们在粗HDR视频的特征空间中进行更精细的对齐和时间融合，以产生更好的重建效果。考虑到目前还没有公开的数据集对HDR视频重建方法进行定量和综合评价，我们收集了这样一个基准数据集，其中包含97个静态场景序列和184个动态场景测试对。大量的实验表明，我们的方法优于以前最先进的方法。我们的数据集、代码和模型将公开提供。

Deep learning has achieved remarkable progress for visual recognition on large-scale balanced datasets but still performs poorly on real-world long-tailed data. Previous methods often adopt class re-balanced training strategies to effectively alleviate the imbalance issue, but might be a risk of over-fitting tail classes. The recent decoupling method overcomes over-fitting issues by using a multi-stage training scheme, yet, it is still incapable of capturing tail class information in the feature learning stage. In this paper, we show that soft label can serve as a powerful solution to incorporate label correlation into a multi-stage training scheme for long-tailed recognition. The intrinsic relation between classes embodied by soft labels turns out to be helpful for long-tailed recognition by transferring knowledge from head to tail classes. Specifically, we propose a conceptually simple yet particularly effective multi-stage training scheme, termed as Self Supervised to Distillation (SSD). This scheme is composed of two parts. First, we introduce a self-distillation framework for long-tailed recognition, which can mine the label relation automatically. Second, we present a new distillation label generation module guided by self-supervision. The distilled labels integrate information from both label and data domains that can model long-tailed distribution effectively. We conduct extensive experiments and our method achieves the state-of-the-art results on three long-tailed recognition benchmarks: ImageNet-LT, CIFAR100-LT and iNaturalist 2018. Our SSD outperforms the strong LWS baseline by from 2.7% to 4.5% on various datasets.

深度学习在大规模平衡数据集的视觉识别方面取得了显著的进展，但在现实世界的长尾数据上仍然表现不佳。以往的方法往往采用班级再平衡的训练策略来有效缓解不平衡问题，但可能存在过度拟合尾部班级的风险。最近的解耦方法通过使用多阶段训练方案克服了过拟合问题，但在特征学习阶段仍然无法捕获尾类信息。在本文中，我们证明了软标签可以作为一个强大的解决方案，将标签相关性纳入长尾识别的多阶段训练方案中。软标签所体现的类之间的内在关系通过将知识从头尾类转移到尾类，有助于长尾识别。具体来说，我们提出了一个概念简单但特别有效的多阶段训练方案，称为自监督蒸馏（SSD）。该方案由两部分组成。首先，我们介绍了一个用于长尾识别的自蒸馏框架，它可以自动挖掘标签关系。其次，我们提出了一种新的基于自我监督的蒸馏标签生成模块。提取的标签整合了来自标签和数据域的信息，可以有效地模拟长尾分布。我们进行了广泛的实验，我们的方法在三个长尾识别基准上取得了最先进的结果：ImageNet LT、CIFAR100-LT和iNaturalist 2018。在各种数据集上，我们的SSD的性能比强大的LWS基线高出2.7%到4.5%。

Spotting objects that are visually adapted to their surroundings is challenging for both humans and AI. Conventional generic / salient object detection techniques are suboptimal for this task because they tend to only discover easy and clear objects, while overlooking the difficult-to-detect ones with inherent uncertainties derived from indistinguishable textures. In this work, we contribute a novel approach using a probabilistic representational model in combination with transformers to explicitly reason under uncertainties, namely uncertainty-guided transformer reasoning (UGTR), for camouflaged object detection. The core idea is to first learn a conditional distribution over the backbone's output to obtain initial estimates and associated uncertainties, and then reason over these uncertain regions with attention mechanism to produce final predictions. Our approach combines the benefits of both Bayesian learning and Transformer-based reasoning, allowing the model to handle camouflaged object detection by leveraging both deterministic and probabilistic information. We empirically demonstrate that our proposed approach can achieve higher accuracy than existing state-of-the-art models on CHAMELEON, CAMO and COD10K datasets. Code is available at <https://github.com/fanyang587/UGTR>.

对人类和人工智能来说，发现视觉上适应周围环境的物体都是一项挑战。传统的通用/显著对象检测技术不适合此任务，因为它们往往只发现容易和清晰的对象，而忽略了难以检测的对象，这些对象具有不可区分纹理的固有不确定性。在这项工作中，我们提出了一种新的方法，使用概率表示模型结合变压器，在不确定性条件下进行显式推理，即不确定性引导变压器推理（UGTR），用于伪装目标检测。其核心思想是首先学习主干输出的条件分布，以获得初始估计和相关的不确定性，然后利用注意机制对这些不确定区域进行推理，以产生最终预测。我们的方法结合了贝叶斯学习和基于变换的推理的优点，允许模型通过利用确定性和概率信息来处理伪装目标检测。我们的经验表明，我们提出的方法可以在变色龙、迷彩和COD1K数据集上实现比现有最先进模型更高的精度。代码可在<https://github.com/fanyang587/UGTR>。

Synchronization refers to the problem of inferring the unknown values attached to vertices of a graph where edges are labelled with the ratio of the incident vertices, and labels belong to a group. This paper addresses the synchronization problem on multi-graphs, that are graphs with more than one edge connecting the same pair of nodes. The problem naturally arises when multiple measures are available to model the relationship between two vertices. This happens when different sensors measure the same quantity, or when the original graph is partitioned into sub-graphs that are solved independently. In this case, the relationships among sub-graphs give rise to multi-edges and the problem can be traced back to a multi-graph synchronization. The baseline solution reduces multi-graphs to simple ones by averaging their multi-edges, however this approach falls short because: i) averaging is well defined only for some groups and ii) the resulting estimator is less precise and accurate, as we prove empirically. Specifically, we present MultiSynch, a synchronization algorithm for multi-graphs that is based on a principled constrained eigenvalue optimization. MultiSynch is a general solution that can cope with any linear group and we show to be profitably usable both on synthetic and real problems.

同步指的是推断附加到图顶点的未知值的问题，其中边用关联顶点的比率进行标记，并且标签属于一个组。本文研究了多个图的同步问题，即多个边连接同一对节点的图。当可以使用多个度量来建模两个顶点之间的关系时，问题自然会出现。当不同的传感器测量相同的量时，或者当原始图被划分为独立求解的子图时，就会发生这种情况。在这种情况下，子图之间的关系会产生多条边，问题可以追溯到多图同步。基线解决方案通过平均多个图的多个边将多个图简化为简单图，但是这种方法存在不足，因为：i) 平均仅适用于某些组，ii) 结果估计的精度和准确性较低，正如我们通过经验证明的那样。具体地说，我们提出了MultiSynch，一种基于约束特征值优化的多图同步算法。多同步是一个通用的解决方案，可以处理任何线性组，我们证明了它在合成和实际问题上都是有益的。

Neural implicit 3D representations have emerged as a powerful paradigm for reconstructing surfaces from multi-view images and synthesizing novel views. Unfortunately, existing methods such as DVR or IDR require accurate per-pixel object masks as supervision. At the same time, neural radiance fields have revolutionized novel view synthesis. However, NeRF's estimated volume density does not admit accurate surface reconstruction. Our key insight is that implicit surface models and radiance fields can be formulated in a unified way, enabling both surface and volume rendering using the same model. This unified perspective enables novel, more efficient sampling procedures and the ability to reconstruct accurate surfaces without input masks. We compare our method on the DTU, BlendedMVS, and a synthetic indoor dataset. Our experiments demonstrate that we outperform NeRF in terms of reconstruction quality while performing on par with IDR without requiring masks.

神经隐式3D表示已成为从多视图图像重建曲面和合成新视图的强大范例。不幸的是，DVR或IDR等现有方法需要精确的每像素对象遮罩作为监控。与此同时，神经辐射场已经彻底改变了新的视图合成。然而，NeRF估计的体积密度不允许精确的曲面重建。我们的主要见解是，隐式曲面模型和辐射场可以以统一的方式表示，从而可以使用相同的模型进行曲面和体渲染。这种统一的透视线可以实现新颖、更高效的采样程序，并能够在不使用输入遮罩的情况下重建精确的曲面。我们在DTU、BlendedMVS和合成室内数据集上比较了我们的方法。我们的实验表明，我们优于NERF在重建质量方面，同时执行与IDR PAR，而不需要口罩。

This paper tackles the problem of table structure parsing (TSP) from images in the wild. In contrast to existing studies that mainly focus on parsing well-aligned tabular images with simple layouts from scanned PDF documents, we aim to establish a practical table structure parsing system for real-world scenarios where tabular input images are taken or scanned with severe deformation, bending or occlusions. For designing such a system, we propose an approach named CycleCenterNet on the top of CenterNet with a novel cycle-pairing module to simultaneously detect and group tabular cells into structured tables. In the cycle-pairing module, a new pairing loss function is proposed for the network training. Alongside with our CycleCenterNet, we also present a large-scale dataset, named Wired Table in the Wild (WTW), which includes well-annotated structure reparsing of multiple style tables in several scenes like photo, scanning files, web pages, etc.. In experiments, we demonstrate that our CycleCenterNet consistently achieves the best accuracy of table structure parsing on the new WTW dataset by 24.6% absolute improvement evaluated by the TEDS metric. A more comprehensive experimental analysis also validates the advantages of our proposed methods for the TSP task.

本文解决了从野外图像中提取表结构的问题。现有的研究主要集中于从扫描的PDF文档中解析具有简单布局的对齐良好的表格，与此相反，我们的目标是建立一个实用的表格结构解析系统，用于在表格输入图像被拍摄或扫描时出现严重变形、弯曲或结语。为了设计这样一个系统，我们在CenterNet的顶部提出了一种称为Cycle CenterNet的方法，该方法使用了一个新的Cycle pairing模块来同时检测表格单元格并将其分组到结构化表格中。在循环配对模块中，提出了一种新的网络训练配对损失函数。除了Cycle CenterNet之外，我们还提供了一个名为Wired Table in the Wild (WTW) 的大规模数据集，其中包括在多个场景（如照片、扫描文件、网页等）中对多个样式表进行注释良好的结构解析。。在实验中，我们证明，我们的Cycle CenterNet在新WTW数据集上始终实现了最佳的表结构解析精度，通过TEDS度量评估，绝对提高了24.6%。更全面的实验分析也验证了我们提出的方法在TSP任务中的优势。

Despite much recent progress in video-based person re-identification (re-ID), the current state-of-the-art still suffers from common real-world challenges such as appearance similarity among various people, occlusions, and frame misalignment. To alleviate these problems, we propose Spatio-Temporal Representation Factorization (STRF), a flexible new computational unit that can be used in conjunction with most existing 3D convolutional neural network architectures for re-ID. The key innovations of STRF over prior work include explicit pathways for learning discriminative temporal and spatial features, with each component further factorized to capture complementary person-specific appearance and motion information. Specifically, temporal factorization comprises two branches, one each for static features (e.g., the color of clothes) that do not change much over time, and dynamic features (e.g., walking patterns) that change over time. Further, spatial factorization also comprises two branches to learn both global (coarse segments) as well as local (finer segments) appearance features, with the local features particularly useful in cases of occlusion or spatial misalignment. These two factorization operations taken together result in a modular architecture for our parameter-wise light STRF unit that can be plugged in between any two 3D convolutional layers, resulting in an end-to-end learning framework. We empirically show that STRF improves performance of various existing baseline architectures while demonstrating new state-of-the-art results using standard person re-ID evaluation protocols on three benchmarks.

尽管基于视频的人物再识别 (re-ID) 技术最近取得了很多进展，但目前的技术水平仍然面临着现实世界中常见的挑战，如不同人物之间的外观相似性、遮挡和帧错位。为了缓解这些问题，我们提出了时空表示因子分解 (STRF)，一种灵活的新计算单元，可与大多数现有的3D卷积神经网络架构结合使用，用于re-ID。STRF在先前工作中的主要创新包括学习辨别性时间和空间特征的明确途径，每个组件进一步分解，以捕获互补的特定于人的外观和运动信息。具体而言，时间分解包括两个分支，一个分支用于静态

特征（例如，衣服的颜色），随着时间的推移变化不大，另一个分支用于动态特征（例如，行走模式）。此外，空间分解还包括两个分支来学习全局（粗段）和局部（细段）外观特征，其中局部特征在遮挡或空间错位的情况下特别有用。这两个因式分解操作结合在一起，为我们的参数化light STRF单元形成了一个模块化架构，该架构可插入任意两个3D卷积层之间，从而形成端到端学习框架。我们的经验表明，STRF提高了各种现有基线体系结构的性能，同时在三个基准上使用标准人员重新识别评估协议展示了最新的结果。

Modern deep-learning-based lane detection methods are successful in most scenarios but struggling for lane lines with complex topologies. In this work, we propose CondLaneNet, a novel top-to-down lane detection framework that detects the lane instances first and then dynamically predicts the line shape for each instance. Aiming to resolve lane instance-level discrimination problem, we introduce a conditional lane detection strategy based on conditional convolution and row-wise formulation. Further, we design the Recurrent Instance Module(RIM) to overcome the problem of detecting lane lines with complex topologies such as dense lines and fork lines. Benefit from the end-to-end pipeline which requires little post-process, our method has real-time efficiency. We extensively evaluate our method on three benchmarks of lane detection. Results show that our method achieves state-of-the-art performance on all three benchmark datasets. Moreover, our method has the coexistence of accuracy and efficiency, e.g. a 78.14 F1 score and 220 FPS on CULane. Our code is available at <https://github.com/aliyun/conditional-lane-detection>.

现代的基于深度学习的车道检测方法在大多数情况下都是成功的，但对于复杂拓扑的车道线来说却很困难。在这项工作中，我们提出了CondLaneNet，一种新的自上而下车道检测框架，该框架首先检测车道实例，然后动态预测每个实例的线形。针对车道实例级判别问题，提出了一种基于条件卷积和行公式的条件车道检测策略。此外，我们还设计了递归实例模块（RIM）来解决具有复杂拓扑结构（如密集线和分叉线）的车道线检测问题。由于采用了端到端的流水线结构，后处理量小，因此该方法具有实时性好的优点。我们在车道检测的三个基准上广泛地评估了我们的方法。结果表明，我们的方法在所有三个基准数据集上都达到了最先进的性能。此外，我们的方法具有准确性和效率并存的特点，例如，在CULane上的F1分数为78.14，FPS为220。我们的代码可在<https://github.com/aliyun/conditional-lane-detection>。

Growing at a fast pace, modern autonomous systems will soon be deployed at scale, opening up the possibility for cooperative multi-agent systems. Sharing information and distributing workloads allow autonomous agents to better perform tasks and increase computation efficiency. However, shared information can be modified to execute adversarial attacks on deep learning models that are widely employed in modern systems. Thus, we aim to study the robustness of such systems and focus on exploring adversarial attacks in a novel multi-agent setting where communication is done through sharing learned intermediate representations of neural networks. We observe that an indistinguishable adversarial message can severely degrade performance, but becomes weaker as the number of benign agents increases. Furthermore, we show that black-box transfer attacks are more difficult in this setting when compared to directly perturbing the inputs, as it is necessary to align the distribution of learned representations with domain adaptation. Our work studies robustness at the neural network level to contribute an additional layer of fault tolerance to modern security protocols for more secure multi-agent systems.

现代自治系统以快速的速度发展，不久将大规模部署，为多智能体协作系统提供了可能。共享信息和分配工作负载使自治代理能够更好地执行任务并提高计算效率。但是，可以修改共享信息，对现代系统中广泛使用的深度学习模型执行对抗性攻击。因此，我们的目标是研究这类系统的鲁棒性，并重点探讨在一种新的多智能体环境中的对抗性攻击，其中通信是通过共享学习到的神经网络中间表示来完成的。我们观察到，无法区分的敌对消息可能会严重降低性能，但随着良性代理数量的增加，其性能会变弱。此

外，我们还表明，与直接扰动输入相比，在这种情况下，黑盒转移攻击更为困难，因为有必要将学习表示的分布与域自适应对齐。我们的工作是研究神经网络层面的稳健性，为现代安全协议提供额外的容错层，以实现更安全的多智能体系统。

High-fidelity face digitization solutions often combine multi-view stereo (MVS) techniques for 3D reconstruction and a non-rigid registration step to establish dense correspondence across identities and expressions. A common problem is the need for manual clean-up after the MVS step, as 3D scans are typically affected by noise and outliers and contain hairy surface regions that need to be cleaned up by artists. Furthermore, mesh registration tends to fail for extreme facial expressions. Most learning-based methods use an underlying 3D morphable model (3DMM) to ensure robustness, but this limits the output accuracy for extreme facial expressions. In addition, the global bottleneck of regression architectures cannot produce meshes that tightly fit the ground truth surfaces. We propose ToFu, Topological consistent Face from multi-view, a geometry inference framework that can produce topologically consistent meshes across facial identities and expressions using a volumetric representation instead of an explicit underlying 3DMM. Our novel progressive mesh generation network embeds the topological structure of the face in a feature volume, sampled from geometry-aware local features. A coarse-to-fine architecture facilitates dense and accurate facial mesh predictions in a consistent mesh topology. ToFu further captures displacement maps for pore-level geometric details and facilitates high-quality rendering in the form of albedo and specular reflectance maps. These high-quality assets are readily usable by production studios for avatar creation, animation and physically-based skin rendering. We demonstrate state-of-the-art geometric and correspondence accuracy, while only taking 0.385 seconds to compute a mesh with 10K vertices, which is three orders of magnitude faster than traditional techniques. The code and the model are available for research purposes at <https://tianyeli.github.io/tofu>.

高保真人脸识别解决方案通常结合用于三维重建的多视图立体（MVS）技术和非刚性配准步骤，以在身份和表情之间建立紧密的对应关系。一个常见的问题是在MVS步骤之后需要手动清理，因为3D扫描通常会受到噪声和异常值的影响，并且包含需要艺术家清理的毛茸茸的曲面区域。此外，对于极端的面部表情，网格配准往往失败。大多数基于学习的方法使用底层3D变形模型（3DMM）来确保鲁棒性，但这限制了极端面部表情的输出精度。此外，回归架构的全局瓶颈无法生成紧密贴合地面真实曲面的网格。我们提出了ToFu，多视图拓扑一致性人脸，这是一个几何推理框架，它可以使用体积表示而不是显式的底层3DMM，跨人脸身份和表情生成拓扑一致的网格。我们的新型渐进式网格生成网络将人脸的拓扑结构嵌入从几何感知局部特征中采样的特征体中。从粗到精的体系结构有助于在一致的网格拓扑中进行密集和精确的面部网格预测。ToFu进一步捕获孔隙级几何细节的位移图，并以反照率和镜面反射图的形式促进高质量渲染。这些高质量资产可供制作工作室用于头像创建、动画和基于物理的皮肤渲染。我们展示了最先进的几何和对应精度，同时只需0.385秒即可计算出具有10K个顶点的网格，比传统技术快三个数量级。有关代码和模型的研究目的，请访问<https://tianyeli.github.io/tofu>。

Video compression is a critical component of Internet video delivery. Recent work has shown that deep learning techniques can rival or outperform human-designed algorithms, but these methods are significantly less compute and power-efficient than existing codecs. This paper presents a new approach that augments existing codecs with a small, content-adaptive super-resolution model that significantly boosts video quality. Our method, SRVC, encodes video into two bitstreams: (i) a content stream, produced by compressing downsampled low-resolution video with the existing codec, (ii) a model stream, which encodes periodic updates to a lightweight super-resolution neural network customized for short segments of the video. SRVC decodes the video by passing the decompressed low-resolution video frames through the (time-varying) super-resolution model to reconstruct high-resolution video frames. Our results show that to achieve the same PSNR, SRVC requires 20% of the bits-per-pixel of H.265 in slow mode, and 3% of the bits-per-pixel of DVC, a recent deep learning-based video compression scheme. SRVC runs at 90 frames per second on an NVIDIA V100 GPU.

视频压缩是互联网视频传输的关键组成部分。最近的研究表明，深度学习技术可以与人类设计的算法相媲美或优于人类设计的算法，但这些方法的计算效率和功耗明显低于现有的编解码器。本文提出了一种新的方法，该方法通过一个小的、内容自适应的超分辨率模型来增强现有的编解码器，从而显著提高视频质量。我们的方法SRVC将视频编码为两个比特流：(i) 内容流，通过使用现有编解码器压缩下采样的低分辨率视频产生；(ii) 模型流，对为视频短片段定制的轻型超分辨率神经网络的周期更新进行编码。SRVC通过将解压缩的低分辨率视频帧通过（时变）超分辨率模型来重构高分辨率视频帧，从而对视频进行解码。我们的结果表明，为了获得相同的峰值信噪比，SRVC在慢模式下需要H.265每像素20%的比特数，而DVC（一种最新的基于深度学习的视频压缩方案）每像素3%的比特数。SRVC在NVIDIA V100 GPU上以每秒90帧的速度运行。

Most monocular depth sensing methods use conventionally captured images that are created without considering scene content. In contrast, animal eyes have fast mechanical motions, called saccades, that control how the scene is imaged by the fovea, where resolution is highest. In this paper, we present the SaccadeCam framework for adaptively distributing resolution onto regions of interest in the scene. Our algorithm for adaptive resolution is a self-supervised network and we demonstrate results for end-to-end learning for monocular depth estimation. We also show preliminary results with a real SaccadeCam hardware prototype.

大多数单目深度传感方法使用常规捕获的图像，这些图像是在不考虑场景内容的情况下创建的。相比之下，动物的眼睛有快速的机械运动，称为扫视，它控制场景如何通过中心凹成像，中心凹的分辨率最高。在本文中，我们提出了一种用于自适应地将分辨率分布到场景中感兴趣区域的扫视摄像机框架。我们的自适应分辨率算法是一个自监督网络，我们展示了单目深度估计的端到端学习结果。我们还展示了一个真实的扫视摄像头硬件原型的初步结果。

Traditional normalization techniques (e.g., Batch Normalization and Instance Normalization) generally and simplistically assume that training and test data follow the same distribution. As distribution shifts are inevitable in real-world applications, well-trained models with previous normalization methods can perform badly in new environments. Can we develop new normalization methods to improve generalization robustness under distribution shifts? In this paper, we answer the question by proposing CrossNorm and SelfNorm. CrossNorm exchanges channel-wise mean and variance between feature maps to enlarge training distribution, while SelfNorm uses attention to recalibrate the statistics to bridge gaps between training and test distributions. CrossNorm and SelfNorm can complement each other, though exploring different directions in statistics usage. Extensive experiments on different fields (vision and language), tasks (classification and segmentation), settings (supervised and semi-supervised), and distribution shift types (synthetic and natural) show the effectiveness. Code is available at <https://github.com/amazon-research/crossnorm-selfnorm>

传统的规范化技术（例如，批量规范化和实例规范化）通常简单地假设训练和测试数据遵循相同的分布。由于分布变化在现实世界的应用中是不可避免的，使用以前的规范化方法的经过良好训练的模型在新环境中可能表现不佳。我们能否开发新的规范化方法来提高分布变化下的泛化鲁棒性？在本文中，我们通过提出交叉范数和自范数来回答这个问题。交叉范数在特征图之间交换通道方向的均值和方差以扩大训练分布，而自范数则利用注意力重新校准统计数据以弥合训练分布和测试分布之间的差距。交叉规范和自我规范可以相互补充，尽管在统计使用方面探索了不同的方向。在不同领域（视觉和语言）、任务（分类和分割）、设置（监督和半监督）和分布转移类型（合成和自然）上的大量实验表明了该方法的有效性。代码可在<https://github.com/amazon-research/crossnorm-selfnorm>

In video highlight detection, the goal is to identify the interesting moments within an unedited video. Although the audio component of the video provides important cues for highlight detection, the majority of existing efforts focus almost exclusively on the visual component. In this paper, we argue that both audio and visual components of a video should be modeled jointly to retrieve its best moments. To this end, we propose an audio-visual network for video highlight detection. At the core of our approach lies a bimodal attention mechanism, which captures the interaction between the audio and visual components of a video, and produces fused representations to facilitate highlight detection. Furthermore, we introduce a noise sentinel technique to adaptively discount a noisy visual or audio modality. Empirical evaluations on two benchmark datasets demonstrate the superior performance of our approach over the state-of-the-art methods.

在视频高光检测中，目标是识别未编辑视频中的有趣时刻。尽管视频的音频成分为高光检测提供了重要的线索，但现有的大部分工作几乎完全集中在视觉成分上。在本文中，我们认为视频的音频和视频组件都应该联合建模，以检索其最佳时刻。为此，我们提出了一种用于视频高光检测的视听网络。我们的方法的核心是一种双峰注意机制，它捕获视频的音频和视频组件之间的交互，并生成融合的表示以便于突出显示检测。此外，我们还引入了一种噪声哨兵技术来自适应地对有噪声的视觉或音频模态进行折扣。对两个基准数据集的实证评估表明，我们的方法优于最先进的方法。

Many state-of-the-art few-shot learners focus on developing effective training procedures for feature representations, before using simple (e.g., nearest centroid) classifiers. We take an approach that is agnostic to the features used, and focus exclusively on meta-learning the final classifier layer. Specifically, we introduce MetaQDA, a Bayesian meta-learning generalisation of the classic quadratic discriminant analysis. This approach has several benefits of interest to practitioners: meta-learning is fast and memory efficient, without the need to fine-tune features. It is agnostic to the off-the-shelf features chosen, and thus will continue to benefit from future advances in feature representations. Empirically, it leads to excellent performance in cross-domain few-shot learning, class-incremental few-shot learning, and crucially for real-world applications, the Bayesian formulation leads to state-of-the-art uncertainty calibration in predictions.

在使用简单（例如，最近质心）分类器之前，许多最先进的少数镜头学习者专注于开发有效的特征表示训练程序。我们采用了一种与所使用的特征无关的方法，并专门关注最终分类器层的元学习。具体来说，我们介绍MetaQDA，一种经典二次判别分析的贝叶斯元学习推广。这种方法有几个实践者感兴趣的好处：元学习速度快，内存效率高，无需微调功能。它与所选择的现成功能无关，因此将继续受益于功能表示的未来发展。从经验上看，它在跨域少镜头学习、类增量少镜头学习方面具有优异的性能，对于现实世界的应用来说，至关重要的是，贝叶斯公式能够在预测中实现最先进的不确定性校准。

Learning to model how the world changes as time elapses has proven a challenging problem for the computer vision community. We introduce a self-supervised approach to this problem that solves a multi-modal temporal cycle consistency objective, MMCC, jointly in vision and language. This objective requires a model to learn modality-agnostic functions to predict the future and past that undo each other when composed. We hypothesize that a model trained on this objective will discover long-term temporal dynamics in video. We verify this hypothesis by using the resultant visual representations and predictive models as-is to solve a variety of downstream tasks. Our method outperforms state-of-the-art self-supervised video prediction methods on future action anticipation, temporal image ordering, and arrow-of-time classification tasks, without training on target datasets or their labels.

对于计算机视觉界来说，学习如何模拟世界随着时间的推移而发生的变化已经证明是一个具有挑战性的问题。我们引入了一种自监督方法来解决这个问题，该方法在视觉和语言上联合解决了多模态时间周期一致性目标MMCC。这一目标需要一个模型来学习模态不可知函数，以预测未来和过去，它们在组合时相互撤销。我们假设在这个目标上训练的模型将在视频中发现长期的时间动态。我们通过使用生成的视觉表示和预测模型来验证这一假设，以解决各种下游任务。在未对目标数据集或其标签进行训练的情况下，我们的方法在未来动作预测、时间图像排序和时间箭头分类任务方面优于最先进的自监督视频预测方法。

Facial expression recognition (FER) has received increasing interest in computer vision. We propose the TRANSFER model which can learn rich relation-aware local representations. It mainly consists of three components: Multi-Attention Dropping (MAD), ViT-FER, and Multi-head Self-Attention Dropping (MSAD). First, local patches play an important role in distinguishing various expressions, however, few existing works can locate discriminative and diverse local patches. This can cause serious problems when some patches are invisible due to pose variations or viewpoint changes. To address this issue, the MAD is proposed to randomly drop an attention map. Consequently, models are pushed to explore diverse local patches adaptively. Second, to build rich relations between different local patches, the Vision Transformers (ViT) are used in FER, called ViT-FER. Since the global scope is used to reinforce each local patch, a better representation is obtained to boost the FER performance. Thirdly, the multi-head self-attention allows ViT to jointly attend to features from different information subspaces at different positions. Given no explicit guidance, however, multiple self-attentions may extract similar relations. To address this, the MSAD is proposed to randomly drop one self-attention module. As a result, models are forced to learn rich relations among diverse local patches. Our proposed TRANSFER model outperforms the state-of-the-art methods on several FER benchmarks, showing its effectiveness and usefulness.

人脸表情识别 (FER) 在计算机视觉中受到越来越多的关注。我们提出了能够学习丰富的关系感知局部表示的迁移模型。它主要由三个部分组成：多注意力下降（MAD）、ViT FER和多头自我注意力下降（MSAD）。首先，局部斑块在区分各种表达中起着重要作用，然而，现有的研究很少能够定位有区别的、多样的局部斑块。当某些面片由于姿势变化或视点更改而不可见时，这可能会导致严重问题。为了解决这个问题，MAD建议随机删除一个注意图。因此，推动模型自适应地探索不同的局部斑块。其次，为了在不同的局部斑块之间建立丰富的关系，视觉变换器（ViT）用于FER，称为ViT FER。由于全局范围用于增强每个局部补丁，因此获得了更好的表示以提高FER性能。第三，多头自我注意使ViT能够在不同位置共同关注来自不同信息子空间的特征。然而，如果没有明确的指导，多重自我关注可能会产生类似的关系。为了解决这个问题，建议MSAD随机丢弃一个自我注意模块。结果，模型被迫学习不同局部斑块之间的丰富关系。我们提出的转移模型在几个外汇储备基准上优于最先进的方法，显示了其有效性和实用性。

We propose a manifold matching approach to generative models which includes a distribution generator (or data generator) and a metric generator. In our framework, we view the real data set as some manifold embedded in a high-dimensional Euclidean space. The distribution generator aims at generating samples that follow some distribution condensed around the real data manifold. It is achieved by matching two sets of points using their geometric shape descriptors, such as centroid and p-diameter, with learned distance metric; the metric generator utilizes both real data and generated samples to learn a distance metric which is close to some intrinsic geodesic distance on the real data manifold. The produced distance metric is further used for manifold matching. The two networks learn simultaneously during the training process. We apply the approach on both unsupervised and supervised learning tasks: in unconditional image generation task, the proposed method obtains competitive results compared with existing generative models; in super-resolution task, we incorporate the framework in perception-based models and improve visual qualities by producing samples with more natural textures. Experiments and analysis demonstrate the feasibility and effectiveness of the proposed framework.

我们提出了一种生成模型的流形匹配方法，该方法包括分布生成器（或数据生成器）和度量生成器。在我们的框架中，我们将真实数据集视为嵌入高维欧几里得空间的流形。分布生成器旨在生成样本，这些样本遵循围绕真实数据流形压缩的某种分布。它是通过使用两组点的几何形状描述符（例如质心和p-直径）和学习的距离度量来匹配两组点来实现的；度量生成器利用真实数据和生成的样本来学习距离度

量，该距离度量接近真实数据流形上的某个固有测地距离。生成的距离度量进一步用于流形匹配。这两个网络在训练过程中同时学习。我们将该方法应用于无监督和有监督学习任务中：在无条件图像生成任务中，与现有的生成模型相比，该方法获得了有竞争力的结果；在超分辨率任务中，我们将该框架融入基于感知的模型中，并通过生成具有更自然纹理的样本来提高视觉质量。实验和分析证明了该框架的可行性和有效性。

The security of Deep Neural Networks (DNNs) is of great importance due to their employment in various safety-critical applications. DNNs are shown to be vulnerable against the Trojan attack that manipulates the model parameters via poisoned training and gets activated by the pre-defined trigger in inputs during inference. In this work, we present ProFlip, the first targeted Trojan attack framework that can divert the prediction of the DNN to the target class by progressively identifying and flipping a small set of bits in model parameters. At its core, ProFlip consists of three key phases: (i) Determining significant neurons in the last layer; (ii) Generating an effective trigger pattern for the target class; (iii) Identifying a sequence of susceptible bits of DNN parameters stored in the main memory (e.g., DRAM). After model deployment, the adversary can insert the Trojan by flipping the critical bits found by ProFlip using bit flip techniques such as Row Hammer or laser beams. As the result, the altered DNN predicts the target class when the trigger pattern is present in any inputs. We perform extensive evaluations of ProFlip on CIFAR10, SVHN, and ImageNet datasets with ResNet-18 and VGG-16 architectures. Empirical results show that, to reach an attack success rate (ASR) of over 94%, ProFlip requires only 12 bit flips out of 88 million parameter bits for ResNet-18 with CIFAR-10, and 15 bit flips for ResNet-18 with ImageNet. Compared to the SOTA, ProFlip reduces the number of required bits flips by 28x 34x while reaching the same level of ASR.

深度神经网络（DNN）由于在各种安全关键应用中的应用，其安全性非常重要。DNN易受特洛伊木马攻击的攻击，特洛伊木马通过中毒训练操纵模型参数，并在推理过程中被输入中的预定义触发器激活。在这项工作中，我们提出了ProFlip，这是第一个有针对性的特洛伊木马攻击框架，它可以通过逐步识别和翻转模型参数中的一小部分位，将DNN的预测转移到目标类。ProFlip的核心包括三个关键阶段：(i) 确定最后一层的重要神经元；(ii) 生成目标类的有效触发模式；(iii) 识别存储在主存储器（例如DRAM）中的DNN参数的敏感位序列。部署模型后，对手可以使用位翻转技术（如行锤或激光束）翻转ProFlip找到的关键位，从而插入特洛伊木马。因此，当触发器模式出现在任何输入中时，修改后的DNN预测目标类。我们对具有ResNet-18和VGG-16体系结构的CIFAR10、SVHN和ImageNet数据集进行了广泛的ProFlip评估。实证结果表明，要达到94%以上的攻击成功率（ASR），ProFlip只需要对使用CIFAR-10的ResNet-18进行8800万个参数位中的12位翻转，对使用ImageNet的ResNet-18进行15位翻转。与SOTA相比，ProFlip在达到相同ASR水平的同时，将所需位翻转次数减少了28x 34x。

Autonomous highlight detection is crucial for enhancing the efficiency of video browsing on social media platforms. To attain this goal in a data-driven way, one may often face the situation where highlight annotations are not available on the target video category used in practice, while the supervision on another video category (named as source video category) is achievable. In such a situation, one can derive an effective highlight detector on target video category by transferring the highlight knowledge acquired from source video category to the target one. We call this problem cross-category video highlight detection, which has been rarely studied in previous works. For tackling such practical problem, we propose a Dual-Learner-based Video Highlight Detection (DL-VHD) framework. Under this framework, we first design a set-based Learning module (SL-module) to improve the conventional pair-based learning by assessing the highlight extent of a video segment under a broader context. Based on such learning manner, we introduce two different learners to acquire the basic distinction of target category videos and the characteristics of highlight moments on source video category, respectively. These two types of highlight knowledge are further consolidated via knowledge distillation. Extensive experiments on three benchmark datasets demonstrate the superiority of the proposed SL-module, and the DL-VHD method outperforms five typical Unsupervised Domain Adaptation (UDA) algorithms on various cross-category highlight detection tasks.

自主亮点检测对于提高社交媒体平台上视频浏览的效率至关重要。为了以数据驱动的方式实现这一目标，人们通常会面临这样的情况：在实际使用的目标视频类别上，高亮注释不可用，而对另一个视频类别（称为源视频类别）的监控是可以实现的。在这种情况下，通过将从源视频类别获取的高光知识转移到目标视频类别，可以导出目标视频类别上的有效高光检测器。我们称之为跨类别视频高光检测，这在以前的工作中很少被研究。为了解决这一实际问题，我们提出了一种基于双学习者的视频高光检测（DL-VHD）框架。在此框架下，我们首先设计了一个基于集合的学习模块（SL模块），通过在更广泛的背景下评估视频片段的突出程度来改进传统的基于配对的学习。基于这种学习方式，我们引入两个不同的学习者，分别获得目标类别视频的基本区别和源视频类别上突出时刻的特征。这两种类型的突出知识通过知识提炼得到进一步的整合。在三个基准数据集上的大量实验证明了所提出的SL模块的优越性，并且DL-VHD方法在各种跨类别高光检测任务上优于五种典型的无监督域自适应（UDA）算法。

Videos flow as the mixture of language, acoustic, and vision modalities. A thorough video understanding needs to fuse time-series data of different modalities for prediction. Due to the variable receiving frequency for sequences from each modality, there usually exists inherent asynchrony across the collected multimodal streams. Towards an efficient multimodal fusion from asynchronous multimodal streams, we need to model the correlations between elements from different modalities. The recent Multimodal Transformer (MuLT) approach extends the self-attention mechanism of the original Transformer network to learn the crossmodal dependencies between elements. However, the direct replication of self-attention will suffer from the distribution mismatch across different modality features. As a result, the learnt crossmodal dependencies can be unreliable. Motivated by this observation, this work proposes the Modality-Invariant Crossmodal Attention (MICA) approach towards learning crossmodal interactions over modality-invariant space in which the distribution mismatch between different modalities is well bridged. To this end, both the marginal distribution and the elements with high-confidence correlations are aligned over the common space of the query and key vectors which are computed from different modalities. Experiments on three standard benchmarks of multimodal video understanding clearly validate the superiority of our approach.

视频是语言、声音和视觉的混合体。全面的视频理解需要融合不同模式的时间序列数据进行预测。由于来自每个模态的序列的接收频率可变，在所收集的多模态流中通常存在固有的异步性。为了从异步多模态流中获得有效的多模态融合，我们需要对来自不同模态的元素之间的相关性进行建模。最近的多模态变压器 (MuLT) 方法扩展了原始变压器网络的自我注意机制，以了解元件之间的交叉模态依赖关系。然而，自我注意的直接复制会受到不同模态特征分布不匹配的影响。因此，学习到的跨模态依赖关系可能不可靠。基于这一观察结果，本研究提出了模态不变跨模态注意 (MICA) 方法，用于在模态不变空间上学习跨模态交互，其中不同模态之间的分布不匹配被很好地桥接。为此，边缘分布和具有高置信度相关性的元素在查询的公共空间和从不同模式计算的关键向量上对齐。在多模态视频理解的三个标准基准上的实验清楚地验证了我们方法的优越性。

Recently, the generalization behavior of Convolutional Neural Networks (CNN) is gradually transparent through explanation techniques with the frequency components decomposition. However, the importance of the phase spectrum of the image for a robust vision system is still ignored. In this paper, we notice that the CNN tends to converge at the local optimum which is closely related to the high-frequency components of the training images, while the amplitude spectrum is easily disturbed such as noises or common corruptions. In contrast, more empirical studies found that humans rely on more phase components to achieve robust recognition. This observation leads to more explanations of the CNN's generalization behaviors in both robustness to common perturbations and out-of-distribution detection, and motivates a new perspective on data augmentation designed by re-combing the phase spectrum of the current image and the amplitude spectrum of the distracter image. That is, the generated samples force the CNN to pay more attention to the structured information from phase components and keep robust to the variation of the amplitude. Experiments on several image datasets indicate that the proposed method achieves state-of-the-art performances on multiple generalizations and calibration tasks, including adaptability for common corruptions and surface variations, out-of-distribution detection, and adversarial attack.

最近，卷积神经网络 (CNN) 的泛化行为通过频率分量分解的解释技术逐渐变得透明。然而，图像相位谱对于鲁棒视觉系统的重要性仍然被忽视。在本文中，我们注意到CNN趋向于收敛于局部最优，这与训练图像的高频成分密切相关，而振幅谱容易受到干扰，例如噪声或常见的损坏。相比之下，更多的实证研究发现，人类依赖更多的相位成分来实现稳健的识别。这一观察结果进一步解释了CNN在对常见扰动和分布外检测的鲁棒性方面的泛化行为，并激发了通过重新组合当前图像的相位谱和干扰图像的振幅谱来设计数据增强的新视角。也就是说，生成的样本迫使CNN更加关注来自相位分量的结构化信息，并对振幅的变化保持鲁棒性。在多个图像数据集上的实验表明，该方法在多个泛化和校准任务上达到了最先进的性能，包括对常见腐蚀和表面变化的适应性、分布外检测和对抗性攻击。

Human can easily recognize visual objects with lost information: even losing most details with only contour reserved, e.g. cartoon. However, in terms of visual perception of Deep Neural Networks (DNNs), the ability for recognizing abstract objects (visual objects with lost information) is still a challenge. In this work, we investigate this issue from an adversarial viewpoint: will the performance of DNNs decrease even for the images only losing a little information? Towards this end, we propose a novel adversarial attack, named AdvDrop, which crafts adversarial examples by dropping existing information of images. Previously, most adversarial attacks add extra disturbing information on clean images explicitly. Opposite to previous works, our proposed work explores the adversarial robustness of DNN models in a novel perspective by dropping imperceptible details to craft adversarial examples. We demonstrate the effectiveness of AdvDrop by extensive experiments, and show that this new type of adversarial examples is more difficult to be defended by current defense systems.

人类可以很容易地识别丢失信息的视觉对象：即使丢失了大部分细节，也只保留了轮廓，例如卡通。然而，就深度神经网络（DNN）的视觉感知而言，识别抽象对象（具有丢失信息的视觉对象）的能力仍然是一个挑战。在这项工作中，我们从一个对立的角度来研究这个问题：即使图像只丢失很少的信息，DNN的性能是否会降低？为此，我们提出了一种新的对抗性攻击，名为AdvDrop，它通过删除图像的现有信息来制作对抗性示例。以前，大多数敌对攻击都会在干净的图像上添加额外的干扰信息。与以前的工作相反，我们提出的工作从一个新的角度探讨了DNN模型的对抗性健壮性，通过丢弃不易察觉的细节来制作对抗性示例。我们通过大量的实验证明了AdvDrop的有效性，并表明这种新型的对抗性示例更难被现有的防御系统防御。

Convolution on 3D point clouds that generalized from 2D grid-like domains is widely researched yet far from perfect. The standard convolution characterises feature correspondences indistinguishably among 3D points, presenting an intrinsic limitation of poor distinctive feature learning. In this paper, we propose Adaptive Graph Convolution (AdaptConv) which generates adaptive kernels for points according to their dynamically learned features. Compared with using a fixed/isotropic kernel, AdaptConv improves the flexibility of point cloud convolutions, effectively and precisely capturing the diverse relations between points from different semantic parts. Unlike popular attentional weight schemes, the proposed AdaptConv implements the adaptiveness inside the convolution operation instead of simply assigning different weights to the neighboring points. Extensive qualitative and quantitative evaluations show that our method outperforms state-of-the-art point cloud classification and segmentation approaches on several benchmark datasets. Our code is available at <https://github.com/hrzhou2/AdaptConv-master>.

从二维网格状区域推广的三维点云卷积已经得到了广泛的研究，但还远远不够完善。标准卷积在3D点之间以特征对应为特征，表现出较差的特征学习的内在局限性。在本文中，我们提出了自适应图卷积（AdaptConv），它根据点的动态学习特征生成自适应核。与使用固定/各向同性核相比，AdaptConv提高了点云卷积的灵活性，有效且精确地捕获了不同语义部分点之间的不同关系。与流行的注意权重方案不同，本文提出的AdaptConv在卷积运算中实现了自适应性，而不是简单地给相邻点分配不同的权重。大量的定性和定量评估表明，在几个基准数据集上，我们的方法优于最先进的点云分类和分割方法。我们的代码可在<https://github.com/hrzhou2/AdaptConv-master>.

Recently, Cross-Modal Hamming space Retrieval (CMHR) regains ever-increasing attention, mainly benefiting from the excellent representation capability of deep neural networks. On the other hand, the vulnerability of deep networks exposes a deep cross-modal retrieval system to various safety risks (e.g., adversarial attack). However, attacking deep cross-modal Hamming retrieval remains underexplored. In this paper, we propose an effective Adversarial Attack on Deep Cross-Modal Hamming Retrieval, dubbed AACH, which fools a target deep CMHR model in a black-box setting. Specifically, given a target model, we first construct its substitute model to exploit cross-modal correlations within hamming space, with which we create adversarial examples by limitedly querying from a target model. Furthermore, to enhance the efficiency of adversarial attacks, we design a triplet construction module to exploit cross-modal positive and negative instances. In this way, perturbations can be learned to fool the target model through pulling perturbed examples far away from the positive instances whereas pushing them close to the negative ones. Extensive experiments on three widely used cross-modal (image and text) retrieval benchmarks demonstrate the superiority of the proposed AACH. We find that AACH can successfully attack a given target deep CMHR model with fewer interactions, and that its performance is on par with previous state-of-the-art attacks.

近年来，跨模态汉明空间检索（CMHR）重新受到人们越来越多的关注，这主要得益于深层神经网络的良好表示能力。另一方面，深层网络的脆弱性使深层跨模式检索系统面临各种安全风险（例如，对抗性攻击）。然而，攻击深度跨模态汉明检索仍然没有得到充分的研究。在本文中，我们提出了一种针对深度交叉模式汉明检索的有效对抗攻击，称为AACH，它在黑盒环境中愚弄目标深度CMHR模型。具体地说，给定一个目标模型，我们首先构造它的替代模型来利用hamming空间中的交叉模态相关性，通过有限地查询目标模型来创建对抗性示例。此外，为了提高对抗性攻击的效率，我们设计了一个三元组构造模块来利用交叉模态的正、负实例。通过这种方式，可以学习扰动，通过将扰动示例拉离正实例，而将其推离负实例，从而愚弄目标模型。在三个广泛使用的跨模式（图像和文本）检索基准上的大量实验证明了所提出的AACH的优越性。我们发现，阿赫可以成功地攻击给定的目标深CMHR模型与较少的相互作用，其性能是等同于以前的最先进的攻击。

Recent studies have shown that cascade cost volume can play a vital role in deep stereo matching to achieve high resolution depth map with efficient hardware usage. However, how to construct good cascade volume as well as effective sampling for them are still under in-depth study. Previous cascade-based methods usually perform uniform sampling in a predicted disparity range based on variance, which easily misses the ground truth disparity and decreases disparity map accuracy. In this paper, we propose an uncertainty adaptive sampling network (UASNet) featuring two modules: an uncertainty distribution-guided range prediction (URP) model and an uncertainty-based disparity sampler (UDS) module. The URP explores the more discriminative uncertainty distribution to handle the complex matching ambiguities and to improve disparity range prediction. The UDS adaptively adjusts sampling interval to localize disparity with improved accuracy. With the proposed modules, our UASNet learns to construct cascade cost volume and predict full-resolution disparity map directly. Extensive experiments show that the proposed method achieves the highest ground truth covering ratio compared with other cascade cost volume based stereo matching methods. Our method also achieves top performance on both SceneFlow dataset and KITTI benchmark.

最近的研究表明，级联代价体积可以在深度立体匹配中发挥重要作用，从而在高效使用硬件的情况下实现高分辨率深度贴图。然而，如何构造良好的叶栅体积以及对其进行有效采样仍在深入研究中。以前基于级联的方法通常在基于方差的预测视差范围内执行均匀采样，这容易遗漏地面真实视差并降低视差图的精度。在本文中，我们提出了一种不确定性自适应采样网络（UASNet），它具有两个模块：不确定性分布引导距离预测（URP）模型和基于不确定性的视差采样器（UDS）模块。URP探索了更具辨别力的不确定性分布，以处理复杂的匹配歧义并改进视差范围预测。UDS自适应调整采样间隔，以提高精度定位视差。利用所提出的模块，我们的UASNet学习构造级联代价体积并直接预测全分辨率视差图。大量实验表明，与其他基于级联代价体的立体匹配方法相比，该方法获得了最高的地面真值覆盖率。我们的方法在SceneFlow数据集和KITTI基准上都达到了最佳性能。

With recent developments of convolutional neural networks, deep learning for 3D point clouds has shown significant progress in various 3D scene understanding tasks, e.g., object recognition, object detection. In a safety-critical environment, it is however not well understood how such deep learning models are vulnerable to adversarial examples. In this work, we explore adversarial attacks for point cloud-based neural networks. We propose a new formulation for adversarial point cloud generation that can generalise two different attack strategies. Our method generates adversarial examples by attacking the classification ability of point cloud-based networks while considering the perceptibility of the examples and ensuring the minimal level of point manipulations. Experimental results show that our method achieves the state-of-the-art performance with higher than 89% and 90% of attack success rate on synthetic and real-world data respectively, while manipulating only about 4% of the total points.

随着卷积神经网络的发展，三维点云的深度学习在各种三维场景理解任务中取得了重大进展，例如，目标识别、目标检测。然而，在一个安全关键的环境中，人们并不十分了解这种深度学习模型如何容易受到对抗性示例的影响。在这项工作中，我们探讨了基于点云的神经网络的对抗性攻击。我们提出了一种新的对抗点云生成公式，可以概括两种不同的攻击策略。我们的方法通过攻击基于点云的网络的分类能力生成对抗性示例，同时考虑示例的可感知性并确保最小程度的点操作。实验结果表明，该方法对合成数据和真实数据的攻击成功率分别达到89%和90%以上，而对总点数的操纵率仅为4%左右。

Existing vanishing point (VP) estimation methods rely on pre-extracted image lines and/or prior knowledge of the number of VPs. However, in practice, this information may be insufficient or unavailable. To solve this problem, we propose a network that treats a perspective image as input and predicts a spherical probability map of VP. Based on this map, we can detect all the VPs. Our method is reliable thanks to four technical novelties. First, we leverage the icosahedral spherical representation to express our probability map. This representation provides uniform pixel distribution, and thus facilitates estimating arbitrary positions of VPs. Second, we design a loss function that enforces the antipodal symmetry and sparsity of our spherical probability map to prevent over-fitting. Third, we generate the ground truth probability map that reasonably expresses the locations and uncertainties of VPs. This map unnecessarily peaks at noisy annotated VPs, and also exhibits various anisotropic dispersions. Fourth, given a predicted probability map, we detect VPs by fitting a Bingham mixture model. This strategy can robustly handle close VPs and provide the confidence level of VP useful for practical applications. Experiments showed that our method achieves the best compromise between generality, accuracy, and efficiency, compared with state-of-the-art approaches.

现有的消失点 (VP) 估计方法依赖于预提取的图像线和/或VP数的先验知识。然而，在实践中，这些信息可能不足或不可用。为了解决这个问题，我们提出了一种将透视图像作为输入并预测VP球形概率图的网络。基于此映射，我们可以检测所有VP。由于四项技术创新，我们的方法是可靠的。首先，我们利用二十面体球面表示来表示我们的概率图。这种表示提供了均匀的像素分布，因此便于估计VP的任意位置。其次，我们设计了一个损失函数，它加强了球面概率图的对极对称性和稀疏性，以防止过度拟合。第三，我们生成了地面真实概率图，合理地表达了VPs的位置和不确定性。该图不必要地在带噪注释的VP处达到峰值，并且还表现出各种各向异性色散。第四，给出一个预测概率图，我们通过拟合宾汉混合模型来检测VPs。该策略能够稳健地处理相近的VP，并提供对实际应用有用的VP置信度。实验表明，与现有的方法相比，我们的方法在通用性、准确性和效率之间取得了最佳的折衷。

Spiking Neural Networks (SNNs) have attracted enormous research interest due to temporal information processing capability, low power consumption, and high biological plausibility. However, the formulation of efficient and high-performance learning algorithms for SNNs is still challenging. Most existing learning methods learn weights only, and require manual tuning of the membrane-related parameters that determine the dynamics of a single spiking neuron. These parameters are typically chosen to be the same for all neurons, which limits the diversity of neurons and thus the expressiveness of the resulting SNNs. In this paper, we take inspiration from the observation that membrane-related parameters are different across brain regions, and propose a training algorithm that is capable of learning not only the synaptic weights but also the membrane time constants of SNNs. We show that incorporating learnable membrane time constants can make the network less sensitive to initial values and can speed up learning. In addition, we reevaluate the pooling methods in SNNs and find that max-pooling will not lead to significant information loss and have the advantage of low computation cost and binary compatibility. We evaluate the proposed method for image classification tasks on both traditional static MNIST, Fashion-MNIST, CIFAR-10 datasets, and neuromorphic N-MNIST, CIFAR10-DVS, DVS128 Gesture datasets. The experiment results show that the proposed method outperforms the state-of-the-art accuracy on nearly all datasets, using fewer time-steps. Our codes are available at <https://github.com/fangwei123456/Parametric-Leaky-Integrate-and-Fire-Spiking-Neuron>.

尖峰神经网络 (SNN) 由于具有时间信息处理能力、低功耗和高生物合理性而引起了人们极大的研究兴趣。然而，为SNN制定高效、高性能的学习算法仍然是一个挑战。大多数现有的学习方法只学习权值，并且需要手动调整膜相关参数，这些参数决定单个脉冲神经元的动力学。这些参数通常被选择为所有神经元的相同参数，这限制了神经元的多样性，从而限制了产生的SNN的表达能力。在本文中，我们从不同脑区膜相关参数的观察中得到启发，并提出了一种训练算法，该算法不仅能够学习SNN的突触权重，而且能够学习SNN的膜时间常数。我们发现，加入可学习的膜时间常数可以使网络对初始值不那么敏感，并且可以加快学习速度。此外，我们重新评估了SNN中的池方法，发现最大池不会导致显著的信息丢失，并且具有低计算成本和二进制兼容性的优势。我们在传统静态MNIST、时尚MNIST、CIFAR-10数据集和神经形态N-MNIST、CIFAR10-DVS、DVS128手势数据集上评估了所提出的图像分类任务方法。实验结果表明，该方法在几乎所有数据集上都优于最新的精度，使用的时间步长更少。我们的代码可在<https://github.com/fangwei123456/Parametric-Leaky-Integrate-and-Fire-Spiking-Neuron>。

We propose a novel approach that integrates under-parameterized RANSAC (UPRANSAC) with Hough Transform to detect vanishing points (VPs) from un-calibrated monocular images. In our algorithm, the UPRANSAC chooses one hypothetical inlier in a sample set to find a portion of the VP's degrees of freedom, which is followed by a highly reliable brute-force voting scheme (1-D Hough Transform) to find the VP's remaining degrees of freedom along the extension line of the hypothetical inlier. Our approach is able to sequentially find a series of VPs by repeatedly removing inliers of any detected VPs from minimal sample sets until the stop criterion is reached. Compared to traditional RANSAC that selects 2 edges as a hypothetical inlier pair to fit a model of VP hypothesis and requires hitting a pair of inliners, the UPRANSAC has a higher likelihood to hit one inliner and is more reliable in VP detection. Meanwhile, the tremendously scaled-down voting space with the requirement of only 1 parameter for processing significantly increased the performance efficiency of Hough Transform in our scheme. Testing results with well-known benchmark datasets show that the detection accuracies of our approach were higher or on par with the SOTA while running in deeply real-time zone.

我们提出了一种新的方法，结合欠参数化RANSAC (UPRANSAC) 和Hough变换从未校准的单眼图像中检测消失点 (VP)。在我们的算法中，UPRANSAC在一个样本集中选择一个假设的内插器来查找VP的一部分自由度，然后是一个高度可靠的蛮力投票方案 (1-D Hough变换)，以沿假设内插器的延长线查找VP的剩余自由度。我们的方法是通过从最小样本集中重复移除任何检测到的VP的内联，直到达到停止标准，从而能够顺序地找到一系列VP。与传统的RANSAC相比，UPRANSAC选择2条边作为假设的内联线对来拟合VP假设模型，并且需要命中一对内联线，UPRANSAC命中一条内联线的可能性更高，VP检测更可靠。同时，在我们的方案中，只需要1个处理参数的投票空间被极大地缩小，大大提高了Hough变换的性能效率。与已知的基准数据集的测试结果表明，我们的方法的检测精度较高或与SOTA相当，而运行在深度实时区。

Recently, the power of unconditional image synthesis has significantly advanced through the use of Generative Adversarial Networks (GANs). The task of inverting an image into its corresponding latent code of the trained GAN is of utmost importance as it allows for the manipulation of real images, leveraging the rich semantics learned by the network. Recognizing the limitations of current inversion approaches, in this work we present a novel inversion scheme that extends current encoder-based inversion methods by introducing an iterative refinement mechanism. Instead of directly predicting the latent code of a given real image using a single pass, the encoder is tasked with predicting a residual with respect to the current estimate of the inverted latent code in a self-correcting manner. Our residual-based encoder, named ReStyle, attains improved accuracy compared to current state-of-the-art encoder-based methods with a negligible increase in inference time. We analyze the behavior of ReStyle to gain valuable insights into its iterative nature. We then evaluate the performance of our residual encoder and analyze its robustness compared to optimization-based inversion and state-of-the-art encoders.

最近，通过使用生成性对抗网络 (GAN)，无条件图像合成的能力显著提高。将图像转化为训练后的GAN的相应潜在代码的任务至关重要，因为它允许操纵真实图像，利用网络学习的丰富语义。认识到当前反演方法的局限性，在这项工作中，我们提出了一种新的反演方案，通过引入迭代细化机制扩展了基于当前编码器的反演方法。编码器的任务不是使用单个过程直接预测给定真实图像的潜在码，而是以自校正方式预测关于反转潜在码的当前估计的残差。我们的基于残差的编码器名为ReStyle，与当前最先进的基于编码器的方法相比，其精度得到了提高，而推理时间的增加可以忽略不计。我们分析ReStyle的行为，以获得对其迭代性质的有价值的见解。然后，我们评估了残差编码器的性能，并与基于优化的反转和最先进的编码器相比，分析了其鲁棒性。

This paper focuses on the problem of low-rank tensor completion, the goal of which is to recover an underlying low-rank tensor from incomplete observations. Our method is motivated by the recently proposed t-product based on any invertible linear transforms. First, we define the new tensor average rank under the invertible real linear transforms. We then propose a new tensor completion model using a nonconvex surrogate to approximate the tensor average rank. This surrogate overcomes the discontinuity of the tensor average rank and alleviates the bias problem caused by the convex relaxation. Further, we develop an efficient algorithm to solve the proposed model and establish its convergence. Finally, experimental results on both synthetic and real data demonstrate the superiority of our method.

本文主要研究低秩张量完备问题，其目标是从不完全观测中恢复一个潜在的低秩张量。我们的方法是基于最近提出的基于任何可逆线性变换的t-积。首先，在可逆实线性变换下定义了新的张量平均秩。然后，我们提出了一个新的张量完成模型，使用非凸代理来近似张量平均秩。该替代项克服了张量平均秩的不连续性，并缓解了由凸松弛引起的偏差问题。此外，我们还开发了一种有效的算法来求解该模型并证明其收敛性。最后，在合成数据和真实数据上的实验结果证明了该方法的优越性。

Vision systems that deploy Deep Neural Networks (DNNs) are known to be vulnerable to adversarial examples. Recent research has shown that checking the intrinsic consistencies in the input data is a promising way to detect adversarial attacks (e.g., by checking the object co-occurrence relationships in complex scenes). However, existing approaches are tied to specific models and do not offer generalizability. Motivated by the observation that language descriptions of natural scene images have already captured the object co-occurrence relationships that can be learned by a language model, we develop a novel approach to perform context consistency checks using such language models. The distinguishing aspect of our approach is that it is independent of the deployed object detector and yet offers very high accuracy in terms of detecting adversarial examples in practical scenes with multiple objects. Experiments on the PASCAL VOC and MS COCO datasets show that our method can outperform state-of-the-art methods in detecting adversarial attacks.

众所周知，部署深度神经网络（DNN）的视觉系统容易受到对抗性示例的攻击。最近的研究表明，检查输入数据的内在一致性是检测对抗性攻击的一种很有前途的方法（例如，通过检查复杂场景中的对象共生关系）。然而，现有的方法与特定的模型相联系，不能提供通用性。基于自然场景图像的语言描述已经捕获了可以通过语言模型学习的对象共生关系，我们开发了一种新的方法来使用这种语言模型执行上下文一致性检查。我们的方法的独特之处在于，它独立于部署的对象检测器，但在实际场景中检测多个对象的对抗性示例时，它提供了非常高的准确性。在PASCAL VOC和MS COCO数据集上的实验表明，我们的方法在检测对抗性攻击方面优于最新的方法。

We present a novel semi-supervised semantic segmentation method which jointly achieves two desiderata of segmentation model regularities: the label-space consistency property between image augmentations and the feature-space contrastive property among different pixels. We leverage the pixel-level L2 loss and the pixel contrastive loss for the two purposes respectively. To address the computational efficiency issue and the false negative noise issue involved in the pixel contrastive loss, we further introduce and investigate several negative sampling techniques. Extensive experiments demonstrate the state-of-the-art performance of our method (PC2Seg) with the DeepLab-v3+ architecture, in several challenging semi-supervised settings derived from the VOC, Cityscapes, and COCO datasets.

我们提出了一种新的半监督语义分割方法，它共同实现了分割模型规则的两个要求：图像增强之间的标签空间一致性和不同像素之间的特征空间对比性。我们分别利用像素级L2损耗和像素对比损耗来实现这两个目的。为了解决像素对比损失中的计算效率问题和假负噪声问题，我们进一步介绍和研究了几种负采样技术。广泛的实验证明了我们的方法（PC2Seg）在DeepLab-v3+体系结构下的最先进性能，在来自VOC、Cityscapes和COCO数据集的几个具有挑战性的半监督设置中。

visual localization and mapping is the key technology underlying the majority of mixed reality and robotics systems. Most state-of-the-art approaches rely on local features to establish correspondences between images. In this paper, we present three novel scenarios for localization and mapping which require the continuous update of feature representations and the ability to match across different feature types. While localization and mapping is a fundamental computer vision problem, the traditional setup supposes the same local features are used throughout the evolution of a map. Thus, whenever the underlying features are changed, the whole process is repeated from scratch. However, this is typically impossible in practice, because raw images are often not stored and re-building the maps could lead to loss of the attached digital content. To overcome the limitations of current approaches, we present the first principled solution to cross-descriptor localization and mapping. Our data-driven approach is agnostic to the feature descriptor type, has low computational requirements, and scales linearly with the number of description algorithms. Extensive experiments demonstrate the effectiveness of our approach on state-of-the-art benchmarks for a variety of handcrafted and learned features.

视觉定位和映射是大多数混合现实和机器人系统的关键技术。大多数最先进的方法依赖于局部特征来建立图像之间的对应关系。在本文中，我们提出了三种新的定位和映射方案，它们需要不断更新特征表示，并能够跨不同的特征类型进行匹配。虽然定位和映射是一个基本的计算机视觉问题，但传统的设置假定在地图的整个演化过程中使用相同的局部特征。因此，只要底层特性发生变化，整个过程就会从头开始重复。然而，这在实践中通常是不可能的，因为原始图像通常不会存储，重新构建地图可能会导致附加数字内容的丢失。为了克服现有方法的局限性，我们提出了跨描述符定位和映射的第一个原则性解决方案。我们的数据驱动方法与特征描述符类型无关，计算要求低，并且与描述算法的数量成线性关系。大量的实验证明了我们的方法对各种手工制作和学习功能的最新基准测试的有效性。

This paper proposes Panoptic Narrative Grounding, a spatially fine and general formulation of the natural language visual grounding problem. We establish an experimental framework for the study of this new task, including new ground truth and metrics, and we propose a strong baseline method to serve as stepping stone for future work. We exploit the intrinsic semantic richness in an image by including panoptic categories, and we approach visual grounding at a fine-grained level by using segmentations. In terms of ground truth, we propose an algorithm to automatically transfer Localized Narratives annotations to specific regions in the panoptic segmentations of the MS COCO dataset. To guarantee the quality of our annotations, we take advantage of the semantic structure contained in WordNet to exclusively incorporate noun phrases that are grounded to a meaningfully related panoptic segmentation region. The proposed baseline achieves a performance of 55.4 absolute Average Recall points. This result is a suitable foundation to push the envelope further in the development of methods for Panoptic Narrative Grounding.

本文提出了全景叙事基础，这是自然语言视觉基础问题在空间上的一个精细和一般的表述。我们为这项新任务的研究建立了一个实验框架，包括新的基本事实和指标，并提出了一个强大的基线方法，作为未来工作的垫脚石。我们通过包含全景类别来利用图像中固有的语义丰富性，并通过分割在细粒度级别上实现视觉基础。在地面真实性方面，我们提出了一种算法来自动将局部叙述注释转移到MS COCO数据集全景分割中的特定区域。为了保证注释的质量，我们利用WordNet中包含的语义结构，专门包含基于有意义的相关全景切分区域的名词短语。建议的基线实现了55.4个绝对平均召回点的性能。这一结果是一个合适的基础，推动信封进一步发展的全景叙事接地的方法。

Anomaly detection with weakly supervised video-level labels is typically formulated as a multiple instance learning (MIL) problem, in which we aim to identify snippets containing abnormal events, with each video represented as a bag of video snippets. Although current methods show effective detection performance, their recognition of the positive instances, i.e., rare abnormal snippets in the abnormal videos, is largely biased by the dominant negative instances, especially when the abnormal events are subtle anomalies that exhibit only small differences compared with normal events. This issue is exacerbated in many methods that ignore important video temporal dependencies. To address this issue, we introduce a novel and theoretically sound method, named Robust Temporal Feature Magnitude learning (RTFM), which trains a feature magnitude learning function to effectively recognise the positive instances, substantially improving the robustness of the MIL approach to the negative instances from abnormal videos. RTFM also adapts dilated convolutions and self-attention mechanisms to capture long- and short-range temporal dependencies to learn the feature magnitude more faithfully. Extensive experiments show that the RTFM-enabled MIL model (i) outperforms several state-of-the-art methods by a large margin on four benchmark data sets (ShanghaiTech, UCF-Crime, XD-Violence and UCSD-Peds) and (ii) achieves significantly improved subtle anomaly discriminability and sample efficiency.

具有弱监督视频级别标签的异常检测通常被描述为多实例学习 (MIL) 问题，其中我们的目标是识别包含异常事件的片段，每个视频都表示为一包视频片段。尽管目前的方法显示出有效的检测性能，但它们对正实例（即异常视频中罕见的异常片段）的识别在很大程度上受到主要负实例的影响，特别是当异常事件是细微异常，与正常事件相比仅表现出微小差异时。这一问题在许多忽略重要视频时间依赖性的方法中更加严重。为了解决这个问题，我们引入了一种新的理论上合理的方法，称为鲁棒时域特征幅度学习 (RTFM)，该方法训练一个特征幅度学习函数来有效地识别正实例，大大提高了MIL方法对来自异常视频的负实例的鲁棒性。RTFM还采用扩展卷积和自我注意机制来捕获长程和短程时间依赖关系，以便更准确地了解特征量。大量实验表明，支持RTFM的MIL模型 (i) 在四个基准数据集（上海理工大学、UCF 犯罪、XD暴力和UCSD Peds）上大大优于几种最先进的方法，(ii) 显著提高了细微异常的可辨别性和样本效率。

Time-to-event analysis is an important statistical tool for allocating clinical resources such as ICU beds. However, classical techniques like the Cox model cannot directly incorporate images due to their high dimensionality. We propose a deep learning approach that naturally incorporates multiple, time-dependent imaging studies as well as non-imaging data into time-to-event analysis. Our techniques are benchmarked on a clinical dataset of 1,894 COVID-19 patients, and show that image sequences significantly improve predictions. For example, classical time-to-event methods produce a concordance error of around 30-40% for predicting hospital admission, while our error is 25% without images and 20% with multiple X-rays included. Ablation studies suggest that our models are not learning spurious features such as scanner artifacts. While our focus and evaluation is on COVID-19, the methods we develop are broadly applicable.

事件时间分析是分配ICU病床等临床资源的重要统计工具。然而，像考克斯模型这样的经典技术由于其高维性而不能直接合并图像。我们提出了一种深度学习方法，自然地将多个时间相关的成像研究以及非成像数据纳入到时间-事件分析中。2019冠状病毒疾病的临床数据集上的技术，并显示图像序列显著改善预测。例如，经典的事件时间方法在预测住院时产生的一致性误差约为30-40%，而我们的误差在没有图像的情况下为25%，在包括多张X光片的情况下为20%。消融研究表明，我们的模型并没有学到诸如扫描仪伪影之类的虚假特征。虽然我们的重点和评价是在COVID-19，我们开发的方法是广泛适用的。

Network quantization, which aims to reduce the bit-lengths of the network weights and activations, has emerged for their deployments to resource-limited devices. Although recent studies have successfully discretized a full-precision network, they still incur large quantization errors after training, thus giving rise to a significant performance gap between a full-precision network and its quantized counterpart. In this work, we propose a novel quantization method for neural networks, Cluster-Promoting Quantization (CPQ) that finds the optimal quantization grids while naturally encouraging the underlying full-precision weights to gather around those quantization grids cohesively during training. This property of CPQ is thanks to our two main ingredients that enable differentiable quantization: i) the use of the categorical distribution designed by a specific probabilistic parametrization in the forward pass and ii) our proposed multi-class straight-through estimator (STE) in the backward pass. Since our second component, multi-class STE, is intrinsically biased, we additionally propose a new bit-drop technique, DropBits, that revises the standard dropout regularization to randomly drop bits instead of neurons. As a natural extension of DropBits, we further introduce the way of learning heterogeneous quantization levels to find proper bit-length for each layer by imposing an additional regularization on DropBits. We experimentally validate our method on various benchmark datasets and network architectures, and also support a new hypothesis for quantization: learning heterogeneous quantization levels outperforms the case using the same but fixed quantization levels from scratch.

网络量化 (networkquantization) 旨在减少网络权重和激活的比特长度，已经出现在资源有限的设备上。尽管最近的研究已经成功地将全精度网络离散化，但它们在训练后仍会产生较大的量化误差，从而导致全精度网络与其量化网络之间存在显著的性能差距。在这项工作中，我们提出了一种新的神经网络量化方法，即聚类促进量化 (CPQ)，它可以找到最佳的量化网格，同时自然地鼓励基本的全精度权重在训练期间聚集在这些量化网格周围。CPQ的这一特性归功于我们实现可微量化的两个主要因素：i) 在前向传递中使用由特定概率参数化设计的分类分布，以及ii) 在后向传递中使用我们提出的多类直通估计器 (STE)。由于我们的第二个组件，多类STE，本质上是有偏差的，我们另外提出了一种新的位丢弃技术DropBits，它修改了标准的丢失正则化，以随机丢弃位而不是神经元。作为DropBits的自然扩展，我们进一步介绍了通过对DropBits施加额外的正则化来学习异构量化级别的方法，以便为每一层找到合适的比特长度。我们在各种基准数据集和网络架构上对我们的方法进行了实验验证，并支持一个新的量化假设：学习异构量化级别优于从头开始使用相同但固定量化级别的情况。

Class-incremental learning (CIL) aims at continuously updating a trained model with new classes (plasticity) without forgetting previously learned old ones (stability). Contemporary studies resort to storing representative exemplars for rehearsal or preventing consolidated model parameters from drifting, but the former requires an additional space for storing exemplars at every incremental phase while the latter usually shows poor model generalization. In this paper, we focus on resolving the stability-plasticity dilemma in class-incremental learning where no exemplars from old classes are stored. To make a trade-off between learning new information and maintaining old knowledge, we reformulate a simple yet effective baseline method based on a cosine classifier framework and reciprocal adaptive weights. With the reformulated baseline, we present two new approaches to CIL by learning class-independent knowledge and multi-perspective knowledge, respectively. The former exploits class-independent knowledge to bridge learning new and old classes, while the latter learns knowledge from different perspectives to facilitate CIL. Extensive experiments on several widely used CIL benchmark datasets show the superiority of our approaches over the state-of-the-art methods.

类增量学习 (CIL) 的目标是在不忘记以前学习过的旧类 (稳定性) 的情况下, 用新类 (可塑性) 不断更新训练模型。当代研究诉诸于存储有代表性的样本进行排练或防止合并模型参数漂移, 但前者在每个增量阶段都需要额外的空间来存储样本, 而后者通常显示出较差的模型泛化。在本文中, 我们致力于解决在没有存储旧类样本的情况下, 类增量学习中的稳定性-可塑性困境。为了在学习新信息和维护旧知识之间进行权衡, 我们基于余弦分类器框架和倒数自适应权重重新构造了一种简单而有效的基线方法。在重新制定的基线下, 我们分别通过学习班级独立知识和多视角知识, 提出了两种新的CIL方法。前者利用与班级无关的知识来沟通新旧班级的学习, 而后者则从不同的角度学习知识以促进CIL。在几个广泛使用的CIL基准数据集上进行的大量实验表明, 我们的方法优于最先进的方法。

This paper presents a neural network for robust normal estimation on point clouds, named AdaFit, that can deal with point clouds with noise and density variations. Existing works use a network to learn point-wise weights for weighted least squares surface fitting to estimate the normals, which has difficulty in finding accurate normals in complex regions or containing noisy points. By analyzing the step of weighted least squares surface fitting, we find that it is hard to determine the polynomial order of the fitting surface and the fitting surface is sensitive to outliers. To address these problems, we propose a simple yet effective solution that adds an additional offset prediction to improve the quality of normal estimation. Furthermore, in order to take advantage of points from different neighborhood sizes, a novel Cascaded Scale Aggregation layer is proposed to help the network predict more accurate point-wise offsets and weights. Extensive experiments demonstrate that AdaFit achieves state-of-the-art performance on both the synthetic PCPNet dataset and the real-word SceneNN dataset.

本文提出了一种用于点云稳健正态估计的神经网络AdaFit, 它可以处理具有噪声和密度变化的点云。现有工作使用网络学习加权最小二乘曲面拟合的逐点权重来估计法线, 这在复杂区域或包含噪声点的区域中很难找到准确的法线。通过分析加权最小二乘曲面拟合的步骤, 发现拟合曲面的多项式阶数难以确定, 且拟合曲面对异常值敏感。为了解决这些问题, 我们提出了一种简单而有效的解决方案, 即增加一个额外的偏移量预测, 以提高正态估计的质量。此外, 为了充分利用来自不同邻域大小的点, 提出了一种新的级联尺度聚集层, 以帮助网络预测更精确的点方向偏移量和权重。大量实验表明, AdaFit在合成PCPNet数据集和真实word SceneNN数据集上都达到了最先进的性能。

We present the first method capable of photorealistically reconstructing deformable scenes using photos/videos captured casually from mobile phones. Our approach augments neural radiance fields (NeRF) by optimizing an additional continuous volumetric deformation field that warps each observed point into a canonical 5D NeRF. We observe that these NeRF-like deformation fields are prone to local minima, and propose a coarse-to-fine optimization method for coordinate-based models that allows for more robust optimization. By adapting principles from geometry processing and physical simulation to NeRF-like models, we propose an elastic regularization of the deformation field that further improves robustness. We show that our method can turn casually captured selfie photos/videos into deformable NeRF models that allow for photorealistic renderings of the subject from arbitrary viewpoints, which we dub "nerfies." We evaluate our method by collecting time-synchronized data using a rig with two mobile phones, yielding train/validation images of the same pose at different viewpoints. We show that our method faithfully reconstructs non-rigidly deforming scenes and reproduces unseen views with high fidelity.

我们提出了第一种方法, 该方法能够使用从手机上随意拍摄的照片/视频真实地重建可变形场景。我们的方法通过优化附加的连续体积变形场来增强神经辐射场 (NeRF), 该变形场将每个观测点扭曲成规范的5D NeRF。我们观察到这些类NeRF变形场容易出现局部极小值, 并针对基于坐标的模型提出了一种从粗到精的优化方法, 以实现更稳健的优化。通过将几何处理和物理模拟的原理应用于类NeRF模型, 我们提

出了变形场的弹性正则化，进一步提高了鲁棒性。我们证明，我们的方法可以将随意拍摄的自拍照片/视频转换为可变形的NeRF模型，从而允许从任意视点对对象进行真实感渲染，我们称之为“nerfies”我们通过使用带有两部手机的试验台收集时间同步数据来评估我们的方法，生成不同视点下相同姿势的训练/验证图像。我们表明，我们的方法忠实地重建非刚性变形场景，并以高保真度再现看不见的视图。

we study the problem of concept induction in visual reasoning, i.e., identifying concepts and their hierarchical relationships from question-answer pairs associated with images; and achieve an interpretable model via working on the induced symbolic concept space. To this end, we first design a new framework named object-centric compositional attention model (OCCAM) to perform the visual reasoning task with object-level visual features. Then, we come up with a method to induce concepts of objects and relations using clues from the attention patterns between objects' visual features and question words. Finally, we achieve a higher level of interpretability by imposing OCCAM on the objects represented in the induced symbolic concept space. Experiments on the CLEVR and GQA datasets demonstrate: 1) our OCCAM achieves a new state of the art without human-annotated functional programs; 2) our induced concepts are both accurate and sufficient as OCCAM achieves an on-par performance on objects represented either in visual features or in the induced symbolic concept space.

我们研究视觉推理中的概念归纳问题，即从与图像相关的问答对中识别概念及其层次关系；通过对归纳符号概念空间的研究，实现了一个可解释的模型。为此，我们首先设计了一个名为“以对象为中心的合成注意模型”（OCCAM）的新框架来执行具有对象级视觉特征的视觉推理任务。然后，我们提出了一种从物体的视觉特征和疑问词之间的注意模式中提取线索来归纳物体和关系概念的方法。最后，我们通过对诱导符号概念空间中表示的对象施加OCCAM来实现更高级别的可解释性。在CLEVR和GQA数据集上的实验表明：1) 我们的OCCAM在没有人工注释的功能程序的情况下达到了最新的水平；2) 我们的诱导概念既准确又充分，因为OCCAM在视觉特征或诱导符号概念空间中表示的对象上实现了不相上下的性能。

Deep neural networks have significantly improved appearance-based gaze estimation accuracy. However, it still suffers from unsatisfactory performance when generalizing the trained model to new domains, e.g., unseen environments or persons. In this paper, we propose a plug-and-play gaze adaptation framework (PnP-GA), which is an ensemble of networks that learn collaboratively with the guidance of outliers. Since our proposed framework does not require ground-truth labels in the target domain, the existing gaze estimation networks can be directly plugged into PnP-GA and generalize the algorithms to new domains. We test PnP-GA on four gaze domain adaptation tasks, ETH-to-MPII, ETH-to-EyeDiap, Gaze360-to-MPII, and Gaze360-to-EyeDiap. The experimental results demonstrate that the PnP-GA framework achieves considerable performance improvements of 36.9%, 31.6%, 19.4%, and 11.8% over the baseline system. The proposed framework also outperforms the state-of-the-art domain adaptation approaches on gaze domain adaptation tasks.

深度神经网络显著提高了基于外观的注视估计精度。然而，当将训练后的模型推广到新的领域（例如，看不见的环境或人）时，它的性能仍然不令人满意。在本文中，我们提出了一个即插即用的注视适应框架（PnP-GA），它是一个在异常值指导下协作学习的网络集合。由于我们提出的框架在目标域不需要地面真值标签，现有的凝视估计网络可以直接插入到PnP GA中，并将算法推广到新的域。我们在四种注视域适应任务上测试了PnP GA，即ETH到MPII、ETH到EyeDiap、GAGE360到MPII和GAGE360到EyeDiap。实验结果表明，与基线系统相比，PnP-GA框架的性能分别提高了36.9%、31.6%、19.4%和11.8%。该框架在注视域适应任务上也优于最新的领域适应方法。

Editing an image automatically via a linguistic request can significantly save laborious manual work and is friendly to photography novice. In this paper, we focus on the task of language-guided global image editing. Existing works suffer from imbalanced data distribution of real-world datasets and thus fail to understand language requests well. To handle this issue, we propose to create a cycle with our image generator by creating another model called Editing Description Network (EDNet) which predicts an editing embedding given a pair of images. Given the cycle, we propose several free augmentation strategies to help our model understand various editing requests given the imbalanced dataset. In addition, two other novel ideas are proposed: an Image-Request Attention (IRA) module which allows our method to edit an image spatial-adaptively when the image requires different editing degree at different regions, as well as a new evaluation metric for this task which is more semantic and reasonable than conventional pixel losses (eg L1). Extensive experiments on two benchmark datasets demonstrate the effectiveness of our method over existing approaches.

通过语言请求自动编辑图像可以显著节省费力的手工工作，并且对摄影新手很友好。在本文中，我们重点研究语言引导的全局图像编辑任务。现有的工作受到现实世界数据集数据分布不平衡的影响，因此无法很好地理解语言请求。为了解决这个问题，我们建议通过创建另一个称为编辑描述网络（Editing Description Network, EDNet）的模型来使用图像生成器创建一个循环，该模型预测给定一对图像的编辑嵌入。鉴于这个周期，我们提出了几种免费的扩充策略，以帮助我们的模型理解给定不平衡数据集的各种编辑请求。此外，还提出了另外两个新的想法：一个图像请求注意（IRA）模块，该模块允许我们的方法在图像在不同区域需要不同编辑程度时自适应地编辑图像空间；以及一个新的评估指标，该指标比传统的像素损失（如L1）更具语义性和合理性。在两个基准数据集上的大量实验证明了我们的方法比现有方法的有效性。

The field of face recognition (FR) has witnessed remarkable progress with the surge of deep learning. The effective loss functions play an important role for FR. In this paper, we observe that a majority of loss functions, including the widespread triplet loss and softmax-based cross-entropy loss, embed inter-class (negative) similarity  $s_n$  and intra-class (positive) similarity  $s_p$  into similarity pairs and optimize to reduce ( $s_n - s_p$ ) in the training process. However, in the verification process, existing metrics directly take the absolute similarity between two features as the confidence of belonging to the same identity, which inevitably causes a gap between the training and verification process. To bridge the gap, we propose a new metric called Discrepancy Alignment Metric (DAM) for verification, which introduces the Local Inter-class Discrepancy (LID) for each face image to normalize the absolute similarity score. To estimate the LID of each face image in the verification process, we propose two types of LID Estimation (LIDE) methods, which are reference-based and learning-based estimation methods, respectively. The proposed DAM is plug-and-play and can be easily applied to the most existing methods. Extensive experiments on multiple popular face recognition benchmark datasets demonstrate the effectiveness of our proposed method.

随着深度学习的兴起，人脸识别领域取得了显著的进展。有效损失函数对FR起着重要作用。在本文中，我们观察到大多数损失函数，包括广泛存在的三重态损失和基于softmax的交叉熵损失，将类间（负）相似度 $s_n$ 和类内（正）相似度 $s_p$ 嵌入到相似度对中，并在训练过程中进行优化以减少( $s_n-s_p$ )。然而，在验证过程中，现有的度量直接将两个特征之间的绝对相似性作为属于同一身份的置信度，这不可避免地导致了训练和验证过程之间的差距。为了弥补这一差距，我们提出了一种新的验证度量，称为差异对齐度量（DAM），该度量为每个人脸图像引入局部类间差异（LID），以规范化绝对相似性分数。为了在验证过程中估计每个人脸图像的LID，我们提出了两种LID估计（LIDE）方法，分别是基于参考的和基于学习的估计方法。拟建大坝即插即用，可轻松应用于大多数现有方法。在多个流行的人脸识别基准数据集上的大量实验证明了该方法的有效性。

This work proposes a novel Deep Neural Network (DNN) quantization framework, namely RMSMP, with a \underline R ow-wise \underline M ixed-\underline S cheme and \underline M ulti-\underline P recision approach. Specifically, this is the first effort to assign mixed quantization schemes and multiple precisions within layers -- among rows of the DNN weight matrix, for simplified operations in hardware inference, while preserving accuracy. Furthermore, this paper makes a different observation from the prior work that the quantization error does not necessarily exhibit the layer-wise sensitivity, and actually can be mitigated as long as a certain portion of the weights in every layer are in higher precisions. This observation enables layer-wise uniformality in the hardware implementation towards guaranteed inference acceleration, while still enjoying row-wise flexibility of mixed schemes and multiple precisions to boost accuracy. The candidates of schemes and precisions are derived practically and effectively with a highly hardware-informative strategy to reduce the problem search space. With the offline determined ratio of different quantization schemes and precisions for all the layers, the RMSMP quantization algorithm uses Hessian and variance based method to effectively assign schemes and precisions for each row. The proposed RMSMP is tested for the image classification and natural language processing (BERT) applications, and achieves the best accuracy performance among state-of-the-arts under the same equivalent precisions. The RMSMP is implemented on FPGA devices, achieving 3.65x speedup in the end-to-end inference time for ResNet-18 on ImageNet, comparing with the 4-bit Fixed-point baseline.

本工作提出了一种新的深度神经网络 (DNN) 量化框架, 即RMSMP, 它采用了一种\underline Row wise \underline M ixed-\underline S模式和\underline M-\underline精度方法。具体地说, 这是第一次在层内 (DNN权重矩阵的行之间) 分配混合量化方案和多精度, 以简化硬件推断中的操作, 同时保持准确性。此外, 本文与以前的工作不同的是, 量化误差不一定表现出分层灵敏度, 并且只要每一层中的某一部分权重具有更高的精度, 实际上可以减轻量化误差。这一观察结果使得硬件实现中的分层一致性能够保证推理加速, 同时仍然享受混合方案的行级灵活性和提高精度的多重精度。采用高度硬件信息策略, 以减少问题搜索空间, 切实有效地推导出候选方案和精度。通过离线确定所有层的不同量化方案和精度的比率, RMSMP量化算法使用Hessian和基于方差的方法有效地为每行分配方案和精度。所提出的RMSMP在图像分类和自然语言处理 (BERT) 应用中进行了测试, 在同等精度下达到了最先进的精度性能。RMSMP在FPGA设备上实现, 与4位定点基线相比, ImageNet上ResNet-18的端到端推断时间达到3.65x speedup。

Although the visual appearances of small-scale objects are not well observed, humans can recognize them by associating the visual cues of small objects from their memorized appearance. It is called cued recall. In this paper, motivated by the memory process of humans, we introduce a novel pedestrian detection framework that imitates cued recall in detecting small-scale pedestrians. We propose a large-scale embedding learning with the large-scale pedestrian recalling memory (LPR Memory). The purpose of the proposed large-scale embedding learning is to memorize and recall the large-scale pedestrian appearance via the LPR Memory. To this end, we employ the large-scale pedestrian exemplar set, so that, the LPR Memory can recall the information of the large-scale pedestrians from the small-scale pedestrians. Comprehensive quantitative and qualitative experimental results validate the effectiveness of the proposed framework with the LPR Memory.

虽然小规模物体的视觉外观没有被很好地观察到, 但人类可以通过将小物体的视觉线索与记忆的外观联系起来来识别它们。这被称为线索回忆。在本文中, 受人类记忆过程的启发, 我们引入了一种新的行人检测框架, 该框架模仿线索回忆来检测小规模的行人。我们提出了一种基于大规模行人回忆记忆 (LPR 记忆) 的大规模嵌入学习算法。大规模嵌入学习的目的是通过LPR记忆对大规模行人外观进行记忆和回

忆。为此，我们采用大规模行人样本集，使得LPR存储器能够从小规模行人中召回大规模行人的信息。综合定量和定性实验结果验证了该框架在LPR存储器中的有效性。

Survival outcome prediction is a challenging weakly-supervised and ordinal regression task in computational pathology that involves modeling complex interactions within the tumor microenvironment in gigapixel whole slide images (WSIs). Despite recent progress in formulating WSIs as bags for multiple instance learning (MIL), representation learning of entire WSIs remains an open and challenging problem, especially in overcoming: 1) the computational complexity of feature aggregation in large bags, and 2) the data heterogeneity gap in incorporating biological priors such as genomic measurements. In this work, we present a Multimodal Co-Attention Transformer (MCAT) framework that learns an interpretable, dense co-attention mapping between WSIs and genomic features formulated in an embedding space. Inspired by approaches in Visual Question Answering (VQA) that can attribute how word embeddings attend to salient objects in an image when answering a question, MCAT learns how histology patches attend to genes when predicting patient survival. In addition to visualizing multimodal interactions, our co-attention transformation also reduces the space complexity of WSI bags, which enables the adaptation of Transformer layers as a general encoder backbone in MIL. We apply our proposed method on five different cancer datasets (4,730 WSIs, 67 million patches). Our experimental results demonstrate that the proposed method consistently achieves superior performance compared to the state-of-the-art methods.

在计算病理学中，生存结果预测是一项具有挑战性的弱监督有序回归任务，它涉及在千兆像素整张幻灯片图像（WSI）中模拟肿瘤微环境中的复杂相互作用。尽管最近在将WSIs描述为用于多实例学习（MIL）的包方面取得了进展，但整个WSIs的表示学习仍然是一个开放且具有挑战性的问题，特别是在克服以下方面：1) 大型包中特征聚合的计算复杂性，2) 在纳入生物先验知识（如基因组测量）方面的数据异质性差距。在这项工作中，我们提出了一个多模态共同注意转换器（MCAT）框架，该框架学习WSIs和嵌入空间中形成的基因组特征之间的可解释、密集的共同注意映射。受视觉问答（VQA）方法的启发，MCAT可以在回答问题时确定单词嵌入如何处理图像中的突出对象，它学习了组织学补丁在预测患者生存时如何处理基因。除了可视化多模态交互，我们的共同注意转换还降低了WSI包的空间复杂性，这使得变压器层能够适应MIL中的通用编码器主干。我们将我们提出的方法应用于五个不同的癌症数据集（4730个WSI，6700万个补丁）。我们的实验结果表明，与目前最先进的方法相比，本文提出的方法始终取得了优异的性能。

Model quantization has emerged as a mandatory technique for efficient inference with advanced Deep Neural Networks (DNN). It converts the model parameters in full precision (32-bit floating point) to the hardware friendly data representation with shorter bit-width, to not only reduce the model size but also simplify the computation complexity. Nevertheless, prior model quantization either suffers from the inefficient data encoding method thus leading to noncompetitive model compression rate, or requires time-consuming quantization aware training process. In this work, we propose a novel Adaptive Floating-Point (AFP) as a variant of standard IEEE-754 floating-point format, with flexible configuration of exponent and mantissa segments. Leveraging the AFP for model quantization (i.e., encoding the parameter) could significantly enhance the model compression rate without accuracy degradation and model re-training. We also want to highlight that our proposed AFP could effectively eliminate the computationally intensive de-quantization step existing in the dynamic quantization technique adopted by the famous machine learning frameworks (e.g., pytorch, tensorRT and etc). Moreover, we develop a framework to automatically optimize and choose the adequate AFP configuration for each layer, thus maximizing the compression efficacy. Our experiments indicate that AFP-encoded ResNet-50/MobileNet-v2 only has ~0.04/0.6% accuracy degradation w.r.t its full-precision counterpart. It outperforms the state-of-the-art works by 1.1% in accuracy using the same bit-width while reducing the energy consumption by 11.2x, which is quite impressive for inference.

模型量化已成为使用高级深度神经网络 (DNN) 进行有效推理的必要技术。它将全精度 (32位浮点) 的模型参数转换为具有较短位宽的硬件友好数据表示形式，不仅减小了模型尺寸，而且简化了计算复杂度。然而，先前的模型量化要么受到低效的数据编码方法的影响，从而导致非竞争性的模型压缩率，要么需要耗时的量化感知训练过程。在这项工作中，我们提出了一种新的自适应浮点 (AFP) 作为标准 IEEE-754浮点格式的变体，具有指数和尾数段的灵活配置。利用AFP进行模型量化（即，编码参数）可以显著提高模型压缩率，而不会降低精度和重新训练模型。我们还想强调的是，我们提出的AFP可以有效地消除著名机器学习框架（如Pytork、tensorRT等）采用的动态量化技术中存在的计算密集型去量化步骤。此外，我们开发了一个框架来自动优化和选择每一层的适当AFP配置，从而最大限度地提高压缩效率。我们的实验表明， AFP编码的ResNet-50/MobileNet-v2与全精度对应物相比，精度降低了约 0.04/0.6%。在使用相同的位宽度时，它的精度比最先进的作品高出1.1%，同时将能耗降低了11.2倍，这对于推理来说是非常令人印象深刻的。

Humans usually explain their reasoning (e.g. classification) by dissecting the image and pointing out the evidence from these parts to the concepts in their minds. Inspired by this cognitive process, several part-level interpretable neural network architectures have been proposed to explain the predictions. However, they suffer from the complex data structure and confusing the effect of the individual part to output category. In this work, an interpretable image recognition deep network is designed by introducing a plug-in transparent embedding space (TesNet) to bridge the high-level input patches (e.g. CNN feature maps) and the output categories. This plug-in embedding space is spanned by transparent basis concepts which are constructed on the Grassmann manifold. These basis concepts are enforced to be category-aware and within-category concepts are orthogonal to each other, which makes sure the embedding space is disentangled. Meanwhile, each basis concept can be traced back to the particular image patches, thus they are transparent and friendly to explain the reasoning process. By comparing with state-of-the-art interpretable methods, TesNet is much more beneficial to classification tasks, esp. providing better interpretability on predictions and improve the final accuracy.

人类通常通过解剖图像并指出从这些部分到他们头脑中概念的证据来解释他们的推理（例如分类）。受这一认知过程的启发，人们提出了几种部分级可解释神经网络结构来解释预测。然而，它们受到复杂数据结构的影响，并混淆了单个部件对输出类别的影响。在这项工作中，通过引入一个插件透明嵌入空间（TesNet）来连接高级输入块（如CNN特征图）和输出类别，设计了一个可解释的图像识别深度网络。这个插件嵌入空间由在格拉斯曼流形上构造的透明基础概念所跨越。这些基本概念被强制为具有类别意识，并且类别内的概念彼此正交，这确保了嵌入空间是分离的。同时，每个基本概念都可以追溯到特定的图像块，因此它们透明且友好地解释了推理过程。与最先进的可解释方法相比，TesNet对分类任务更为有利，尤其是提供更好的预测可解释性和提高最终精度。

Recently, deep learning based point cloud descriptors have achieved impressive results in the place recognition task. Nonetheless, due to the sparsity of point clouds, how to extract discriminative local features of point clouds to efficiently form a global descriptor is still a challenging problem. In this paper, we propose a pyramid point cloud transformer network (PPT-Net) to learn the discriminative global descriptors from point clouds for efficient retrieval. Specifically, we first develop a pyramid point transformer module that adaptively learns the spatial relationship of the different local k-NN graphs of point clouds, where the grouped self-attention is proposed to extract discriminative local features of the point clouds. Furthermore, the grouped self-attention not only enhances long-term dependencies of the point clouds, but also reduces the computational cost. In order to obtain discriminative global descriptors, we construct a pyramid VLAD module to aggregate the multi-scale feature maps of point clouds into the global descriptors. By applying VLAD pooling on multi-scale feature maps, we utilize the context gating mechanism on the multiple global descriptors to adaptively weight the multi-scale global context information into the final global descriptor. Experimental results on the Oxford dataset and three in-house datasets show that our method achieves the state-of-the-art on the point cloud based place recognition task. Code is available at <https://github.com/fpthink/PPT-Net>.

最近，基于深度学习的点云描述符在位置识别任务中取得了令人印象深刻的结果。然而，由于点云的稀疏性，如何提取点云的局部特征以有效地形成全局描述符仍然是一个具有挑战性的问题。在本文中，我们提出了一个金字塔点云变换网络（PPT网络），从点云中学习有区别的全局描述符，以实现高效检索。具体来说，我们首先开发了一个金字塔点变换模块，该模块自适应地学习点云的不同局部k-NN图的空间关系，其中提出了分组自我注意来提取点云的区分性局部特征。此外，分组自我注意不仅增强了点云的长期依赖性，而且降低了计算成本。为了获得有区别的全局描述符，我们构造了一个金字塔VLAD模块，将点云的多尺度特征映射聚合到全局描述符中。通过在多尺度特征映射上应用VLAD池，我们利用多个全局描述符上的上下文选通机制，自适应地将多尺度全局上下文信息加权到最终的全局描述符中。在Oxford数据集和三个内部数据集上的实验结果表明，我们的方法实现了基于点云的位置识别任务。代码可在<https://github.com/fpthink/PPT-Net>。

This paper aims to explain adversarial attacks in terms of how adversarial perturbations contribute to the attacking task. We estimate attributions of different image regions to the decrease of the attacking cost based on the Shapley value. We define and quantify interactions among adversarial perturbation pixels, and decompose the entire perturbation map into relatively independent perturbation components. The decomposition of the perturbation map shows that adversarially-trained DNNs have more perturbation components in the foreground than normally-trained DNNs. Moreover, compared to the normally-trained DNN, the adversarially-trained DNN have more components which mainly decrease the score of the true category. Above analyses provide new insights into the understanding of adversarial attacks.

本文旨在从对抗性干扰如何影响攻击任务的角度来解释对抗性攻击。我们根据Shapley值估计不同图像区域的属性以降低攻击成本。我们定义并量化对抗扰动像素之间的相互作用，并将整个扰动图分解为相对独立的扰动分量。对扰动图的分解表明，对抗训练的DNN比正常训练的DNN在前景中具有更多的扰动分量。此外，与正常训练的DNN相比，对抗训练的DNN有更多的成分，这主要降低了真实类别的得分。上述分析为理解对抗性攻击提供了新的见解。

We propose a novel method that leverages human fixations to visually decode the image a person has in mind into a photofit (facial composite). Our method combines three neural networks: An encoder, a scoring network, and a decoder. The encoder extracts image features and predicts a neural activation map for each face looked at by a human observer. A neural scoring network compares the human and neural attention and predicts a relevance score for each extracted image feature. Finally, image features are aggregated into a single feature vector as a linear combination of all features weighted by relevance which a decoder decodes into the final photofit. We train the neural scoring network on a novel dataset containing gaze data of 19 participants looking at collages of synthetic faces. We show that our method significantly outperforms a mean baseline predictor and report on a human study that shows that we can decode photofits that are visually plausible and close to the observer's mental image.

我们提出了一种新的方法，利用人类的注视，将一个人心目中的图像视觉解码为一个photofit（面部合成）。我们的方法结合了三个神经网络：编码器、评分网络和解码器。编码器提取图像特征并预测人类观察者看到的每个人脸的神经激活图。神经评分网络比较人类和神经注意，并预测每个提取的图像特征的相关评分。最后，图像特征被聚合成单个特征向量，作为所有特征的线性组合，通过相关性加权，解码器解码成最终的光拟合。我们在一个新的数据集上训练神经评分网络，该数据集包含19名观看合成人脸拼贴的参与者的凝视数据。我们表明，我们的方法明显优于平均基线预测值，并报告了一项人类研究，该研究表明，我们可以解码视觉上合理且接近观察者心理图像的照片拟合。

A lifespan face synthesis (LFS) model aims to generate a set of photo-realistic face images of a person's whole life, given only one snapshot as reference. The generated face image given a target age code is expected to be age-sensitive reflected by bio-plausible transformations of shape and texture, while being identity preserving. This is extremely challenging because the shape and texture characteristics of a face undergo separate and highly nonlinear transformations w.r.t. age. Most recent LFS models are based on generative adversarial networks (GANs) whereby age code conditional transformations are applied to a latent face representation. They benefit greatly from the recent advancements of GANs. However, without explicitly disentangling their latent representations into the texture, shape and identity factors, they are fundamentally limited in modeling the nonlinear age-related transformation on texture and shape whilst preserving identity. In this work, a novel LFS model is proposed to disentangle the key face characteristics including shape, texture and identity so that the unique shape and texture age transformations can be modeled effectively. This is achieved by extracting shape, texture and identity features separately from an encoder. Critically, two transformation modules, one conditional convolution based and the other channel attention based, are designed for modeling the nonlinear shape and texture feature transformations respectively. This is to accommodate their rather distinct aging processes and ensure that our synthesized images are both age-sensitive and identity preserving. Extensive experiments show that our LFS model is clearly superior to the state-of-the-art alternatives.

寿命人脸合成 (LFS) 模型旨在生成一组照片逼真的脸图像，这些图像只提供一张快照作为参考。给定目标年龄代码生成的人脸图像预计会通过形状和纹理的生物合理变换反映年龄敏感，同时保持身份。这是一项极具挑战性的工作，因为人脸的形状和纹理特征会随着年龄的增长而发生独立的、高度非线性的变化。最新的LFS模型基于生成性对抗网络 (GAN)，其中年龄代码条件转换应用于潜在人脸表示。他

们从GANs的最新发展中受益匪浅。然而，如果没有明确地将它们的潜在表征分解为纹理、形状和身份因素，它们在建模纹理和形状上的非线性年龄相关变换时基本上是有限的，同时保持身份。在这项工作中，提出了一种新的LFS模型来分离关键的人脸特征，包括形状、纹理和身份，从而有效地模拟独特的形状和纹理年龄变换。这是通过从编码器中分别提取形状、纹理和身份特征来实现的。关键的是，设计了两个变换模块，一个基于条件卷积，另一个基于通道注意，分别用于建模非线性形状和纹理特征变换。这是为了适应它们相当独特的老化过程，并确保我们的合成图像既对年龄敏感又保持身份。大量实验表明，我们的LFS模型明显优于最先进的替代方案。

Event cameras are novel sensors that perceive the per-pixel intensity changes and output asynchronous event streams with high dynamic range and less motion blur. It has been shown that events alone can be used for end-task learning, e.g., semantic segmentation, based on encoder-decoder-like networks. However, as events are sparse and mostly reflect edge information, it is difficult to recover original details merely relying on the decoder. Moreover, most methods resort to the pixel-wise loss alone for supervision, which might be insufficient to fully exploit the visual details from sparse events, thus leading to less optimal performance. In this paper, we propose a simple yet flexible two-stream framework named Dual Transfer Learning (DTL) to effectively enhance the performance on the end-tasks without adding extra inference cost. The proposed approach consists of three parts: event to end-task learning (EEL) branch, event to image translation (EIT) branch, and transfer learning (TL) module that simultaneously explores the feature-level affinity information and pixel-level knowledge from the EIT branch to improve the EEL branch. This simple yet novel method leads to strong representation learning from events and is evidenced by the significant performance boost on the end-tasks such as semantic segmentation and depth estimation.

事件摄影机是一种新型传感器，可感知每像素强度变化并输出具有高动态范围和较少运动模糊的异步事件流。已经证明，事件本身可以用于最终任务学习，例如，基于编码器-解码器类网络的语义分割。然而，由于事件稀疏且大多反映边缘信息，仅依靠解码器很难恢复原始细节。此外，大多数方法仅依靠像素损失进行监控，这可能不足以充分利用稀疏事件中的视觉细节，从而导致较差的最佳性能。在本文中，我们提出了一个简单而灵活的双流框架，称为双重迁移学习（DTL），在不增加额外推理成本的情况下有效地提高了终端任务的性能。该方法由三个部分组成：事件到结束任务学习（EEL）分支、事件到图像翻译（EIT）分支和迁移学习（TL）模块，该模块同时探索EIT分支中的特征级亲和力信息和像素级知识，以改进EEL分支。这种简单而新颖的方法可以从事件中获得很强的表示性学习，并在语义分割和深度估计等最终任务中获得显著的性能提升。

Deep neural networks achieve great success in many visual recognition tasks. However, the model deployment is usually subject to some computational resources. Model pruning under computational budget has attracted growing attention. In this paper, we focus on the discrimination-aware compression of Convolutional Neural Networks (CNNs). In prior arts, directly searching the optimal sub-network is an integer programming problem, which is non-smooth, non-convex, and NP-hard. Meanwhile, the heuristic pruning criterion lacks clear interpretability and doesn't generalize well in applications. To address this problem, we formulate sub-networks as samples from a multivariate Bernoulli distribution and resort to the approximation of continuous problem. We propose a new flexible search scheme via alternating exploration and estimation. In the exploration step, we employ stochastic gradient Hamiltonian Monte Carlo with budget-awareness to generate sub-networks, which allows large search space with efficient computation. In the estimation step, we deduce the sub-network sampler to a near-optimal point, to promote the generation of high-quality sub-networks. Unifying the exploration and estimation, our approach avoids early falling into local minimum via a fast gradient-based search in a larger space. Extensive experiments on CIFAR-10 and ImageNet show that our method achieves state-of-the-art performances on pruning several popular CNNs.

深度神经网络在许多视觉识别任务中取得了巨大的成功。然而，模型部署通常需要一些计算资源。计算预算下的模型修剪越来越受到人们的关注。本文主要研究卷积神经网络 (CNN) 的识别感知压缩。在现有技术中，直接搜索最优子网络是一个非光滑、非凸、NP难的整数规划问题。同时，启发式剪枝准则缺乏明确的解释性，在实际应用中推广效果不好。为了解决这个问题，我们从多元贝努利分布中构造子网络作为样本，并求助于连续问题的近似。我们提出了一个新的灵活的搜索方案，通过交替探索和估计。在探索步骤中，我们使用具有预算意识的随机梯度哈密顿蒙特卡罗生成子网络，这允许使用高效计算的大搜索空间。在估计步骤中，我们将子网络采样器推到接近最优点，以促进高质量子网络的生成。我们的方法通过在更大的空间中基于梯度的快速搜索，将搜索和估计统一起来，避免了早期陷入局部极小。在CIFAR-10和ImageNet上的大量实验表明，我们的方法在修剪几个流行的CNN时达到了最先进的性能。

Re-assembling multiple pots accurately from numerous 3D scanned fragments remains a challenging task to this date. Previous methods extract all potential matching pairs of pot sherds and consider them simultaneously to search for an optimal global pot configuration. In this work, we empirically show such global approach greatly suffers from false positive matches between sherds inflicted by indistinctive sharp fracture surfaces in pot fragments. To mitigate this problem, we take inspirations from the field of structure-from-motion (SfM), where many pipelines have matured in reconstructing a 3D scene from multiple images. Motivated by the success of the incremental approach in robust SfM, we present an efficient reassembly method for axially symmetric pots based on iterative registration of one sherd at a time. Our method goes beyond replicating incremental SfM and addresses indistinguishable false matches by embracing beam search to explore multitudes of registration possibilities. Additionally, we utilize multiple roots in each step to allow simultaneous reassembly of multiple pots. The proposed approach shows above 80% reassembly accuracy on a dataset of real 80 fragments mixed from 5 pots, pushing the state-of-the-art and paving the way towards the goal of large-scale pot reassembly. Our code and preprocessed data will be made available for research.

迄今为止，从大量3D扫描碎片中准确地重新组装多个POT仍然是一项具有挑战性的任务。以前的方法提取所有可能匹配的壶碎片对，并同时考虑它们以搜索最优的全局壶配置。在这项工作中，我们的经验表明，这种全局方法极大地受到由罐碎片中模糊的尖锐断裂面造成的碎片之间的假阳性匹配的影响。为了缓解这个问题，我们从运动结构 (SfM) 领域获得了灵感，在该领域，许多管道已经成熟，可以从多幅图像重建3D场景。基于增量方法在鲁棒SfM中的成功，我们提出了一种基于一次迭代配准一个碎片的轴

对称pots的有效重组方法。我们的方法超越了复制增量SfM，并通过采用波束搜索来探索多种注册可能性，从而解决了无法区分的错误匹配。此外，我们在每个步骤中使用多个根，以允许同时重新组装多个罐。所提出的方法在一个由5个罐中混合的80个片段组成的数据集上显示了80%以上的重组精度，推动了最先进的技术，为实现大规模罐重组的目标铺平了道路。我们的代码和预处理数据将用于研究。

In a typical multi-label setting, a picture contains on average few positive labels, and many negative ones. This positive-negative imbalance dominates the optimization process, and can lead to under-emphasizing gradients from positive labels during training, resulting in poor accuracy. In this paper, we introduce a novel asymmetric loss ("ASL"), which operates differently on positive and negative samples. The loss enables to dynamically down-weights and hard-thresholds easy negative samples, while also discarding possibly mislabeled samples. We demonstrate how ASL can balance the probabilities of different samples, and how this balancing is translated to better mAP scores. With ASL, we reach state-of-the-art results on multiple popular multi-label datasets: MS-COCO, Pascal-VOC, NUS-WIDE and Open Images. We also demonstrate ASL applicability for other tasks, such as single-label classification and object detection. ASL is effective, easy to implement, and does not increase the training time or complexity. Implementation is available at: <https://github.com/Alibaba-MIIL/ASL>.

在典型的多标签设置中，一张图片平均包含少数正面标签和许多负面标签。这种正-负不平衡主导了优化过程，并可能导致在训练过程中对正标签的梯度强调不足，导致精度低下。在本文中，我们介绍了一种新的不对称损失（“ASL”），它对正样本和负样本的操作不同。这种损失可以动态地降低权重和硬阈值，同时丢弃可能标记错误的样本。我们演示了ASL如何平衡不同样本的概率，以及如何将这种平衡转化为更好的mAP分数。使用ASL，我们可以在多个流行的多标签数据集上获得最先进的结果：MS-COCO、Pascal VOC、NUS-WIDE和开放图像。我们还演示了ASL对其他任务的适用性，如单标签分类和目标检测。ASL有效，易于实现，并且不会增加培训时间或复杂性。可在以下网址获得实施：<https://github.com/Alibaba-MIIL/ASL>。

Accurate video understanding involves reasoning about the relationships between actors, objects and their environment, often over long temporal intervals. In this paper, we propose a message passing graph neural network that explicitly models these spatio-temporal relations and can use explicit representations of objects, when supervision is available, and implicit representations otherwise. Our formulation generalises previous structured models for video understanding, and allows us to study how different design choices in graph structure and representation affect the model's performance. We demonstrate our method on two different tasks requiring relational reasoning in videos -- spatio-temporal action detection on AVA and UCF101-24, and video scene graph classification on the recent Action Genome dataset -- and achieve state-of-the-art results on all three datasets. Furthermore, we show quantitatively and qualitatively how our method is able to more effectively model relationships between relevant entities in the scene.

准确的视频理解涉及到对演员、对象及其环境之间关系的推理，通常是在很长的时间间隔内。在本文中，我们提出了一种消息传递图神经网络，它可以显式地建模这些时空关系，并且可以在监控可用时使用对象的显式表示，在其他情况下使用隐式表示。我们的公式概括了以前用于视频理解的结构化模型，并允许我们研究图形结构和表示中的不同设计选择如何影响模型的性能。我们在视频中需要关系推理的两个不同任务上演示了我们的方法——AVA和UCF101-24上的时空动作检测，以及最新动作基因组数据集上的视频场景图分类——并在所有三个数据集上实现了最先进的结果。此外，我们还从定量和定性两方面展示了我们的方法如何能够更有效地建模场景中相关实体之间的关系。

We propose three novel solvers for estimating the relative pose of a multi-camera system from affine correspondences (ACs). A new constraint is derived interpreting the relationship of ACs and the generalized camera model. Using the constraint, we demonstrate efficient solvers for two types of motions assumed. Considering that the cameras undergo planar motion, we propose a minimal solution using a single AC and a solver with two ACs to overcome the degenerate case. Also, we propose a minimal solution using two ACs with known vertical direction, e.g., from an IMU. Since the proposed methods require significantly fewer correspondences than state-of-the-art algorithms, they can be efficiently used within RANSAC for outlier removal and initial motion estimation. The solvers are tested both on synthetic data and on real-world scenes from the KITTI odometry benchmark. It is shown that the accuracy of the estimated poses is superior to the state-of-the-art techniques.

我们提出了三种新的解算器，用于从仿射对应（ACs）估计多摄像机系统的相对姿态。推导了一个新的约束条件，解释了ACs与广义摄像机模型之间的关系。使用该约束，我们演示了两种假设运动的有效解算器。考虑到摄像机经历平面运动，我们提出了一种使用单个AC和一个带有两个AC的解算器的最小解决来克服退化情况。此外，我们还提出了一个最小解决方案，使用两个垂直方向已知的ACs，例如来自IMU的ACs。由于所提出的方法比最先进的算法需要更少的对应关系，因此它们可以在RANSAC中有效地用于异常值去除和初始运动估计。解算器在KITTI里程计基准的合成数据和真实场景上进行测试。结果表明，估计姿势的精度优于最先进的技术。

We aim at improving the computational efficiency of graph convolutional networks (GCNs) for learning on point clouds. The basic graph convolution that is composed of a K-nearest neighbor (KNN) search and a multilayer perceptron (MLP) is examined. By mathematically analyzing the operations there, two findings to improve the efficiency of GCNs are obtained. (1) The local geometric structure information of 3D representations propagates smoothly across the GCN that relies on KNN search to gather neighborhood features. This motivates the simplification of multiple KNN searches in GCNs. (2) Shuffling the order of graph feature gathering and an MLP leads to equivalent or similar composite operations. Based on those findings, we optimize the computational procedure in GCNs. A series of experiments show that the optimized networks have reduced computational complexity, decreased memory consumption, and accelerated inference speed while maintaining comparable accuracy for learning on point clouds.

我们的目标是提高在点云上学习的图卷积网络（GCN）的计算效率。研究了由K近邻（KNN）搜索和多层次感知器（MLP）组成的基本图卷积。通过对这些操作的数学分析，得出了提高GCN效率的两个发现。

(1) 三维表示的局部几何结构信息通过依赖KNN搜索来收集邻域特征的GCN平滑传播。这推动了GCN中多个KNN搜索的简化。(2) 改变图形特征收集和MLP的顺序会导致等效或类似的复合操作。基于这些发现，我们优化了GCNs中的计算过程。一系列实验表明，优化后的网络降低了计算复杂度，减少了内存消耗，加快了推理速度，同时保持了在点云上学习的同等精度。

Model compression aims to deploy deep neural networks (DNN) on mobile devices with limited computing and storage resources. However, most of the existing model compression methods rely on manually defined rules, which require domain expertise. DNNs are essentially computational graphs, which contain rich structural information. In this paper, we aim to find a suitable compression policy from DNNs' structural information. We propose an automatic graph encoder-decoder model compression (AGMC) method combined with graph neural networks (GNN) and reinforcement learning (RL). We model the target DNN as a graph and use GNN to learn the DNN's embeddings automatically. We compared our method with rule-based DNN embedding model compression methods to show the effectiveness of our method. Results show that our learning-based DNN embedding achieves better performance and a higher compression ratio with fewer search steps. We evaluated our method on over-parameterized and mobile-friendly DNNs and compared our method with handcrafted and learning-based model compression approaches. On over parameterized DNNs, such as ResNet-56, our method outperformed handcrafted and learning-based methods with 4.36% and 2.56% higher accuracy, respectively. Furthermore, on MobileNet-v2, we achieved a higher compression ratio than state-of-the-art methods with just 0.93% accuracy loss.

模型压缩的目的是在计算和存储资源有限的移动设备上部署深度神经网络（DNN）。然而，大多数现有的模型压缩方法依赖于手动定义的规则，这需要领域的专业知识。DNN本质上是计算图，包含丰富的结构信息。在本文中，我们的目标是从DNN的结构信息中找到合适的压缩策略。提出了一种结合图形神经网络（GNN）和强化学习（RL）的自动图形编解码模型压缩（AGMC）方法。我们将目标DNN建模为一个图，并使用GNN自动学习DNN的嵌入。我们将该方法与基于规则的DNN嵌入模型压缩方法进行了比较，证明了该方法的有效性。结果表明，基于学习的DNN嵌入在搜索步骤较少的情况下，获得了更好的性能和更高的压缩比。我们在参数化和移动友好的DNN上评估了我们的方法，并将我们的方法与手工制作和基于学习的模型压缩方法进行了比较。在过度参数化的DNN（如ResNet-56）上，我们的方法优于手工和基于学习的方法，准确率分别高出4.36%和2.56%。此外，在MobileNet-v2上，我们实现了比最新方法更高的压缩比，精度损失仅为0.93%。

We propose generalized convolutional kernels for 3D reconstruction with ConvNets from point clouds. Our method uses multiscale convolutional kernels that can be applied to adaptive grids as generated with octrees. In addition to standard kernels in which each element has a distinct spatial location relative to the center, our elements have a distinct relative location as well as a relative scale level. Making our kernels span multiple resolutions allows us to apply ConvNets to adaptive grids for large problem sizes where the input data is sparse but the entire domain needs to be processed. Our ConvNet architecture can predict the signed and unsigned distance fields for large data sets with millions of input points and is faster and more accurate than classic energy minimization or recent learning approaches. We demonstrate this in a zero-shot setting where we only train on synthetic data and evaluate on the Tanks and Temples dataset of real-world large-scale 3D scenes.

我们提出了广义卷积核的三维重建与convnet从点云。我们的方法使用多尺度卷积核，可以应用于自适应网格生成的八叉树。除了标准内核（每个元素相对于中心有一个不同的空间位置）之外，我们的元素还有一个不同的相对位置和相对比例级别。使内核跨越多个分辨率使我们能够将CONVNET应用于大型问题的自适应网格，其中输入数据稀疏，但整个域需要处理。我们的ConvNet体系结构可以预测具有数百万个输入点的大型数据集的有符号和无符号距离场，并且比经典的能量最小化或最近的学习方法更快、更准确。我们在一个零镜头设置中演示了这一点，在该设置中，我们仅对合成数据进行训练，并对真实世界大规模3D场景的坦克和庙宇数据集进行评估。

SmartShadow is a deep learning application for digital painting artists to draw shadows on line drawings, with three proposed tools. (1) Shadow brush: artists can draw scribbles to coarsely indicate the areas inside or outside their wanted shadows, and the application will generate the shadows in real-time. (2) Shadow boundary brush: this brush can precisely control the boundary of any specific shadow. (3) Global shadow generator: this tool can estimate the global shadow direction from input brush scribbles, and then consistently propagate local shadows to the entire image. These three tools can not only speed up the shadow drawing process (by 3.1 times as experiments validate), but also allow for the flexibility to achieve various shadow effects and facilitate richer artistic creations. To this end, we train Convolutional Neural Networks (CNNs) with a collected large-scale dataset of both real and synthesized data, and especially, we collect 1670 shadow samples drawn by real artists. Both qualitative analysis and user study show that our approach can generate high-quality shadows that are practically usable in the daily works of digital painting artists. We present 30 additional results and 15 visual comparisons in the supplementary material.

SmartShadow是一款深度学习应用程序，可供数字绘画艺术家在线条图上绘制阴影，并提供三种建议工具。（1）阴影画笔：艺术家可以绘制涂鸦，粗略地指示他们想要的阴影内外的区域，应用程序将实时生成阴影。（2）阴影边界笔刷：该笔刷可以精确控制任何特定阴影的边界。（3）全局阴影生成器：该工具可以根据输入笔刷涂鸦估计全局阴影方向，然后将局部阴影一致地传播到整个图像。这三种工具不仅可以加快阴影绘制过程（实验证为3.1倍），还可以灵活地实现各种阴影效果，促进更丰富的艺术创作。为此，我们使用收集的真实和合成数据的大规模数据集来训练卷积神经网络（CNN），特别是我们收集了真实艺术家绘制的1670个阴影样本。定性分析和用户研究都表明，我们的方法可以生成高质量的阴影，这些阴影实际上可以在数字绘画艺术家的日常作品中使用。我们在补充装备中提供了30个附加结果和15个视觉比较。

A table arranging data in rows and columns is a very effective data structure, which has been widely used in business and scientific research. Considering large-scale tabular data in online and offline documents, automatic table recognition has attracted increasing attention from the document analysis community. Though human can easily understand the structure of tables, it remains a challenge for machines to understand that, especially due to a variety of different table layouts and styles. Existing methods usually model a table as either the markup sequence or the adjacency matrix between different table cells, failing to address the importance of the logical location of table cells, e.g., a cell is located in the first row and the second column of the table. In this paper, we reformulate the problem of table structure recognition as the table graph reconstruction, and propose an end-to-end trainable table graph reconstruction network (TGRNet) for table structure recognition. Specifically, the proposed method has two main branches, a cell detection branch and a cell logical location branch, to jointly predict the spatial location and the logical location of different cells. Experimental results on three popular table recognition datasets and a new dataset with table graph annotations (TableGraph-350K) demonstrate the effectiveness of the proposed TGRNet for table structure recognition. Code and annotations will be made publicly available.

按行和列排列数据的表是一种非常有效的数据结构，在商业和科学的研究中得到了广泛的应用。考虑到在线和离线文档中的大规模表格数据，自动表格识别已引起文档分析界越来越多的关注。虽然人类可以很容易地理解表格的结构，但是机器理解表格结构仍然是一个挑战，特别是由于各种不同的表格布局和样式。现有方法通常将表建模为不同表单元之间的标记序列或邻接矩阵，无法解决表单元逻辑位置的重要性，例如，单元格位于表的第一行和第二列。本文将表结构识别问题转化为表图重构问题，提出了一种用于表结构识别的端到端可训练表图重构网络（TGRNet）。具体来说，该方法有两个主要分支，一个是小区检测分支，一个是小区逻辑位置分支，用于联合预测不同小区的空间位置和逻辑位置。在三个流行

的表识别数据集和一个带有表图注释的新数据集 (TableGraph-350K) 上的实验结果证明了所提出的 TGRNet 在表结构识别中的有效性。代码和注释将公开提供。

This paper aims at addressing the problem of substantial performance degradation at extremely low computational cost (e.g. 5M FLOPs on ImageNet classification). We found that two factors, sparse connectivity and dynamic activation function, are effective to improve the accuracy. The former avoids the significant reduction of network width, while the latter mitigates the detriment of reduction in network depth. Technically, we propose micro-factorized convolution, which factorizes a convolution matrix into low rank matrices, to integrate sparse connectivity into convolution. We also present a new dynamic activation function, named Dynamic Shift Max, to improve the non-linearity via maxing out multiple dynamic fusions between an input feature map and its circular channel shift. Building upon these two new operators, we arrive at a family of networks, named MicroNet, that achieves significant performance gains over the state of the art in the low FLOP regime. For instance, under the constraint of 12M FLOPs, MicroNet achieves 59.4% top-1 accuracy on ImageNet classification, outperforming MobileNetV3 by 9.6%. Source code is at <https://github.com/liyunsheng13/micronet>.

本文旨在以极低的计算成本（例如ImageNet分类上的500万次浮点运算）解决性能大幅下降的问题。我们发现两个因素，稀疏连通性和动态激活函数，是有效的提高精度。前者避免了网络宽度的显著减小，而后者减轻了网络深度减小的不利影响。在技术上，我们提出了微因子化卷积，将卷积矩阵分解为低秩矩阵，将稀疏连通性集成到卷积中。我们还提出了一个新的动态激活函数dynamicshift-Max，通过最大化输入特征映射与其循环通道移位之间的多个动态融合来改善非线性。在这两个新运营商的基础上，我们形成了一个名为MicroNet的网络家族，该家族在低FLOP模式下实现了超过最先进水平的显著性能提升。例如，在12M触发器的限制下，MicroNet在ImageNet分类方面达到59.4%的顶级精度，比MobileNetV3高9.6%。源代码位于<https://github.com/liyunsheng13/micronet>.

We propose a novel framework for fine-grained object recognition that learns to recover object variation in 3D space from a single image, trained on an image collection without using any ground-truth 3D annotation. We accomplish this by representing an object as a composition of 3D shape and its appearance, while eliminating the effect of camera viewpoint, in a canonical configuration. Unlike conventional methods modeling spatial variation in 2D images only, our method is capable of reconfiguring the appearance feature in a canonical 3D space, thus enabling the subsequent object classifier to be invariant under 3D geometric variation. Our representation also allows us to go beyond existing methods, by incorporating 3D shape variation as an additional cue for object recognition. To learn the model without ground-truth 3D annotation, we deploy a differentiable renderer in an analysis-by-synthesis framework. By incorporating 3D shape and appearance jointly in a deep representation, our method learns the discriminative representation of the object and achieves competitive performance on fine-grained image recognition and vehicle re-identification. We also demonstrate that the performance of 3D shape reconstruction is improved by learning fine-grained shape deformation in a boosting manner.

我们提出了一种新的细粒度对象识别框架，该框架学习从单个图像恢复三维空间中的对象变化，在图像集合上进行训练，而不使用任何地面真实三维注释。我们通过将对象表示为三维形状及其外观的组合来实现这一点，同时消除相机视点在规范配置中的影响。与传统的仅在二维图像中建模空间变化的方法不同，我们的方法能够在规范的三维空间中重新配置外观特征，从而使后续的对象分类器在三维几何变化下保持不变。我们的表示法还允许我们超越现有的方法，将3D形状变化作为物体识别的额外线索。为了学习没有地面真实三维注释的模型，我们在综合分析框架中部署了一个可微渲染器。通过将三维形状和外观结合在一个深度表示中，我们的方法学习了对象的区分性表示，并在细粒度图像识别和车辆重新识

别方面取得了有竞争力的性能。我们还证明，通过以增强方式学习细粒度形状变形，可以提高三维形状重建的性能。

Image generation has rapidly evolved in recent years. Modern architectures for adversarial training allow to generate even high resolution images with remarkable quality. At the same time, more and more effort is dedicated towards controlling the content of generated images. In this paper, we take one further step in this direction and propose a conditional generative adversarial network (GAN) that generates images with a defined number of objects from given classes. This entails two fundamental abilities (1) being able to generate high-quality images given a complex constraint and (2) being able to count object instances per class in a given image. Our proposed model modularly extends the successful StyleGAN2 architecture with a count-based conditioning as well as with a regression sub-network to count the number of generated objects per class during training. In experiments on three different datasets, we show that the proposed model learns to generate images according to the given multiple-class count condition even in the presence of complex backgrounds. In particular, we propose a new dataset, CityCount, which is derived from the Cityscapes street scenes dataset, to evaluate our approach in a challenging and practically relevant scenario. An implementation is available at <https://github.com/boschresearch/MCCGAN>.

近年来，图像生成技术发展迅速。现代对抗性训练体系结构允许生成高分辨率图像，质量卓越。同时，越来越多的工作致力于控制生成图像的内容。在本文中，我们朝着这个方向迈出了进一步的一步，并提出了一种条件生成对抗网络（GAN），它可以从给定的类中生成具有定义数量的对象的图像。这需要两个基本能力（1）能够在给定复杂约束的情况下生成高质量图像，以及（2）能够在给定图像中计算每个类的对象实例。我们提出的模型模块化地扩展了成功的StyleGAN2体系结构，使用基于计数的条件化以及回归子网络来计算训练期间每个类生成的对象数。在三个不同数据集上的实验表明，即使在复杂背景下，该模型也能根据给定的多类计数条件学习生成图像。特别是，我们提出了一个新的数据集CityCount，它来自Cityscapes street scenes数据集，用于在一个具有挑战性且实际相关的场景中评估我们的方法。可在以下网址获得实施：<https://github.com/boschresearch/MCCGAN>。

We present a novel pyramidal output representation to ensure parsimony with our "specialize and fuse" process for semantic segmentation. A pyramidal "output" representation consists of coarse-to-fine levels, where each level is "specialize" in a different class distribution (e.g., more stuff than things classes at coarser levels). Two types of pyramidal outputs (i.e., unity and semantic pyramid) are "fused" into the final semantic output, where the unity pyramid indicates unity-cells (i.e., all pixels in such cell share the same semantic label). The process ensures parsimony by predicting a relatively small number of labels for unity-cells (e.g., a large cell of grass) to build the final semantic output. In addition to the "output" representation, we design a coarse-to-fine contextual module to aggregate the "features" representation from different levels. We validate the effectiveness of each key module in our method through comprehensive ablation studies. Finally, our approach achieves state-of-the-art performance on three widely-used semantic segmentation datasets---ADE20K, COCO-Stuff, and Pascal-Context.

我们提出了一种新的金字塔输出表示，以确保我们的“专门化和融合”语义分割过程的简约性。金字塔“输出”表示由粗到细的级别组成，其中每个级别在不同的类分布中“专门化”（例如，在较粗级别的类中，内容多于内容）。两种类型的金字塔输出（即统一和语义金字塔）被“融合”到最终语义输出中，其中统一金字塔表示统一单元（即该单元中的所有像素共享相同的语义标签）。该过程通过预测unity单元（例如，一个大的草单元）相对较少的标签来构建最终的语义输出，从而确保节约。除了“输出”表示之外，我们还设计了一个从粗到精的上下文模块，从不同的层次聚合“特征”表示。我们通过全面的消融研究来

验证我们方法中每个关键模块的有效性。最后，我们的方法在三个广泛使用的语义分割数据集 ADE20K、COCO Stuff 和 Pascal Context 上实现了最先进的性能。

Shadow removal from a single image is generally still an open problem. Most existing learning-based methods use supervised learning and require a large number of paired images (shadow and corresponding non-shadow images) for training. A recent unsupervised method, Mask-ShadowGAN, addresses this limitation. However, it requires a binary mask to represent shadow regions, making it inapplicable to soft shadows. To address the problem, in this paper, we propose an unsupervised domain-classifier guided shadow removal network, DC-ShadowNet. Specifically, we propose to integrate a shadow/shadow-free domain classifier into a generator and its discriminator, enabling them to focus on shadow regions. To train our network, we introduce novel losses based on physics-based shadow-free chromaticity, shadow-robust perceptual features, and boundary smoothness. Moreover, we show that our network being unsupervised can be used for test-time training that further improves the results. Our experiments show that all these novel components allow our method to handle soft shadows, and also to perform better on hard shadows both quantitatively and qualitatively than the existing state-of-the-art shadow removal methods.

从单个图像中去除阴影通常仍然是一个开放的问题。大多数现有的基于学习的方法使用监督学习，需要大量成对图像（阴影和相应的非阴影图像）进行训练。最近的一种无监督方法Mask ShadowGAN解决了这一限制。但是，它需要一个二进制遮罩来表示阴影区域，这使得它不适用于软阴影。为了解决这个问题，在本文中，我们提出了一种无监督的领域分类器引导的阴影去除网络，DC ShadowNet。具体来说，我们建议将阴影/无阴影域分类器集成到生成器及其鉴别器中，使它们能够聚焦阴影区域。为了训练我们的网络，我们引入了基于物理的无阴影色度、阴影鲁棒感知特征和边界平滑度的新损耗。此外，我们还表明，我们的无监督网络可以用于测试时训练，从而进一步提高测试结果。我们的实验表明，所有这些新的组件允许我们的方法处理软阴影，并且在数量和质量上都比现有的最先进的阴影去除方法在硬阴影上表现更好。

The recently proposed visual image Transformers (ViT) with pure attention have achieved promising performance on image recognition tasks, such as image classification. However, the routine of the current ViT model is to maintain a full-length patch sequence during inference, which is redundant and lacks hierarchical representation. To this end, we propose a Hierarchical Visual Transformer (HVT) which progressively pools visual tokens to shrink the sequence length and hence reduces the computational cost, analogous to the feature maps downsampling in Convolutional Neural Networks (CNNs). It brings a great benefit that we can increase the model capacity by scaling dimensions of depth/width/resolution/patch size without introducing extra computational complexity due to the reduced sequence length. Moreover, we empirically find that the average pooled visual tokens contain more discriminative information than the single class token. To demonstrate the improved scalability of our HVT, we conduct extensive experiments on the image classification task. With comparable FLOPs, our HVT outperforms the competitive baselines on ImageNet and CIFAR-100 datasets. Code is available at <https://github.com/MonashAI/HVT>.

最近提出的纯注意力视觉图像转换器 (ViT) 在图像识别任务 (如图像分类) 中取得了良好的性能。然而，当前ViT模型的常规是在推理过程中保持完整的补丁序列，这是冗余的，并且缺乏层次表示。为此，我们提出了一种分层视觉变换器 (HVT)，该变换器可逐步汇集视觉标记以缩短序列长度，从而降低计算成本，类似于卷积神经网络 (CNN) 中的特征映射下采样。它带来了一个巨大的好处，即我们可以通过缩放深度/宽度/分辨率/面片大小的维度来增加模型容量，而不会由于减少的序列长度而引入额外的计算复杂性。此外，我们实证发现，平均集合视觉标记比单一类别标记包含更多的鉴别信息。为了证明我

们的HVT改进的可伸缩性，我们对图像分类任务进行了广泛的实验。通过类似的失败，我们的HVT优于ImageNet和CIFAR-100数据集上的竞争基线。代码可在<https://github.com/MonashAI/HVT>。

Deformable templates are essential to large-scale medical image registration, segmentation, and population analysis. Current conventional and deep network-based methods for template construction use only regularized registration objectives and often yield templates with blurry and/or anatomically implausible appearance, confounding downstream biomedical interpretation. We reformulate deformable registration and conditional template estimation as an adversarial game wherein we encourage realism in the moved templates with a generative adversarial registration framework conditioned on flexible image covariates. The resulting templates exhibit significant gain in specificity to attributes such as age and disease, better fit underlying group-wise spatiotemporal trends, and achieve improved sharpness and centrality. These improvements enable more accurate population modeling with diverse covariates for standardized downstream analyses and easier anatomical delineation for structures of interest.

可变形模板是大规模医学图像配准、分割和总体分析的基础。当前传统的和基于深度网络的模板构建方法仅使用正则化的注册目标，并且通常生成外观模糊和/或解剖学上不可信的模板，混淆了下游生物医学解释。我们将可变形配准和条件模板估计重新表述为一个对抗性游戏，其中我们鼓励移动模板中的真实性，并使用一个以灵活的图像协变量为条件的生成性对抗性配准框架。由此产生的模板在年龄和疾病等属性的特异性方面表现出显著的增益，更好地适应潜在的群体时空趋势，并实现更好的清晰度和中心性。这些改进使得使用不同的协变量进行更精确的群体建模，以进行标准化的下游分析，并更容易对感兴趣的结构进行解剖学描述。

We propose an Auto-Parsing Network (APN) to discover and exploit the input data's hidden tree structures for improving the effectiveness of the Transformer-based vision-language systems. Specifically, we impose a Probabilistic Graphical Model (PGM) parameterized by the attention operations on each self-attention layer to incorporate sparse assumption. We use this PGM to softly segment an input sequence into a few clusters where each cluster can be treated as the parent of the inside entities. By stacking these PGM constrained self-attention layers, the clusters in a lower layer compose into a new sequence, and the PGM in a higher layer will further segment this sequence. Iteratively, a sparse tree can be implicitly parsed, and this tree's hierarchical knowledge is incorporated into the transformed embeddings, which can be used for solving the target vision-language tasks. Specifically, we showcase that our APN can strengthen Transformer based networks in two major vision-language tasks: Captioning and Visual Question Answering. Also, a PGM probability-based parsing algorithm is developed by which we can discover what the hidden structure of input is during the inference.

为了提高基于转换器的视觉语言系统的有效性，我们提出了一种自动解析网络（APN）来发现和利用输入数据的隐藏树结构。具体来说，我们在每个自我注意层上施加一个由注意操作参数化的概率图形模型（PGM），以合并稀疏假设。我们使用此PGM将输入序列软分割为几个簇，其中每个簇都可以被视为内部实体的父级。通过堆叠这些PGM约束的自我注意层，较低层中的簇组成一个新序列，较高层中的PGM将进一步分割该序列。通过迭代，可以隐式解析稀疏树，并将该树的层次知识合并到转换后的嵌入中，用于解决目标视觉语言任务。具体来说，我们展示了我们的APN可以在两个主要的视觉语言任务中加强基于转换器的网络：字幕和视觉问答。此外，我们还开发了一种基于概率的PGM解析算法，通过该算法可以在推理过程中发现输入的隐藏结构。

Generative adversarial networks built from deep convolutional neural networks (GANs) lack the ability to exactly replicate the high-frequency components of natural images. To alleviate this issue, we introduce two novel training techniques called frequency dropping (F-Drop) and frequency matching (F-Match). The key idea of F-Drop is to filter out unnecessary high-frequency components from the input images of the discriminators. This simple modification prevents the discriminators from being confused by perturbations of the high-frequency components. In addition, F-Drop makes the GANs focus on fitting in the low-frequency domain, in which there are the dominant components of natural images. F-Match minimizes the difference between real and fake images in the frequency domain for generating more realistic images. F-Match is implemented as a regularization term in the objective functions of the generators; it penalizes the batch mean error in the frequency domain. F-Match helps the generators to fit in the high-frequency domain filtered out by F-Drop to the real image. We experimentally demonstrate that the combination of F-Drop and F-Match improves the generative performance of GANs in both the frequency and spatial domain on multiple image benchmarks.

由深度卷积神经网络 (GANs) 构建的生成性对抗网络缺乏精确复制自然图像高频成分的能力。为了缓解这个问题，我们引入了两种新的训练技术，称为频率下降 (F-Drop) 和频率匹配 (F-Match)。F-Drop 的关键思想是从鉴别器的输入图像中滤除不必要的高频成分。这种简单的修改可防止鉴别器因高频分量的扰动而混淆。此外，F-Drop使GANs专注于低频域的拟合，其中自然图像的主要成分是低频域。F-Match在频域中最小化真实图像和虚假图像之间的差异，以生成更真实的图像。F-匹配作为正则化项在生成器的目标函数中实现；它在频域中惩罚批次平均误差。F-Match有助于生成器适应F-Drop过滤出的高频域，以适应真实图像。我们的实验表明，F-Drop和F-Match的组合在多个图像基准上提高了GANs在频域和空域的生成性能。

Protein structure determination from cryo-EM data requires reconstructing a 3D volume (or distribution of volumes) from many noisy and randomly oriented 2D projection images. While the standard homogeneous reconstruction task aims to recover a single static structure, recently-proposed neural and non-neural methods can reconstruct distributions of structures, thereby enabling the study of protein complexes that possess intrinsic structural or conformational heterogeneity. These heterogeneous reconstruction methods, however, require fixed image poses, which are typically estimated from an upstream homogeneous reconstruction and are not guaranteed to be accurate under highly heterogeneous conditions. In this work we describe cryoDRGN2, an ab initio reconstruction algorithm, which can jointly estimate image poses and learn a neural model of a distribution of 3D structures on real heterogeneous cryo-EM data. To achieve this, we adapt search algorithms from the traditional cryo-EM literature, and describe the optimizations and design choices required to make such a search procedure computationally tractable in the neural model setting. We show that cryoDRGN2 is robust to the high noise levels of real cryo-EM images, trains faster than earlier neural methods, and achieves state-of-the-art performance on real cryo-EM datasets.

从cryo EM数据确定蛋白质结构需要从许多噪声和随机定向的2D投影图像重建3D体积（或体积分布）。虽然标准的同质重建任务旨在恢复单个静态结构，但最近提出的神经和非神经方法可以重建结构的分布，从而能够研究具有内在结构或构象异质性的蛋白质复合物。然而，这些非均匀重建方法需要固定的图像姿势，这些姿势通常是从上游均匀重建估计的，并且不能保证在高度非均匀的条件下是准确的。在这项工作中，我们描述了cryoDRGN2，一种从头算重建算法，它可以联合估计图像姿态并学习真实非均匀cryo EM数据上三维结构分布的神经模型。为了实现这一点，我们采用了传统cryo EM文献中的搜索算法，并描述了使这种搜索过程在神经模型设置中计算可处理所需的优化和设计选择。我们表明

cryoDRGN2对真实cryo EM图像的高噪声水平具有鲁棒性，训练速度比早期的神经方法更快，并且在真实cryo EM数据集上实现了最先进的性能。

We consider the shuffled linear regression problem where the correspondences between covariates and responses are unknown. While the existing formulation assumes an ideal underlying bijection in which all pieces of data should match, such an assumption barely holds in real-world applications due to either missing data or outliers. Therefore, in this work, we generalize the formulation of shuffled linear regression to a broader range of conditions where only part of the data should correspond. Moreover, we present a remarkably simple yet effective optimization algorithm with guaranteed global convergence. Distinct tasks validate the effectiveness of the proposed method.

我们考虑混合变量和响应之间的对应关系未知的混洗线性回归问题。虽然现有公式假设所有数据块都应该匹配的理想基础双射，但由于缺少数据或异常值，这种假设在实际应用中几乎不成立。因此，在这项工作中，我们将混洗线性回归公式推广到更广泛的条件，其中只有部分数据应该对应。此外，我们提出了一个非常简单但有效的优化算法，保证了全局收敛性。不同的任务验证了该方法的有效性。

Visual storytelling and story comprehension are uniquely human skills that play a central role in how we learn about and experience the world. Despite remarkable progress in recent years in synthesis of visual and textual content in isolation and learning effective joint visual-linguistic representations, existing systems still operate only at a superficial, factual level. With the goal of developing systems that are able to comprehend rich human-generated narratives, and co-create new stories, we introduce AESOP: a new dataset that captures the creative process associated with visual storytelling. Visual panels are composed of clip-art objects with specific attributes enabling a broad range of creative expression. Using AESOP, we propose foundational storytelling tasks that are generative variants of story cloze tests, to better measure the creative and causal reasoning ability required for visual storytelling. We further develop a generalized story completion framework that models stories as the co-evolution of visual and textual concepts. We benchmark the proposed approach with human baselines and evaluate using comprehensive qualitative and quantitative metrics. Our results highlight key insights related to the dataset, modelling and evaluation of visual storytelling for future research in this promising field of study.

视觉故事讲述和故事理解是独特的人类技能，在我们如何了解和体验世界中起着核心作用。尽管近年来在孤立地合成视觉和文本内容以及学习有效的联合视觉语言表达方面取得了显著进展，但现有系统仍然只在表面的、事实的层面上运行。为了开发能够理解丰富的人类故事并共同创造新故事的系统，我们引入了伊索：一个新的数据集，它捕捉了与视觉故事讲述相关的创作过程。视觉面板由具有特定属性的剪贴画对象组成，可实现广泛的创造性表达。使用伊索，我们提出了基本的故事讲述任务，这些任务是故事完形填空测试的生成变体，以更好地衡量视觉故事讲述所需的创造性和因果推理能力。我们进一步开发了一个通用的故事完成框架，该框架将故事建模为视觉和文本概念的共同进化。我们用人类基线对提议的方法进行基准测试，并使用综合的定性和定量指标进行评估。我们的研究结果突出了与数据集、可视化讲故事建模和评估相关的关键见解，为这一前景广阔的研究领域的未来研究提供了参考。

Class Activation Mapping (CAM) is a powerful technique used to understand the decision making of Convolutional Neural Network (CNN) in computer vision. Recently, there have been attempts not only to generate better visual explanations, but also to improve classification performance using visual explanations. However, previous works still have their own drawbacks. In this paper, we propose a novel architecture, LFI-CAM\*\*\* (Learning Feature Importance Class Activation Mapping), which is trainable for image classification and visual explanation in an end-to-end manner. LFI-CAM generates attention map for visual explanation during forward propagation, and simultaneously uses attention map to improve classification performance through the attention mechanism. Feature Importance Network (FIN) focuses on learning the feature importance instead of directly learning the attention map to obtain a more reliable and consistent attention map. We confirmed that LFI-CAM is optimized not only by learning the feature importance but also by enhancing the backbone feature representation to focus more on important features of the input image. Experiments show that LFI-CAM outperforms baseline models' accuracy on classification tasks as well as significantly improves on previous works in terms of attention map quality and stability over different hyper-parameters.

类激活映射 (CAM) 是理解计算机视觉中卷积神经网络 (CNN) 决策的一种强有力的技术。最近，人们不仅尝试生成更好的视觉解释，而且还尝试使用视觉解释提高分类性能。然而，以前的工作仍然有自己的缺点。在本文中，我们提出了一种新的体系结构LFI-CAM\*\*\*（学习特征重要性类激活映射），该体系结构可用于端到端的图像分类和视觉解释。LFI-CAM在前向传播过程中生成用于视觉解释的注意图，同时使用注意图通过注意机制提高分类性能。特征重要性网络 (FIN) 关注的是特征重要性的学习，而不是直接学习注意图，以获得更可靠和一致的注意图。我们证实，LFI-CAM不仅通过学习特征重要性进行优化，而且通过增强主干特征表示来更加关注输入图像的重要特征。实验表明，LFI-CAM在分类任务上优于基线模型，并且在不同超参数下的注意图质量和稳定性方面显著提高。

Compared with the visual grounding on 2D images, the natural-language-guided 3D object localization on point clouds is more challenging. In this paper, we propose a new model, named InstanceRefer, to achieve a superior 3D visual grounding through the grounding-by-matching strategy. In practice, our model first predicts the target category from the language descriptions using a simple language classification model. Then based on the category, our model sifts out a small number of instance candidates (usually less than 20) from the panoptic segmentation on point clouds. Thus, the non-trivial 3D visual grounding task has been effectively re-formulated as a simplified instance-matching problem, considering that instance-level candidates are more rational than the redundant 3D object proposals. Subsequently, for each candidate, we perform the multi-level contextual inference, i.e., referring from instance attribute perception, instance-to-instance relation perception, and instance-to-background global localization perception, respectively. Eventually, the most relevant candidate is selected and localized by ranking confidence scores, which are obtained by the cooperative holistic visual-language feature matching. Experiments confirm that our method outperforms previous state-of-the-arts on ScanRefer online benchmark (ranked 1st place) and Nr3D/Sr3D datasets.

与二维图像的视觉基础相比，自然语言引导的点云三维目标定位更具挑战性。在本文中，我们提出了一种新的模型，名为InstanceRefer，通过匹配策略实现良好的三维视觉接地。在实践中，我们的模型首先使用一个简单的语言分类模型从语言描述中预测目标类别。然后根据类别，我们的模型从点云的全景分割中筛选出少量候选实例（通常少于20个）。因此，考虑到实例级候选方案比冗余的三维对象方案更合理，非平凡的三维视觉基础任务被有效地重新表述为一个简化的实例匹配问题。随后，我们对每个候选对象进行多层次的上下文推理，即分别从实例属性感知、实例到实例关系感知和实例到背景全局定位感知进行引用。最后，通过合作整体视觉语言特征匹配获得置信度评分，选择最相关的候选对象并进行定

位。实验证实，在ScanRefer在线基准测试（排名第一）和Nr3D/Sr3D数据集上，我们的方法优于以前的最新技术。

Video-and-Language Inference is a recently proposed task for joint video-and-language understanding. This new task requires a model to draw inference on whether a natural language statement entails or contradicts a given video clip. In this paper, we study how to address three critical challenges for this task: judging the global correctness of the statement involved multiple semantic meanings, joint reasoning over video and subtitles, and modeling long-range relationships and complex social interactions. First, we propose an adaptive hierarchical graph network that achieves in-depth understanding of the video over complex interactions. Specifically, it performs joint reasoning over video and subtitles in three hierarchies, where the graph structure is adaptively adjusted according to the semantic structures of the statement. Secondly, we introduce semantic coherence learning to explicitly encourage the semantic coherence of the adaptive hierarchical graph network from three hierarchies. The semantic coherence learning can further improve the alignment between vision and linguistics, and the coherence across a sequence of video segments. Experimental results show that our method significantly outperforms the baseline by a large margin.

视频和语言推理是最近提出的一项视频和语言联合理解任务。这项新任务需要一个模型来推断自然语言陈述是否包含或与给定的视频剪辑相矛盾。在本文中，我们研究了如何解决这项任务的三个关键挑战：判断涉及多种语义的语句的全局正确性，对视频和字幕进行联合推理，以及对长期关系和复杂社会互动进行建模。首先，我们提出了一种自适应的层次图网络，该网络能够深入理解复杂交互中的视频。具体来说，它在三个层次中对视频和字幕执行联合推理，其中图形结构根据语句的语义结构自适应调整。其次，我们引入语义一致性学习，从三个层次明确鼓励自适应层次图网络的语义一致性。语义连贯学习可以进一步提高视觉和语言学之间的一致性，以及视频片段序列之间的连贯性。实验结果表明，我们的方法明显优于基线。

Existing salient object detection (SOD) models usually focus on either backbone feature extractors or saliency heads, ignoring their relations. A powerful backbone could still achieve sub-optimal performance with a weak saliency head and vice versa. Moreover, the balance between model performance and inference latency poses a great challenge to model design, especially when considering different deployment scenarios. Considering all components in an integral neural architecture search (iNAS) space, we propose a flexible device-aware search scheme that only trains the SOD model once and quickly finds high-performance but low-latency models on multiple devices. An evolution search with latency-group sampling (LGS) is proposed to explore the entire latency area of our enlarged search space. Models searched by iNAS achieve similar performance with SOTA methods but reduce the 3.8x, 3.3x, 2.6x, 1.9x latency on Huawei Nova6 SE, Intel Core CPU, the Jetson Nano, and Nvidia Titan Xp. The code is released at <https://mmcheng.net/inas/>.

现有的显著目标检测（SOD）模型通常只关注主干特征提取或显著性头部，而忽略了它们之间的关系。强大的主干网仍然可以在显著性头部较弱的情况下实现次优性能，反之亦然。此外，模型性能和推理延迟之间的平衡对模型设计提出了巨大挑战，特别是在考虑不同部署场景时。考虑到整体神经架构搜索（iNAS）空间中的所有组件，我们提出了一种灵活的设备感知搜索方案，该方案只需训练SOD模型一次，并在多个设备上快速找到高性能但低延迟的模型。提出了一种基于延迟群抽样（LGS）的进化搜索方法，以探索我们扩大的搜索空间的整个延迟区域。iNAS搜索的型号与SOTA方法的性能相似，但在华为Nova6 SE、Intel Core CPU、Jetson Nano和Nvidia Titan Xp上减少了3.8倍、3.3倍、2.6倍和1.9倍的延迟。该代码发布于<https://mmcheng.net/inas/>。

We present a flexible and high-performance framework, named Pyramid R-CNN, for two-stage 3D object detection from point clouds. Current approaches generally rely on the points or voxels of interest for ROI feature extraction on the second stage, but cannot effectively handle the sparsity and non-uniform distribution of those points, and this may result in failures in detecting objects that are far away. To resolve the problems, we propose a novel second-stage module, named pyramid ROI head, to adaptively learn the features from the sparse points of interest. The pyramid ROI head consists of three key components. Firstly, we propose the ROI-grid Pyramid, which addresses the sparsity problem by extensively collecting points of interest for each ROI in a pyramid manner. Secondly, we propose ROI-grid Attention, a new operation that can encode richer information from sparse points by incorporating conventional attention-based and graph-based point operators into a unified formulation. Thirdly, we propose the Density-Aware Radius Prediction (DARP) module, which can adapt to different point density levels by dynamically adjusting the focusing range of ROIs. Combining the three components, our pyramid ROI head is robust to the sparse and imbalanced circumstances, and can be applied upon various 3D backbones to consistently boost the detection performance. Extensive experiments show that Pyramid R-CNN outperforms the state-of-the-art 3D detection models by a large margin on both the KITTI dataset and the Waymo Open dataset.

我们提出了一个灵活的高性能框架，称为金字塔R-CNN，用于从点云进行两阶段3D目标检测。当前的方法通常在第二阶段依赖感兴趣的点或体素进行ROI特征提取，但不能有效地处理这些点的稀疏性和非均匀分布，这可能导致无法检测远处的对象。为了解决这些问题，我们提出了一种新的第二阶段模块，称为金字塔ROI头，用于从稀疏的兴趣点自适应地学习特征。金字塔ROI头部由三个关键组件组成。首先，我们提出了ROI网格金字塔，它通过以金字塔的方式广泛收集每个ROI的兴趣点来解决稀疏性问题。其次，我们提出了ROI网格注意，这是一种新的操作，通过将传统的基于注意和基于图形的点算子合并到一个统一的公式中，可以从稀疏点编码更丰富的信息。第三，我们提出了密度感知半径预测（DARP）模块，该模块通过动态调整ROI的聚焦范围来适应不同的点密度水平。结合这三个组件，我们的金字塔ROI头对稀疏和不平衡的情况具有鲁棒性，并且可以应用于各种3D主干上，以持续提高检测性能。大量实验表明，在KITTI数据集和Waymo开放数据集上，金字塔R-CNN的性能大大优于最先进的3D检测模型。

Semi-supervised learning (SSL) algorithms have attracted much attentions in medical image segmentation by leveraging unlabeled data, which challenge in acquiring massive pixel-wise annotated samples. However, most of the existing SSLs neglected the geometric shape constraint in object, leading to unsatisfactory boundary and non-smooth of object. In this paper, we propose a novel boundary-aware semi-supervised medical image segmentation network, named Graph-BAS3Net, which incorporates the boundary information and learns duality constraints between semantics and geometrics in the graph domain. Specifically, the proposed method consists of two components: a multi-task learning framework BAS3Net and a graph-based cross-task module BGCM. The BAS3Net improves the existing GAN-based SSL by adding a boundary detection task, which encodes richer features of object shape and surface. Moreover, the BGCM further explores the co-occurrence relations between the semantics segmentation and boundary detection task, so that the network learns stronger semantic and geometric correspondences from both labeled and unlabeled data. Experimental results on the LiTS dataset and COVID-19 dataset confirm that our proposed Graph-BAS3 Net outperforms the state-of-the-art methods in semi-supervised segmentation task.

半监督学习（SSL）算法在利用未标记数据进行医学图像分割方面受到了广泛关注，这对获取大量像素级标注样本提出了挑战。然而，现有的SSL大多忽略了物体的几何形状约束，导致物体边界不理想、不光滑。在本文中，我们提出了一种新的边界感知半监督医学图像分割网络Graph-BAS3Net，该网络融合了边界信息，并在图域中学习语义和几何之间的对偶约束。具体而言，该方法由两部分组成：一个多任务学习框架BAS3Net和一个基于图形的跨任务模块BGCM。BAS3Net通过添加边界检测任务改进了现有的

基于GAN的SSL，该任务编码了更丰富的对象形状和表面特征。此外，BGCM进一步探索了语义分割和边界检测任务之间的共生关系，从而使网络从标记和未标记的数据中学习到更强的语义和几何对应。LITS 2019冠状病毒疾病数据集和COVID-19数据集的实验结果证实了我们所提出的GROMA-BAS3网络优于半监督分割任务中的最先进的方法。

The recently developed vision transformer (ViT) has achieved promising results on image classification compared to convolutional neural networks. Inspired by this, in this paper, we study how to learn multi-scale feature representations in transformer models for image classification. To this end, we propose a dual-branch transformer to combine image patches (i.e., tokens in a transformer) of different sizes to produce stronger image features. Our approach processes small-patch and large-patch tokens with two separate branches of different computational complexity and these tokens are then fused purely by attention multiple times to complement each other. Furthermore, to reduce computation, we develop a simple yet effective token fusion module based on cross attention, which uses a single token for each branch as a query to exchange information with other branches. Our proposed cross-attention only requires linear time for both computational and memory complexity instead of quadratic time otherwise. Extensive experiments demonstrate that our approach performs better than or on par with several concurrent works on vision transformer, in addition to efficient CNN models. For example, on the ImageNet1K dataset, with some architectural changes, our approach outperforms the recent DeiT by a large margin of 2% with a small to moderate increase in FLOPs and model parameters. Our source codes and models are available at <https://github.com/IBM/CrossViT>.

与卷积神经网络相比，最近开发的视觉变换器（ViT）在图像分类方面取得了很好的效果。受此启发，本文研究了如何在变压器模型中学习多尺度特征表示用于图像分类。为此，我们提出了一种双分支变换器来组合不同大小的图像块（即变换器中的标记），以产生更强的图像特征。我们的方法处理具有两个不同计算复杂度的独立分支的小补丁和大补丁令牌，然后这些令牌纯粹通过注意多次融合以相互补充。此外，为了减少计算量，我们开发了一个简单而有效的基于交叉注意的令牌融合模块，该模块使用每个分支的单个令牌作为查询，与其他分支交换信息。我们提出的交叉注意只需要计算和内存复杂性的线性时间，而不是二次时间。大量的实验表明，除了有效的美国有线电视新闻网模型之外，我们的方法比在视觉变压器上的多个并行作品表现更好或更好。例如，在ImageNet1K数据集上，通过一些架构更改，我们的方法比最近的DeiT有2%的大幅度优势，触发器和模型参数有小到中等程度的增加。我们的源代码和模型可在<https://github.com/IBM/CrossViT>。

The point cloud representation of an object can have a large geometric variation in view of inconsistent data acquisition procedure, which thus leads to domain discrepancy due to diverse and uncontrollable shape representation cross datasets. To improve discrimination on unseen distribution of point-based geometries in a practical and feasible perspective, this paper proposes a new method of geometry-aware self-training (GAST) for unsupervised domain adaptation of object point cloud classification. Specifically, this paper aims to learn a domain-shared representation of semantic categories, via two novel self-supervised geometric learning tasks as feature regularization. On one hand, the representation learning is empowered by a linear mixup of point cloud samples with their self-generated rotation labels, to capture a global topological configuration of local geometries. On the other hand, a diverse point distribution across datasets can be normalized with a novel curvature-aware distortion localization. Experiments on the PointDA-10 dataset show that our GAST method can significantly outperform the state-of-the-art methods.

鉴于不一致的数据采集过程，对象的点云表示可能会有很大的几何变化，因此，由于不同和不可控的形状表示跨数据集，导致域差异。为了从实用和可行的角度提高对基于点的几何体不可见分布的识别，本文提出了一种新的几何感知自训练（GAST）方法，用于目标点云分类的无监督域自适应。具体而言，本文旨在通过两个新的自监督几何学习任务（如特征正则化）来学习语义类别的领域共享表示。一方面，通过点云样本与其自身生成的旋转标签的线性混合来实现表示学习，以捕获局部几何的全局拓扑结构。另一方面，跨数据集的不同点分布可以通过一种新颖的曲率感知失真定位进行规范化。在PointDA-10数据集上的实验表明，我们的GAST方法可以显著优于最新的方法。

Neural networks trained with class-imbalanced data are known to perform poorly on minor classes of scarce training data. Several recent works attribute this to over-fitting to minor classes. In this paper, we provide a novel explanation of this issue. We found that a neural network tends to first under-fit the minor classes by classifying most of their data into the major classes in early training epochs. To correct these wrong predictions, the neural network then must focus on pushing features of minor class data across the decision boundaries between major and minor classes, leading to much larger gradients for features of minor classes. We argue that such an under-fitting phase over-emphasizes the competition between major and minor classes, hinders the neural network from learning the discriminative knowledge that can be generalized to test data, and eventually results in over-fitting. To address this issue, we propose a novel learning strategy to equalize the training progress across classes. We mix features of the major class data with those of other data in a mini-batch, intentionally weakening their features to prevent a neural network from fitting them first. We show that this strategy can largely balance the training accuracy and feature gradients across classes, effectively mitigating the under-fitting then over-fitting problem for minor class data. On several benchmark datasets, our approach achieves the state-of-the-art accuracy, especially for the challenging step-imbalanced cases.

众所周知，使用类不平衡数据训练的神经网络在少量稀缺训练数据上表现不佳。最近的几项工作将这归因于对次要课程的过度拟合。在本文中，我们提供了一个新的解释这个问题。我们发现，在早期训练阶段，神经网络倾向于首先通过将小类的大部分数据分类到大类来对小类进行欠拟合。为了纠正这些错误的预测，神经网络必须专注于将次要类别数据的特征推过主要类别和次要类别之间的决策边界，从而导致次要类别特征的最大梯度。我们认为，这种欠拟合阶段过分强调了大类和小类之间的竞争，阻碍了神经网络学习可推广到测试数据的区分性知识，并最终导致过度拟合。为了解决这个问题，我们提出了一种新的学习策略来平衡不同班级的培训进度。我们将主要类别数据的特征与其他数据的特征混合在一个小批量中，故意削弱它们的特征，以防止神经网络首先拟合它们。我们表明，该策略可以在很大程度上平衡不同类别的训练精度和特征梯度，有效地缓解次要类别数据的欠拟合和过拟合问题。在几个基准数据集上，我们的方法达到了最先进的精度，特别是对于具有挑战性的阶跃不平衡情况。

Deep convolutional networks have recently achieved great success in video recognition, yet their practical realization remains a challenge due to the large amount of computational resources required to achieve robust recognition. Motivated by the effectiveness of quantization for boosting efficiency, in this paper, we propose a dynamic network quantization framework, that selects optimal precision for each frame conditioned on the input for efficient video recognition. Specifically, given a video clip, we train a very lightweight network in parallel with the recognition network, to produce a dynamic policy indicating which numerical precision to be used per frame in recognizing videos. We train both networks effectively using standard backpropagation with a loss to achieve both competitive performance and resource efficiency required for video recognition. Extensive experiments on four challenging diverse benchmark datasets demonstrate that our proposed approach provides significant savings in computation and memory usage while outperforming the existing state-of-the-art methods. Project page: <https://cs-people.bu.edu/sunxm/videoIQ/project.html>.

深度卷积网络最近在视频识别方面取得了巨大的成功，但由于实现鲁棒识别需要大量的计算资源，因此其实现仍然是一个挑战。基于量化对提高效率的有效性，在本文中，我们提出了一种动态网络量化框架，该框架根据输入为每个帧选择最佳精度，以实现高效的视频识别。具体地说，给定一个视频片段，我们训练一个与识别网络并行的非常轻量级的网络，以生成一个动态策略，指示在识别视频时每帧使用的数值精度。我们使用标准的有损反向传播有效地训练这两个网络，以实现视频识别所需的竞争性能和资源效率。在四个具有挑战性的不同基准数据集上进行的大量实验表明，我们提出的方法在计算和内存使用方面提供了显著的节省，同时优于现有的最新方法。项目页面：<https://cs-people.bu.edu/sunxm/VideoIQ/project.html>。

Transformer-based detector is a new paradigm in object detection, which aims to achieve pretty-well performance while eliminates the priori knowledge driven components, e.g., anchors, proposals and the NMS. DETR, the state-of-the-art model among them, is composed of three sub-modules, i.e., a CNN-based backbone and paired transformer encoder-decoder. The CNN is applied to extract local features and the transformer is used to capture global contexts. This pipeline, however, is not concise enough. In this paper, we propose WB-DETR (DETR-based detector without Backbone) to prove that the reliance on CNN features extraction for a transformer-based detector is not necessary. Unlike the original DETR, WB-DETR is composed of only an encoder and a decoder without CNN backbone. For an input image, WB-DETR serializes it directly to encode the local features into each individual token. To make up the deficiency of transformer in modeling local information, we design an LIE-T2T (local information enhancement tokens to token) module to enhance the internal information of tokens after unfolding. Experimental results demonstrate that WB-DETR, the first pure-transformer detector without CNN to our knowledge, yields on par accuracy and faster inference speed with only half number of parameters compared with DETR baseline.

基于转换器的检测器是一种新的目标检测范式，其目的是在消除先验知识驱动的组件（如锚定、建议和NMS）的同时实现良好的性能。DETR是其中最先进的模型，由三个子模块组成，即基于CNN的主干和成对变压器编码器-解码器。CNN用于提取局部特征，transformer用于捕获全局上下文。然而，这条管道不够简洁。在本文中，我们提出WB-DETR（基于DETR的无主干检测器）来证明基于变压器的检测器不需要依赖CNN特征提取。与原始DETR不同，WB-DETR仅由一个编码器和一个解码器组成，没有CNN主干。对于输入图像，WB-DETR将其直接序列化，以将局部特征编码到每个单独的标记中。为了弥补transformer在局部信息建模方面的不足，我们设计了一个LIE-T2T（local information enhancement tokens To token，局部信息增强token To token）模块来增强令牌展开后的内部信息。实验结果表明，WB-DETR，第一个纯粹的变压器检测器没有美国有线电视新闻网，我们的知识，在PAR准确度和更快的推断速度，只有一半的参数与DETR基线相比。

Deep learning algorithms have made significant progress in dynamic scene deblurring. However, several challenges are still unsettled: 1) The degree and scale of blur in different regions of a blurred image can have a considerable variation in a large range. However, the traditional input pyramid or downscaling-upscaling, is designed to have limited and inflexible perceptual variousness to cope with large blur scale variation. 2) The nonlocal block is proved to be effective in the image enhancement tasks, but it requires high computation and memory cost. In this paper, we are the first to propose a light-weight globally-analyzing module into the image deblurring field, named Light Global Context Refinement (LGCR) module. with exponentially lower cost, it achieves even better performance than the nonlocal unit. Moreover, we propose the Perceptual Variousness Block (PVB) and PVB-piling strategy. By placing PVB repeatedly, the whole method possesses abundant reception field spectrum to be aware of the blur with various degrees and scales. Comprehensive experimental results from the different benchmarks and assessment metrics show that our method achieves excellent performance to set a new state-of-the-art in motion deblurring.

深度学习算法在动态场景去模糊方面取得了重大进展。然而，有几个挑战尚未解决：1) 模糊图像不同区域的模糊程度和规模可能在很大范围内有相当大的变化。然而，传统的输入金字塔或降尺度-升尺度设计为具有有限且不灵活的感知多样性，以应对较大的模糊尺度变化。2) 非局部块被证明在图像增强任务中是有效的，但它需要较高的计算和存储成本。在本文中，我们首次在图像去模糊领域提出了一个轻量级的全局分析模块，称为轻型全局上下文细化 (LGCR) 模块。它以指数级的低成本实现了比非本地单元更好的性能。此外，我们还提出了感知变异块 (PVB) 和PVB堆积策略。通过反复放置PVB，整个方法具有丰富的接收场频谱，能够感知不同程度和尺度的模糊。来自不同基准和评估指标的综合实验结果表明，我们的方法取得了优异的性能，建立了一个新的运动去模糊技术。

In this work, we propose a camera self-calibration algorithm for generic cameras with arbitrary non-linear distortions. We jointly learn the geometry of the scene and the accurate camera parameters without any calibration objects. Our camera model consists of a pinhole model, a fourth order radial distortion, and a generic noise model that can learn arbitrary non-linear camera distortions. While traditional self-calibration algorithms mostly rely on geometric constraints, we additionally incorporate photometric consistency. This requires learning the geometry of the scene, and we use Neural Radiance Fields (NeRF). We also propose a new geometric loss function, viz., projected ray distance loss, to incorporate geometric consistency for complex non-linear camera models. We validate our approach on standard real image datasets and demonstrate that our model can learn the camera intrinsics and extrinsics (pose) from scratch without COLMAP initialization. Also, we show that learning accurate camera models in a differentiable manner allows us to improve PSNR over baselines.

在这项工作中，我们提出了一种适用于具有任意非线性畸变的普通摄像机的摄像机自标定算法。我们在没有任何校准对象的情况下共同学习场景的几何体和精确的相机参数。我们的相机模型由针孔模型、四阶径向畸变和可学习任意非线性相机畸变的通用噪声模型组成。虽然传统的自校准算法主要依赖于几何约束，但我们还加入了光度一致性。这需要学习场景的几何结构，我们使用神经辐射场 (NeRF)。我们还提出了一个新的几何损失函数，即：。投影光线距离损失，以合并复杂非线性相机模型的几何一致性。我们在标准真实图像数据集上验证了我们的方法，并证明了我们的模型可以从头开始学习相机的内部和外部（姿势），而无需COLMAP初始化。此外，我们还表明，以可微的方式学习精确的相机模型可以使我们在基线上提高峰值信噪比。

Even though CCTV cameras are widely deployed for traffic surveillance and have therefore the potential of becoming cheap automated sensors for traffic speed analysis, their large-scale usage toward this goal has not been reported yet. A key difficulty lies in fact in the camera calibration phase. Existing state-of-the-art methods perform the calibration using image processing or keypoint detection techniques that require high-quality video streams, yet typical CCTV footage is low-resolution and noisy. As a result, these methods largely fail in real-world conditions. In contrast, we propose two novel calibration techniques whose only inputs come from an off-the-shelf object detector. Both methods consider multiple detections jointly, leveraging the fact that cars have similar and well-known 3D shapes with normalized dimensions. The first one is based on minimizing an energy function corresponding to a 3D reprojection error, the second one instead learns from synthetic training data to predict the scene geometry directly. Noticing the lack of speed estimation benchmarks faithfully reflecting the actual quality of surveillance cameras, we introduce a novel dataset collected from public CCTV streams. Experimental results conducted on three diverse benchmarks demonstrate excellent speed estimation accuracy that could enable the wide use of CCTV cameras for traffic analysis, even in challenging conditions where state-of-the-art methods completely fail. Additional information can be found on our project web page: <https://rebrand.ly/nle-cctv>

尽管闭路电视摄像机被广泛用于交通监控，因此有可能成为廉价的交通速度分析自动传感器，但它们在实现这一目标方面的大规模使用尚未报道。实际上，一个关键的困难在于摄像机校准阶段。现有最先进的方法使用需要高质量视频流的图像处理或关键点检测技术执行校准，但典型的CCTV镜头分辨率低且噪声大。因此，这些方法在现实环境中基本上是失败的。相比之下，我们提出了两种新的校准技术，其唯一输入来自现成的目标检测器。这两种方法共同考虑多个检测，利用汽车具有相似的和众所周知的具有标准化尺寸的3D形状的事实。第一种方法基于最小化与3D重投影误差相对应的能量函数，第二种方法从合成训练数据中学习，直接预测场景几何。注意到缺乏真实反映监控摄像机实际质量的速度估计基准，我们引入了一个从公共CCTV流中收集的新数据集。在三个不同基准上进行的实验结果表明，即使在最先进的方法完全失败的具有挑战性的条件下，速度估计精度也非常高，可以广泛使用闭路电视摄像机进行交通分析。更多信息可在我们的项目网页上找到：<https://rebrand.ly/nle-cctv>

Most existing brain tumor segmentation methods usually exploit multi-modal magnetic resonance imaging (MRI) images to achieve high segmentation performance. However, the problem of missing certain modality images often happens in clinical practice, thus leading to severe segmentation performance degradation. In this work, we propose a Region-aware Fusion Network (RFNet) that is able to exploit different combinations of multi-modal data adaptively and effectively for tumor segmentation. Considering different modalities are sensitive to different brain tumor regions, we design a Region-aware Fusion Module (RFM) in RFNet to conduct modal feature fusion from available image modalities according to disparate regions. Benefiting from RFM, RFNet can adaptively segment tumor regions from an incomplete set of multi-modal images by effectively aggregating modal features. Furthermore, we also develop a segmentation-based regularizer to prevent RFNet from the insufficient and unbalanced training caused by the incomplete multi-modal data. Specifically, apart from obtaining segmentation results from fused modal features, we also segment each image modality individually from the corresponding encoded features. In this manner, each modal encoder is forced to learn discriminative features, thus improving the representation ability of the fused features. Remarkably, extensive experiments on BRATS2020, BRATS2018 and BRATS2015 datasets demonstrate that our RFNet outperforms the state-of-the-art significantly.

现有的大多数脑肿瘤分割方法通常利用多模态磁共振成像（MRI）图像来实现高分割性能。然而，在临床实践中经常会出现丢失某些模态图像的问题，从而导致分割性能严重下降。在这项工作中，我们提出了一种区域感知融合网络（RFNet），该网络能够自适应和有效地利用多模态数据的不同组合进行肿瘤分割。考虑到不同的模式对不同的脑肿瘤区域敏感，我们在RFNet中设计了一个区域感知融合模块

（RFM），根据不同的区域对可用的图像模式进行模式特征融合。得益于RFM，RFNet可以通过有效地聚集模态特征，从不完整的多模态图像集中自适应地分割肿瘤区域。此外，我们还开发了一种基于分段的正则化器，以防止RFNet由于不完整的多模态数据而导致训练不足和不平衡。具体地说，除了从融合的模态特征中获得分割结果外，我们还从相应的编码特征中分别分割每个图像模态。以这种方式，每个模态编码器被迫学习鉴别特征，从而提高融合特征的表示能力。值得注意的是，在BRATS2020、BRATS2018和BRATS2015数据集上进行的大量实验表明，我们的RFNet显著优于最新技术。

The mainstream image captioning models rely on Convolutional Neural Network (CNN) image features to generate captions via recurrent models. Recently, image scene graphs have been used to augment captioning models so as to leverage their structural semantics such as object entities, relationships and attributes. Several studies have noted that naive use of scene graphs from a black-box scene graph generator harms image captioning performance, and scene graph-based captioning models have to incur the overhead of explicit use of image features to generate decent captions. Addressing these challenges, we propose a framework, SG2Caps, that utilizes only the scene graph labels for competitive image captioning performance. The basic idea is to close the semantic gap between two scene graphs - one derived from the input image and the other one from its caption. In order to achieve this, we leverage the spatial location of objects and the Human-Object-Interaction (HOI) labels as an additional HOI graph. Our framework outperforms existing scene graph-only captioning models by a large margin indicating scene graphs as a promising representation for image captioning. Direct utilization of the scene graph labels avoids expensive graph convolutions over high-dimensional CNN features resulting in 49% fewer trainable parameters. The code is available at: <https://github.com/Kien085/SG2Caps>.

主流的图像字幕模型依靠卷积神经网络（CNN）图像特征通过递归模型生成字幕。最近，图像场景图被用于增强字幕模型，以便利用其结构语义，如对象实体、关系和属性。一些研究已经指出，从黑匣子场景图生成器中天真地使用场景图会损害图像字幕性能，基于场景图的字幕模型必须产生显式使用图像特征以生成像样字幕的开销。为了应对这些挑战，我们提出了一个框架SG2Caps，该框架仅利用场景图标签来提高图像字幕的性能。其基本思想是弥合两个场景图之间的语义鸿沟——一个来自输入图像，另一个来自其标题。为了实现这一点，我们利用对象的空间位置和人机交互（HOI）标签作为附加的HOI图。我们的框架在很大程度上优于现有的仅场景图字幕模型，表明场景图是一种很有前景的图像字幕表示。直接使用场景图标签可以避免在高维CNN特征上进行昂贵的图卷积，从而减少49%的可训练参数。该代码可从以下网址获取：<https://github.com/Kien085/SG2Caps>。

We propose Rank & Sort (RS) Loss, a ranking-based loss function to train deep object detection and instance segmentation methods (i.e. visual detectors). RS Loss supervises the classifier, a sub-network of these methods, to rank each positive above all negatives as well as to sort positives among themselves with respect to (wrt.) their localisation qualities (e.g. Intersection-over-Union - IoU). To tackle the non-differentiable nature of ranking and sorting, we reformulate the incorporation of error-driven update with backpropagation as Identity Update, which enables us to model our novel sorting error among positives. With RS Loss, we significantly simplify training: (i) Thanks to our sorting objective, the positives are prioritized by the classifier without an additional auxiliary head (e.g. for centerness, IoU, mask-IoU), (ii) due to its ranking-based nature, RS Loss is robust to class imbalance, and thus, no sampling heuristic is required, and (iii) we address the multi-task nature of visual detectors using tuning-free task-balancing coefficients. Using RS Loss, we train seven diverse visual detectors only by tuning the learning rate, and show that it consistently outperforms baselines: e.g. our RS Loss improves (i) Faster R-CNN by 3 box AP and aLRP Loss (ranking-based baseline) by 2 box AP on COCO dataset, (ii) Mask R-CNN with repeat factor sampling (RFS) by 3.5 mask AP (7 AP for rare classes) on LVIS dataset; and also outperforms all counterparts. Code is available at: <https://github.com/kemaloksuz/RankSortLoss>.

我们提出了秩和排序 (RS) 损失，这是一种基于排名的损失函数，用于训练深度目标检测和实例分割方法（即视觉检测器）。RS Loss监督分类器，即这些方法的子网络，将每个正的排序置于所有负的之上，并根据 (wrt) 对正的排序它们的定位特性（例如，在联合上的交叉-IoU）。为了解决排序和排序的不可微性，我们将错误驱动的更新与反向传播合并为身份更新，这使我们能够对新的排序错误进行建模。有了RS损失，我们大大简化了培训：(i) 由于我们的分类目标，分类器在没有额外辅助头的情况下（例如，对于中心度、IoU、掩码IoU）对积极因素进行优先排序，(ii) 由于RS损失的排名性质，它对类别不平衡具有鲁棒性，因此，不需要任何抽样启发，(iii) 我们使用无需调整的任务平衡系数来解决视觉检测器的多任务性质。使用RS损失，我们仅通过调整学习率来训练七种不同的视觉检测器，并表明其始终优于基线：例如，我们的RS损失在COCO数据集上提高了(i) 更快的R-CNN 3框AP和aLRP损失（基于排名的基线）2框AP，(ii) 通过LVIS数据集上的3.5掩码AP（稀有类为7 AP），使用重复因子抽样(RFS) 屏蔽R-CNN；而且表现也优于所有对手。代码可从以下网址获取：<https://github.com/kemaloksuz/RankSortLoss>。

How can we animate 3D-characters from a movie script or move robots by simply telling them what we would like them to do?" How unstructured and complex can we make a sentence and still generate plausible movements from it?" These are questions that need to be answered in the long-run, as the field is still in its infancy. Inspired by these problems, we present a new technique for generating compositional actions, which handles complex input sentences. Our output is a 3D pose sequence depicting the actions in the input sentence. We propose a hierarchical two-stream sequential model to explore a finer joint-level mapping between natural language sentences and 3D pose sequences corresponding to the given motion. We learn two manifold representations of the motion, one each for the upper body and the lower body movements. Our model can generate plausible pose sequences for short sentences describing single actions as well as long complex sentences describing multiple sequential and compositional actions. We evaluate our proposed model on the publicly available KIT Motion-Language Dataset containing 3D pose data with human-annotated sentences. Experimental results show that our model advances the state-of-the-art on text-based motion synthesis in objective evaluations by a margin of 50%. Qualitative evaluations based on a user study indicate that our synthesized motions are perceived to be the closest to the ground-truth motion captures for both short and compositional sentences.

我们如何通过简单地告诉3D角色我们希望他们做什么来制作电影脚本中的3D角色动画或移动机器人？“我们能造出一个多么无结构和复杂的句子，并且还能从中产生似是而非的动作？”这些都是长期需要回答的问题，因为该领域仍处于初级阶段。受这些问题的启发，我们提出了一种生成合成动作的新技术，用于处理复杂的输入句子。我们的输出是一个3D姿势序列，描述输入句子中的动作。我们提出了一个层次化的两流序列模型，以探索自然语言句子和对应于给定运动的三维姿势序列之间更精细的关节级映射。我们学习两种运动的流形表示，一种是上半身运动，另一种是下半身运动。我们的模型可以为描述单个动作的短句以及描述多个连续动作和合成动作的长句生成合理的姿势序列。我们在公开的KIT运动语言数据集上评估了我们提出的模型，该数据集包含带有人类注释句子的3D姿势数据。实验结果表明，我们的模型将基于文本的运动合成在客观评价方面的最新进展提高了50%。基于用户研究的定性评估表明，对于短句和合成句，我们的合成运动被认为是最接近地面真实运动捕捉的。

Domain Adaptive Object Detection (DAOD) relieves the reliance on large-scale annotated data by transferring the knowledge learned from a labeled source domain to a new unlabeled target domain. Recent DAOD approaches resort to local feature alignment in virtue of domain adversarial training in conjunction with the ad-hoc detection pipelines to achieve feature adaptation. However, these methods are limited to adapt the specific types of object detectors and do not explore the cross-domain topological relations. In this paper, we first formulate DAOD as an open-set domain adaptation problem in which foregrounds (pixel or region) can be seen as the "known class", while backgrounds (pixel or region) are referred to as the "unknown class". To this end, we present a new and general perspective for DAOD named Dual Bipartite Graph Learning (DBGL), which captures the cross-domain interactions on both pixel-level and semantic-level via increasing the distinction between foregrounds and backgrounds and modeling the cross-domain dependencies among different semantic categories. Experiments reveal that the proposed DBGL in conjunction with one-stage and two-stage detectors exceeds the state-of-the-art performance on standard DAOD benchmarks.

域自适应目标检测 (DAOD) 通过将从标记的源域学习到的知识转移到新的未标记的目标域，减轻了对大规模标注数据的依赖。最近的DAOD方法借助于域对抗性训练和自组织检测管道，借助于局部特征对齐来实现特征自适应。然而，这些方法仅限于适应特定类型的对象检测器，并且没有探索跨域拓扑关系。在本文中，我们首先将DAOD描述为一个开放集域自适应问题，其中前景（像素或区域）可视为“已知类”，而背景（像素或区域）可视为“未知类”。为此，我们提出了DAOD的一个新的、通用的视角，称为双二部图学习 (DBGL)，它通过增加前景和背景之间的区别以及建模不同语义类别之间的跨域依赖关系，捕获像素级和语义级上的跨域交互。实验表明，所提出的DBGL结合一级和两级检测器，在标准DAOD基准上超过了最先进的性能。

In this paper, we propose Parametric Contrastive Learning (PaCo) to tackle long-tailed recognition. Based on theoretical analysis, we observe supervised contrastive loss tends to bias on high-frequency classes and thus increases the difficulty of imbalanced learning. We introduce a set of parametric class-wise learnable centers to rebalance from an optimization perspective. Further, we analyze our PaCo loss under a balanced setting. Our analysis demonstrates that PaCo can adaptively enhance the intensity of pushing samples of the same class close as more samples are pulled together with their corresponding centers and benefit hard example learning. Experiments on long-tailed CIFAR, ImageNet, Places, and iNaturalist 2018 manifest the new state-of-the-art for long-tailed recognition. On full ImageNet, models trained with PaCo loss surpass supervised contrastive learning across various ResNet backbones, e.g., our ResNet-200 achieves 81.8% top-1 accuracy. Our code is available at <https://github.com/dvlab-research/Parametric-Contrastive-Learning>.

在本文中，我们提出了参数对比学习（PaCo）来解决长尾识别问题。基于理论分析，我们发现监督对比缺失倾向于偏向高频课堂，从而增加了不平衡学习的难度。我们引入一组参数化的类学习中心，从优化的角度重新平衡。此外，我们还分析了平衡设置下的PaCo损失。我们的分析表明，当更多样本与其对应的中心拉在一起时，PaCo可以自适应地增强将同一类样本推近的强度，并有利于硬示例学习。在长尾CIFAR、ImageNet、Places和iNaturalist 2018上的实验表明了长尾识别的新技术。在full ImageNet上，使用PaCo loss训练的模型超过了各种ResNet主干的监督对比学习，例如，我们的ResNet-200达到81.8%的top-1精度。我们的代码可在<https://github.com/dvlab-research/Parametric-Contrastive-Learning>。

A recent trend in computer vision is to replace convolutions with transformers. However, the performance gain of transformers is attained at a steep cost, requiring GPU years and hundreds of millions of samples for training. This excessive resource usage compensates for a misuse of transformers: Transformers densely model relationships between its inputs -- ideal for late stages of a neural network, when concepts are sparse and spatially-distant, but extremely inefficient for early stages of a network, when patterns are redundant and localized. To address these issues, we leverage the respective strengths of both operations, building convolution-transformer hybrids. Critically, in sharp contrast to pixel-space transformers, our Visual Transformer (VT) operates in a semantic token space, judiciously attending to different image parts based on context. Our VTs significantly outperforms baselines: On ImageNet, our VT-ResNets outperform convolution-only ResNet by 4.6 to 7 points and transformer-only ViT-B by 2.6 points with 2.5 times fewer FLOPs, 2.1 times fewer parameters. For semantic segmentation on LIP and COCO-stuff, VT-based feature pyramid networks (FPN) achieve 0.35 points higher mIoU while reducing the FPN module's FLOPs by 6.5x.

计算机视觉的一个最新趋势是用变压器代替卷积。然而，变压器的性能增益是以高昂的成本实现的，需要GPU年和数亿个样本进行培训。这种过度的资源使用弥补了变压器的误用：变压器对其输入之间的关系进行了密集的建模——非常适合神经网络的后期阶段，当概念稀疏且空间距离较远时，但对于网络的早期阶段，当模式冗余且局部时，效率极低。为了解决这些问题，我们利用两种操作各自的优势，构建卷积变压器混合。关键的是，与像素空间变换器形成鲜明对比的是，我们的视觉变换器（VT）在语义标记空间中运行，根据上下文明智地处理不同的图像部分。我们的VTs显著优于基线：在ImageNet上，我们的VT ResNet优于仅卷积ResNet 4.6到7个点，仅变压器ViT-B 2.6个点，触发器减少2.5倍，参数减少2.1倍。对于LIP和COCO内容的语义分割，基于VT的特征金字塔网络（FPN）实现了高0.35点的mIoU，同时将FPN模块的失败次数减少了6.5倍。

Recent progress in 3D object detection from single images leverages monocular depth estimation as a way to produce 3D pointclouds, turning cameras into pseudo-lidar sensors. These two-stage detectors improve with the accuracy of the intermediate depth estimation network, which can itself be improved without manual labels via large-scale self-supervised learning. However, they tend to suffer from overfitting more than end-to-end methods, are more complex, and the gap with similar lidar-based detectors remains significant. In this work, we propose an end-to-end, single stage, monocular 3D object detector, DD3D, that can benefit from depth pre-training like pseudo-lidar methods, but without their limitations. Our architecture is designed for effective information transfer between depth estimation and 3D detection, allowing us to scale with the amount of unlabeled pre-training data. Our method achieves state-of-theart results on two challenging benchmarks, with 16.34% and 9.28% AP for Cars and Pedestrians (respectively) on the KITTI-3D benchmark, and 41.5% mAP on NuScenes.

从单个图像进行三维目标检测的最新进展利用单目深度估计作为生成三维点云的一种方法，将相机转变为伪激光雷达传感器。这些两级检测器提高了中间深度估计网络的精度，而中间深度估计网络本身可以通过大规模自监督学习在无需手动标记的情况下进行改进。然而，与端到端方法相比，它们更容易受到过度拟合的影响，更为复杂，与类似的基于激光雷达的探测器之间的差距仍然很大。在这项工作中，我们提出了一种端到端、单级、单目3D目标探测器DD3D，它可以像伪激光雷达方法一样受益于深度预训练，但没有其局限性。我们的架构设计用于深度估计和3D检测之间的有效信息传输，允许我们根据未标记的预训练数据量进行缩放。我们的方法在两个具有挑战性的基准上实现了最先进的结果，在KITTI-3D基准上汽车和行人的AP分别为16.34%和9.28%，在NuScenes上为41.5%。

weakly supervised object localization (WSOL) is a challenging problem when given image category labels but requires to learn object localization models. Optimizing a convolutional neural network (CNN) for classification tends to activate local discriminative regions while ignoring complete object extent, causing the partial activation issue. In this paper, we argue that partial activation is caused by the intrinsic characteristics of CNN, where the convolution operations produce local receptive fields and experience difficulty to capture long-range feature dependency among pixels. We introduce the token semantic coupled attention map (TS-CAM) to take full advantage of the self-attention mechanism in visual transformer for long-range dependency extraction. TS-CAM first splits an image into a sequence of patch tokens for spatial embedding, which produce attention maps of long-range visual dependency to avoid partial activation. TS-CAM then re-allocates category-related semantics for patch tokens, enabling each of them to be aware of object categories. TS-CAM finally couples the patch tokens with the semantic-agnostic attention map to achieve semantic-aware localization. Experiments on the ILSVRC/CUB-200-2011 datasets show that TS-CAM outperforms its CNN-CAM counterparts by 7.1%/27.1% for WSOL, achieving state-of-the-art performance. Code is available at <https://github.com/vasgaowei/TS-CAM>

当给定图像类别标签但需要学习目标定位模型时，弱监督目标定位（WSOL）是一个具有挑战性的问题。优化用于分类的卷积神经网络（CNN）倾向于激活局部判别区域，而忽略完整的对象范围，从而导致部分激活问题。在本文中，我们认为部分激活是由CNN的固有特性引起的，其中卷积运算产生局部感受野，并且难以捕获像素之间的长距离特征依赖性。我们引入标记语义耦合注意图（TS-CAM）来充分利用视觉转换器中的自我注意机制进行远程依赖提取。TS-CAM首先将图像分割为一系列用于空间嵌入的补丁标记，这些标记生成具有远程视觉依赖性的注意图，以避免部分激活。TS-CAM然后为补丁令牌重新分配与类别相关的语义，使每个令牌都能够知道对象类别。TS-CAM最终将补丁标记与语义不可知注意图耦合，以实现语义感知定位。在ILSVRC/CUB-200-2011数据集上的实验表明，TS-CAM比CNN-CAM的性能高7%。WSOL为1%/27.1%，达到最先进的性能。代码可在<https://github.com/vasgaowei/TS-CAM>

Image view synthesis has seen great success in reconstructing photorealistic visuals, thanks to deep learning and various novel representations. The next key step in immersive virtual experiences is view synthesis of dynamic scenes. However, several challenges exist due to the lack of high-quality training datasets, and the additional time dimension for videos of dynamic scenes. To address this issue, we introduce a multi-view video dataset, captured with a custom 10-camera rig in 120FPS. The dataset contains 96 high-quality scenes showing various visual effects and human interactions in outdoor scenes. We develop a new algorithm, Deep 3D Mask Volume, which enables temporally-stable view extrapolation from binocular videos of dynamic scenes, captured by static cameras. Our algorithm addresses the temporal inconsistency of disocclusions by identifying the error-prone areas with a 3D mask volume, and replaces them with static background observed throughout the video. Our method enables manipulation in 3D space as opposed to simple 2D masks. We demonstrate better temporal stability than frame-by-frame static view synthesis methods, or those that use 2D masks. The resulting view synthesis videos show minimal flickering artifacts and allow for larger translational movements.

由于深入的学习和各种新颖的表示，图像视图合成在重建照片级真实感视觉效果方面取得了巨大成功。沉浸式虚拟体验的下一个关键步骤是动态场景的视图合成。然而，由于缺乏高质量的训练数据集，以及动态场景视频的额外时间维度，存在一些挑战。为了解决这个问题，我们引入了一个多视图视频数据集，它是用一个定制的10摄像头设备以120FPS的速度拍摄的。该数据集包含96个高质量场景，显示了户外场景中的各种视觉效果和人类交互。我们开发了一种新的算法，深三维掩模体积，它可以从静态摄像机捕获的动态场景的双目视频中进行时间稳定的视图外推。我们的算法通过用3D掩模体积识别容易出错的区域，并用在整个视频中观察到的静态背景替换它们，从而解决了不一致性的时间问题。我们的方法可以在3D空间进行操作，而不是简单的2D遮罩。我们展示了比逐帧静态视图合成方法或使用2D遮罩的方法更好的时间稳定性。由此产生的视图合成视频显示最小的闪烁伪影，并允许更大的平移运动。

Building an interactive artificial intelligence that can ask questions about the real world is one of the biggest challenges for vision and language problems. In particular, goal-oriented visual dialogue, where the aim of the agent is to seek information by asking questions during a turn-taking dialogue, has been gaining scholarly attention recently. While several existing models based on the Guesswhat?! dataset have been proposed, the Questioner typically asks simple category-based questions or absolute spatial questions. This might be problematic for complex scenes where the objects share attributes or in cases where descriptive questions are required to distinguish objects. In this paper, we propose a novel Questioner architecture, called Unified Questioner Transformer (UniQer), for descriptive question generation with referring expressions. In addition, we build a goal-oriented visual dialogue task called CLEVR Ask. It synthesizes complex scenes that require the Questioner to generate descriptive questions. We train our model with two variants of CLEVR Ask datasets. The results of the quantitative and qualitative evaluations show that UniQer outperforms the baseline.

构建一个能够询问真实世界问题的交互式人工智能是视觉和语言问题面临的最大挑战之一。特别是，目标导向的视觉对话，即代理的目的是通过在轮流对话中提问来寻求信息，最近受到了学术界的关注。而现有的几种模型基于猜测什么？！在提出数据集时，提问者通常会提出简单的基于类别的问题或绝对空间问题。对于对象共享属性的复杂场景，或者在需要描述性问题来区分对象的情况下，这可能会有问题。在本文中，我们提出了一种新的提问器架构，称为统一提问器转换器（UniQer），用于使用引用表达式生成描述性问题。此外，我们还构建了一个面向目标的视觉对话任务，称为CLEVR Ask。它综合了需要提问者生成描述性问题的复杂场景。我们使用两种不同的CLEVR Ask数据集来训练我们的模型。定量和定性评估的结果表明，UniQer优于基线。

Attention mechanism has demonstrated great potential in fine-grained visual recognition tasks. In this paper, we present a counterfactual attention learning method to learn more effective attention based on causal inference. Unlike most existing methods that learn visual attention based on conventional likelihood, we propose to learn the attention with counterfactual causality, which provides a tool to measure the attention quality and a powerful supervisory signal to guide the learning process. Specifically, we analyze the effect of the learned visual attention on network prediction through counterfactual intervention and maximize the effect to encourage the network to learn more useful attention for fine-grained image recognition. Empirically, we evaluate our method on a wide range of fine-grained visual recognition tasks where attention plays a crucial role, including fine-grained image categorization, person re-identification, and vehicle re-identification. The consistent improvement on all benchmarks demonstrates the effectiveness of our method.

注意机制在细粒度视觉识别任务中显示出巨大的潜力。本文提出了一种基于因果推理的反事实注意学习方法来学习更有效的注意。与大多数现有的基于传统似然理论的视觉注意学习方法不同，我们提出用反事实因果关系来学习注意，它提供了一种测量注意质量的工具，并提供了一个强大的监督信号来指导学习过程。具体来说，我们通过反事实干预来分析学到的视觉注意对网络预测的影响，并最大限度地提高影响，以鼓励网络学习更多有用的注意，用于细粒度图像识别。根据经验，我们在广泛的细粒度视觉识别任务中评估了我们的方法，其中注意力起着至关重要的作用，包括细粒度图像分类、人员重新识别和车辆重新识别。所有基准的持续改进证明了我们方法的有效性。

Measuring similarity between two images often requires performing complex reasoning along different axes (e.g., color, texture, or shape). Insights into what might be important for measuring similarity can be provided by annotated attributes, but prior work tends to view these annotations as complete, resulting in them using a simplistic approach of predicting attributes on single images, which are, in turn, used to measure similarity. However, it is impractical for a dataset to fully annotate every attribute that may be important. Thus, only representing images based on these incomplete annotations may miss out on key information. To address this issue, we propose the Pairwise Attribute-informed similarity Network (PAN), which breaks similarity learning into capturing similarity conditions and relevance scores from a joint representation of two images. This enables our model to identify that two images contain the same attribute, but can have it deemed irrelevant (e.g., due to fine-grained differences between them) and ignored for measuring similarity between the two images. Notably, while prior methods of using attribute annotations are often unable to outperform prior art, PAN obtains a 4-9% improvement on compatibility prediction between clothing items on Polyvore Outfits, a 5% gain on few shot classification of images using Caltech-UCSD Birds (CUB), and over 1% boost to Recall@1 on In-Shop Clothes Retrieval.

测量两幅图像之间的相似性通常需要沿不同的轴（例如颜色、纹理或形状）执行复杂的推理。带注释的属性可以提供对测量相似性可能很重要的见解，但之前的工作倾向于将这些注释视为完整的，导致它们使用一种简单的方法来预测单个图像上的属性，而这些属性反过来又用于测量相似性。然而，对于数据集来说，完全注释每个可能重要的属性是不切实际的。因此，仅基于这些不完整注释表示图像可能会遗漏关键信息。为了解决这个问题，我们提出了成对属性通知相似性网络（PAN），它将相似性学习分解为从两幅图像的联合表示中获取相似性条件和相关性分数。这使我们的模型能够识别两个图像包含相同的属性，但可以将其视为不相关（例如，由于它们之间的细粒度差异），并在测量两个图像之间的相似性时将其忽略。值得注意的是，虽然使用属性注释的现有方法通常无法超越现有技术，但PAN在多食动物服装上的服装项目之间的兼容性预测方面获得了4-9%的改进，在使用加州理工大学UCSD Birds (CUB) 的图像的少镜头分类方面获得了5%的改进，并且在性能方面提高了1%以上Recall@1关于店内服装检索。

With the development of 3D scanning technologies, 3D vision tasks have become a popular research area. Owing to the large amount of data acquired by sensors, unsupervised learning is essential for understanding and utilizing point clouds without an expensive annotation process. In this paper, we propose a novel framework and an effective auto-encoder architecture named "PSG-Net" for reconstruction-based learning of point clouds. Unlike existing studies that used fixed or random 2D points, our framework generates input-dependent point-wise features for the latent point set. PSG-Net uses the encoded input to produce point-wise features through the seed generation module and extracts richer features in multiple stages with gradually increasing resolution by applying the seed feature propagation module progressively. We prove the effectiveness of PSG-Net experimentally; PSG-Net shows state-of-the-art performances in point cloud reconstruction and unsupervised classification, and achieves comparable performance to counterpart methods in supervised completion.

随着三维扫描技术的发展，三维视觉任务已经成为一个热门的研究领域。由于传感器采集大量数据，无监督学习对于理解和利用点云而无需昂贵的注释过程至关重要。在本文中，我们提出了一个新的框架和一个有效的自动编码器架构“PSG网”，用于基于重建的点云学习。与使用固定或随机2D点的现有研究不同，我们的框架为潜在点集生成与输入相关的逐点特征。PSG网络使用编码输入通过种子生成模块生成点特征，并通过逐步应用种子特征传播模块，在多个阶段以逐渐增加的分辨率提取更丰富的特征。实验证明了PSG网络的有效性；PSG网络在点云重建和无监督分类方面显示了最先进的性能，并在监督完成方面实现了与对应方法相当的性能。

Image hiding aims to hide a secret image into a cover image in an imperceptible way, and then recover the secret image perfectly at the receiver end. Capacity, invisibility and security are three primary challenges in image hiding task. This paper proposes a novel invertible neural network (INN) based framework, HiNet, to simultaneously overcome the three challenges in image hiding. For large capacity, we propose an inverse learning mechanism by simultaneously learning the image concealing and revealing processes. Our method is able to achieve the concealing of a full-size secret image into a cover image with the same size. For high invisibility, instead of pixel domain hiding, we propose to hide the secret information in wavelet domain. Furthermore, we propose a new low-frequency wavelet loss to constrain that secret information is hidden in high-frequency wavelet sub-bands, which significantly improves the hiding security. Experimental results show that our HiNet significantly outperforms other state-of-the-art image hiding methods, with more than 10 dB PSNR improvement in secret image recovery on ImageNet, COCO and DIV2K datasets. Codes are available at <https://github.com/TomTomTomi/HiNet>.

图像隐藏的目的是以不可察觉的方式将秘密图像隐藏到封面图像中，然后在接收端完美地恢复秘密图像。容量、不可见性和安全性是图像隐藏任务的三个主要挑战。本文提出了一种新的基于可逆神经网络（INN）的图像隐藏框架HiNet，以同时克服图像隐藏中的三个挑战。对于大容量图像，我们提出了一种同时学习图像隐藏和显示过程的逆学习机制。我们的方法能够实现将全尺寸秘密图像隐藏到相同尺寸的封面图像中。为了提高不可见性，我们提出在小波域隐藏秘密信息，而不是在像素域隐藏。此外，我们还提出了一种新的低频小波丢失算法来约束秘密信息隐藏在高频小波子带中，从而显著提高了隐藏的安全性。实验结果表明，在ImageNet、COCO和DIV2K数据集上，我们的HiNet显著优于其他最先进的图像隐藏方法，在秘密图像恢复方面的峰值信噪比提高了10 dB以上。代码可在<https://github.com/TomTomTomi/HiNet>。

A cornerstone of geometric reconstruction, rotation averaging seeks the set of absolute rotations that optimally explains a set of measured relative orientations between them. In spite of being an integral part of bundle adjustment and structure-from-motion, averaging rotations is both a nonconvex and high-dimensional optimization problem. In this paper, we address it from a maximum likelihood estimation standpoint and make a twofold contribution.

Firstly, we set forth a novel initialization-free primal-dual method which we show empirically to converge to the global optimum. Further, we derive what is to our knowledge, the first optimal closed-form solution for rotation averaging in cycle graphs and contextualize this result within spectral graph theory. Our proposed methods achieve a significant gain both in precision and performance.

旋转平均是几何重建的基石，它寻求一组绝对旋转，以最佳方式解释它们之间的一组测量相对方向。尽管是束调整和运动结构的一个组成部分，平均旋转是一个非凸和高维优化问题。在本文中，我们从最大似然估计的角度来处理它，并做出了双重贡献。首先，我们提出了一种新的无初始化原始-对偶方法，并通过实验证明了该方法收敛于全局最优解。此外，我们推导出了我们所知的循环图中旋转平均的第一个最优闭式解，并将此结果与谱图论联系起来。我们提出的方法在精度和性能上都取得了显著的提高。

Computing dense pixel-to-pixel image correspondences is a fundamental task of computer vision. Often, the objective is to align image pairs from the same semantic category for manipulation or segmentation purposes. Despite achieving superior performance, existing deep learning alignment methods cannot cluster images; consequently, clustering and pairing images needed to be a separate laborious and expensive step. Given a dataset with diverse semantic categories, we propose a multi-task model, Jim-Net, that can directly learn to cluster and align images without any pixel-level or image-level annotations. We design a pair-matching alignment unsupervised training algorithm that selectively matches and aligns image pairs from the clustering branch. Our unsupervised Jim-Net achieves comparable accuracy with state-of-the-art supervised methods on benchmark 2D image alignment dataset PF-PASCAL. Specifically, we apply Jim-Net to cryo-electron tomography, a revolutionary 3D microscopy imaging technique of native subcellular structures. After extensive evaluation on seven datasets, we demonstrate that Jim-Net enables systematic discovery and recovery of representative macromolecular structures *in situ*, which is essential for revealing molecular mechanisms underlying cellular functions. To our knowledge, Jim-Net is the first end-to-end model that can simultaneously align and cluster images, which significantly improves the performance as compared to performing each task alone.

计算密集的像素图像对应是计算机视觉的一项基本任务。通常，目标是为了操作或分割目的，对齐来自同一语义类别的图像对。尽管取得了优异的性能，但现有的深度学习对齐方法无法对图像进行聚类；因此，对图像进行聚类和配对需要一个单独的费力且昂贵的步骤。针对具有不同语义类别的数据集，我们提出了一个多任务模型Jim Net，该模型可以直接学习聚类和对齐图像，而无需任何像素级或图像级注释。我们设计了一种对匹配对齐的无监督训练算法，该算法对来自聚类分支的图像对进行选择性匹配和对齐。我们的无监督Jim网络在基准2D图像对齐数据集PF-PASCAL上实现了与最先进的监督方法相当的精度。具体而言，我们将Jim Net应用于冷冻电子断层扫描，这是一种革命性的天然亚细胞结构3D显微成像技术。在对七个数据集进行广泛评估后，我们证明Jim Net能够在原位系统地发现和恢复具有代表性的大分子结构，这对于揭示细胞功能的分子机制至关重要。据我们所知，Jim Net是第一个可以同时对齐和群集映像的端到端模型，与单独执行每个任务相比，它显著提高了性能。

Recently, deep learning-based image denoising methods have achieved significant improvements over traditional methods. Due to the hardware limitation, most deep learning-based image denoising methods utilize cropped small patches to train a convolutional neural network to infer the clean images. However, the real noisy images in practical are mostly of high resolution rather than the cropped small patches and the vanilla training strategies ignore the cross-patch contextual dependency in the whole image. In this paper, we propose Cross-Patch Net (CPNet), which is the first deep- learning-based real image denoising method for HR (high resolution) input. Furthermore, we design a novel loss guided by the noise level map to obtain better performance. Compared with the vanilla patch-based training strategies, our approach effectively exploits the cross-patch contextual dependency. effective method to generate realistic sRGB noisy images from their corresponding clean sRGB images for denoiser training. Denoising experiments on real-world sRGB images show the effectiveness of the proposed method. More importantly, our method achieves state-of-the-art performance on practical sRGB noisy image denoising.

近年来，基于深度学习的图像去噪方法与传统方法相比有了显著的改进。由于硬件的限制，大多数基于深度学习的图像去噪方法都是利用裁剪过的小块来训练卷积神经网络来推断干净的图像。然而，实际的噪声图像大多是高分辨率的，而不是裁剪过的小面片，而普通的训练策略忽略了整个图像中的跨面片上下文依赖性。在本文中，我们提出了交叉面片网（CPNet），这是第一个基于深度学习的真实图像去噪方法的HR（高分辨率）输入。此外，为了获得更好的性能，我们还设计了一种新的由噪声级图引导的损耗。与普通的基于补丁的训练策略相比，我们的方法有效地利用了跨补丁的上下文依赖性。从相应的干净sRGB图像生成真实sRGB噪声图像的有效方法，用于去噪训练。对真实sRGB图像的去噪实验表明了该方法的有效性。更重要的是，我们的方法在实际的sRGB噪声图像去噪方面达到了最先进的性能。

Co-occurrent visual pattern makes aggregating contextual information a common paradigm to enhance the pixel representation for semantic image segmentation. The existing approaches focus on modeling the context from the perspective of the whole image, i.e., aggregating the image-level contextual information. Despite impressive, these methods weaken the significance of the pixel representations of the same category, i.e., the semantic-level contextual information. To address this, this paper proposes to augment the pixel representations by aggregating the image-level and semantic-level contextual information, respectively. First, an image-level context module is designed to capture the contextual information for each pixel in the whole image. Second, we aggregate the representations of the same category for each pixel where the category regions are learned under the supervision of the ground-truth segmentation. Third, we compute the similarities between each pixel representation and the image-level contextual information, the semantic-level contextual information, respectively. At last, a pixel representation is augmented by weighted aggregating both the image-level contextual information and the semantic-level contextual information with the similarities as the weights. Integrating the image-level and semantic-level context allows this paper to report state-of-the-art accuracy on four benchmarks, i.e., ADE20K, LIP, COCOSTuff and Cityscapes.

共现视觉模式使得聚集上下文信息成为增强语义图像分割像素表示的常用范例。现有的方法侧重于从整个图像的角度对上下文进行建模，即聚合图像级上下文信息。尽管令人印象深刻，但这些方法削弱了同一类别像素表示的重要性，即语义级上下文信息。为了解决这个问题，本文提出通过分别聚合图像级和语义级上下文信息来增强像素表示。首先，设计了一个图像级上下文模块，用于捕获整个图像中每个像素的上下文信息。其次，我们为每个像素聚合相同类别的表示，其中类别区域在地面真值分割的监督下学习。第三，我们分别计算每个像素表示和图像级上下文信息、语义级上下文信息之间的相似度。最后，通过将图像级上下文信息和语义级上下文信息以相似度作为权重进行加权聚合，增强像素表示。结

合图像级和语义级上下文，本文可以报告四个基准的最先进的准确性，即ADE20K、LIP、COCOStuff和Cityscapes。

Detecting pedestrians and their associated faces jointly is a challenging task. On one hand, body or face could be absent because of occlusion or non-frontal human pose. On the other hand, the association becomes difficult or even miss-leading in crowded scenes due to the lack of strong correlational evidence. This paper proposes Body-Face Joint (BFJ) detector, a novel framework for detecting bodies and their faces with accurate correspondance. We follow the classical multi-class detector design by detecting body and face in parallel but with two key contributions. First, we propose an Embedding Matching Loss (EML) to learn an associative embedding for matching body and face of the same person. Second, we introduce a novel concept, "head hook", to bridge the gap of matching body and faces spatially. With the new semantical and geometrical sources of information, BFJ greatly reduces the difficulty of detecting body and face in pairs. Since the problem is unexplored yet, we design a new metric named log-average miss matching rate ( $mMR^{‐2}$ ) to evaluate the association performance and extend the CrowdHuman and CityPersons benchmarks by annotating each face box. Experiments show that our BFJ detector can maintain state-of-the-art performance in pedestrian detection on both one-stage and two-stage structures while greatly outperform various body-face association strategies. Code and datasets will be released soon.

联合检测行人及其相关人脸是一项具有挑战性的任务。一方面，由于遮挡或非正面人体姿势，身体或面部可能不存在。另一方面，由于缺乏强有力的相关证据，在拥挤的场景中，关联变得很困难，甚至错过了引导。本文提出了一种新的检测人体和人脸的框架——人体-人脸关节检测器（BFJ）。我们遵循经典的多类检测器设计，并行检测身体和面部，但有两个关键贡献。首先，我们提出了一种嵌入匹配损失（EML）来学习一种关联嵌入来匹配同一个人的身体和脸。其次，我们引入了一个新的概念“头钩”，以弥补匹配的身体和脸空间上的差距。借助新的语义和几何信息源，BFJ大大降低了成对检测身体和人脸的难度。由于这个问题尚未得到解决，我们设计了一个名为log average miss matching rate ( $mMR^{‐2}$ )的新指标来评估关联性能，并通过注释每个面框来扩展CrowdHuman和CityPersons基准。实验表明，我们的BFJ检测器可以在一级和两级结构上保持最先进的行人检测性能，同时大大优于各种身体-面部关联策略。代码和数据集将很快发布。

The crux of self-supervised video representation learning is to build general features from unlabeled videos. However, most recent works have mainly focused on high-level semantics and neglected lower-level representations and their temporal relationship which are crucial for general video understanding. To address these challenges, this paper proposes a multi-level feature optimization framework to improve the generalization and temporal modeling ability of learned video representations. Concretely, high-level features obtained from naive and prototypical contrastive learning are utilized to build distribution graphs, guiding the process of low-level and mid-level feature learning. We also devise a simple temporal modeling module from multi-level features to enhance motion pattern learning. Experiments demonstrate that multi-level feature optimization with the graph constraint and temporal modeling can greatly improve the representation ability in video understanding. Code is available at <https://github.com/shvdiwnkozbw/Video-Representation-via-Multi-level-Optimization>.

自监督视频表示学习的关键是从未标记的视频中建立一般特征。然而，最近的研究主要集中在高层语义上，而忽略了对一般视频理解至关重要的底层表示及其时间关系。为了应对这些挑战，本文提出了一个多层次的特征优化框架，以提高学习视频表示的泛化和时序建模能力。具体而言，利用原始对比学习和原型对比学习获得的高级特征构建分布图，指导低级和中级特征学习过程。我们还从多层次特征中设计了一个简单的时间建模模块来增强运动模式学习。实验表明，基于图约束和时态建模的多层次特征优化

可以极大地提高视频理解中的表示能力。代码可在<https://github.com/shvdiwnkozbw/Video-Representation-via-Multi-level-Optimization>。

Stereo-based 3D detection aims at detecting 3D object bounding boxes from stereo images using intermediate depth maps or implicit 3D geometry representations, which provides a low-cost solution for 3D perception. However, its performance is still inferior compared with LiDAR-based detection algorithms. To detect and localize accurate 3D bounding boxes, LiDAR-based models can encode accurate object boundaries and surface normal directions from LiDAR point clouds. However, the detection results of stereo-based detectors are easily affected by the erroneous depth features due to the limitation of stereo matching. To solve the problem, we propose LIGA-Stereo (LiDAR Geometry Aware Stereo Detector) to learn stereo-based 3D detectors under the guidance of high-level geometry-aware representations of LiDAR-based detection models. In addition, we found existing voxel-based stereo detectors failed to learn semantic features effectively from indirect 3D supervisions. We attach an auxiliary 2D detection head to provide direct 2D semantic supervisions. Experiment results show that the above two strategies improved the geometric and semantic representation capabilities. Compared with the state-of-the-art stereo detector, our method has improved the 3D detection performance of cars, pedestrians, cyclists by 10.44%, 5.69%, 5.97% mAP respectively on the official KITTI benchmark. The gap between stereo-based and LiDAR-based 3D detectors is further narrowed.

基于立体的三维检测旨在使用中间深度贴图或隐式三维几何表示从立体图像中检测三维对象边界框，这为三维感知提供了一种低成本的解决方案。然而，与基于激光雷达的检测算法相比，其性能仍然较差。为了检测和定位精确的三维边界框，基于激光雷达的模型可以从激光雷达点云编码精确的对象边界和曲面法线方向。然而，由于立体匹配的局限性，立体检测器的检测结果容易受到错误深度特征的影响。为了解决这个问题，我们提出LIGA-Stereo (LiDAR-Geometry-Aware-stereodetector) 在基于LiDAR的检测模型的高级几何感知表示的指导下学习基于立体的3D检测器。此外，我们发现现有的基于体素的立体检测器无法从间接的三维监控中有效地学习语义特征。我们附加了一个辅助的2D检测头来提供直接的2D语义监控。实验结果表明，上述两种策略提高了图像的几何和语义表示能力。与最先进的立体探测器相比，我们的方法在官方KITTI基准上分别提高了汽车、行人和自行车的3D检测性能10.44%、5.69%和5.97%。基于立体和基于激光雷达的3D探测器之间的差距进一步缩小。

Sparse voxel-based 3D convolutional neural networks (CNNs) are widely used for various 3D vision tasks. Sparse voxel-based 3D CNNs create sparse non-empty voxels from input point clouds and perform standard convolution operations on them only. We propose a simple and effective padding scheme --- interpolation-aware padding to pad a few empty voxels adjacent to the non-empty voxels and involving them in the CNN computation so that all neighboring voxels exist when computing point-wise features via the trilinear interpolation. For fine-grained 3D vision tasks where point-wise features are essential, like semantic segmentation and 3D detection, our network achieves higher prediction accuracy than the existing networks using the nearest neighbor interpolation or normalized trilinear interpolation with the zero-padding or the octree-padding scheme. Through extensive comparisons on various 3D segmentation and detection tasks, we demonstrate the superiority of 3D sparse CNNs with our sparse padding scheme in conjunction with feature interpolation.

基于稀疏体素的三维卷积神经网络 (CNN) 广泛应用于各种三维视觉任务。基于稀疏体素的三维CNN从输入点云创建稀疏非空体素，并仅对其执行标准卷积运算。我们提出了一种简单有效的填充方案——插值感知填充，将非空体素相邻的几个空体素填充到CNN计算中，使所有相邻体素在通过三线性插值计算逐点特征时都存在。对于细粒度的3D视觉任务，如语义分割和3D检测等关键的逐点特征，我们的网络实现了比现有网络更高的预测精度，使用最近邻插值或标准化三线性插值加零填充或八叉树填充方案。通

通过对各种3D分割和检测任务的广泛比较，我们证明了3D稀疏CNN与我们的稀疏填充方案结合特征插值的优越性。

Quantization Neural Networks (QNN) have attracted a lot of attention due to their high efficiency. To enhance the quantization accuracy, prior works mainly focus on designing advanced quantization algorithms but still fail to achieve satisfactory results under the extremely low-bit case. In this work, we take an architecture perspective to investigate the potential of high-performance QNN. Therefore, we propose to combine Network Architecture Search methods with quantization to enjoy the merits of the two sides. However, a naive combination inevitably faces unacceptable time consumption or unstable training problem. To alleviate these problems, we first propose the joint training of architecture and quantization with a shared step size to acquire a large number of quantized models. Then a bit-inheritance scheme is introduced to transfer the quantized models to the lower bit, which further reduces the time cost and meanwhile improves the quantization accuracy. Equipped with this overall framework, dubbed as Once Quantization-Aware Training (OQAT), our searched model family, OQATNets, achieves a new state-of-the-art compared with various architectures under different bit-widths. In particular, OQAT-2bit-M achieves 61.6% ImageNet Top-1 accuracy, outperforming 2-bit counterpart MobileNetV3 by a large margin of 9% with 10% less computation cost. A series of quantization-friendly architectures are identified easily and extensive analysis can be made to summarize the interaction between quantization and neural architectures. Codes and models are released at <https://github.com/LaVieEnRoseSMZ/OQA>

量化神经网络 (QNN) 因其高效性而受到广泛关注。为了提高量化精度，以往的工作主要集中在设计先进的量化算法上，但在极低比特率的情况下仍未能取得令人满意的结果。在这项工作中，我们从架构的角度来研究高性能QNN的潜力。因此，我们建议将网络架构搜索方法与量化相结合，以充分发挥两者的优点。然而，一个简单的组合不可避免地会面临不可接受的时间消耗或不稳定的训练问题。为了缓解这些问题，我们首先提出了共享步长的架构和量化联合训练，以获得大量量化模型。然后引入位继承机制将量化模型转移到较低的位，进一步降低了时间开销，同时提高了量化精度。我们的搜索模型系列OQATNets配备了这个整体框架，称为一次性量化感知训练 (OQAT)，与不同比特宽度下的各种体系结构相比，它实现了一种新的最先进水平。特别是，OQA-2bit-M实现了61.6%的ImageNet Top-1精度，比2位对应的MobileNetV3高出9%，计算成本降低了10%。一系列量化友好的结构很容易识别，并且可以进行广泛的分析来总结量化和神经结构之间的相互作用。代码和型号发布于<https://github.com/LaVieEnRoseSMZ/OQA>

Current state-of-the-art two-stage detectors generate oriented proposals through time-consuming schemes. This diminishes the detectors' speed, thereby becoming the computational bottleneck in advanced oriented object detection systems. This work proposes an effective and simple oriented object detection framework, termed Oriented R-CNN, which is a general two-stage oriented detector with promising accuracy and efficiency. To be specific, in the first stage, we propose an oriented Region Proposal Network (oriented RPN) that directly generates high-quality oriented proposals in a nearly cost-free manner. The second stage is oriented R-CNN head for refining oriented Regions of Interest (oriented RoIs) and recognizing them. Without tricks, oriented R-CNN with ResNet50 achieves state-of-the-art detection accuracy on two commonly-used datasets for oriented object detection including DOTA (75.87% mAP) and HRSC2016 (96.50% mAP), while having a speed of 15.1 FPS with the image size of 1024x1024 on a single RTX 2080Ti. We hope our work could inspire rethinking the design of oriented detectors and serve as a baseline for oriented object detection. Code is available at <https://github.com/jbwang1997/OBBDetection>.

当前最先进的两级检测器通过耗时的方案生成定向方案。这降低了检测器的速度，从而成为高级面向对象检测系统的计算瓶颈。本文提出了一种有效且简单的面向对象检测框架，称为面向R-CNN，它是一种通用的两级面向对象检测器，具有良好的准确性和效率。具体来说，在第一阶段，我们提出了一个面向区域的提案网络（oriented RPN），该网络以几乎无成本的方式直接生成高质量的面向区域的提案。第二阶段是定向R-CNN头部，用于细化定向感兴趣区域（定向ROI）并识别它们。在没有技巧的情况下，带ResNet50的oriented R-CNN在两个常用的面向对象检测数据集上实现了最先进的检测精度，包括DOTA（75.87%mAP）和HRSC2016（96.50%mAP），同时在单个RTX 2080Ti上的速度为15.1 FPS，图像大小为1024x1024。我们希望我们的工作能够启发人们重新思考面向对象检测器的设计，并作为面向对象检测的基线。代码位于<https://github.com/jbwang1997/obb检测>。

In this paper, we present a neat yet effective transformer-based framework for visual grounding, namely TransVG, to address the task of grounding a language query to the corresponding region onto an image. The state-of-the-art methods, including two-stage or one-stage ones, rely on a complex module with manually-designed mechanisms to perform the query reasoning and multi-modal fusion. However, the involvement of certain mechanisms in fusion module design, such as query decomposition and image scene graph, makes the models easily overfit to datasets with specific scenarios, and limits the plenitudinous interaction between the visual-linguistic context. To avoid this caveat, we propose to establish the multi-modal correspondence by leveraging transformers, and empirically show that the complex fusion modules (e.g., modular attention network, dynamic graph, and multi-modal tree) can be replaced by a simple stack of transformer encoder layers with higher performance. Moreover, we re-formulate the visual grounding as a direct coordinates regression problem and avoid making predictions out of a set of candidates (i.e., region proposals or anchor boxes). Extensive experiments are conducted on five widely used datasets, and a series of state-of-the-art records are set by our TransVG. We build the benchmark of transformer-based visual grounding framework and make the code available at <https://github.com/djiajunustc/TransVG>.

在本文中，我们提出了一个简洁而有效的基于转换器的可视接地框架，即TransVG，以解决将语言查询接地到图像上相应区域的任务。最先进的方法，包括两阶段或一阶段的方法，依赖于一个复杂的模块和手动设计的机制来执行查询推理和多模式融合。然而，在融合模块设计中，由于查询分解和图像场景图等机制的参与，使得模型很容易过度适应特定场景的数据集，限制了视觉语言环境之间的充分交互。为了避免这一警告，我们建议通过利用变压器建立多模态对应关系，并根据经验证明，复杂的融合模块（例如，模块化注意网络、动态图和多模态树）可以被具有更高性能的变压器-编码器层的简单堆栈所取代。此外，我们将视觉基础重新表述为一个直接坐标回归问题，并避免对一组候选对象（即区域建议或锚定框）进行预测。在五个广泛使用的数据集上进行了广泛的实验，我们的TransVG创造了一系列最先进的记录。我们构建了基于变压器的可视化接地框架的基准测试，并在<https://github.com/djiajunustc/TransVG>。

vision-and-language navigation (VLN) aims to enable embodied agents to navigate in realistic environments using natural language instructions. Given the scarcity of domain-specific training data and the high diversity of image and language inputs, the generalization of VLN agents to unseen environments remains challenging. Recent methods explore pretraining to improve generalization, however, the use of generic image-caption datasets or existing small-scale VLN environments is suboptimal and results in limited improvements. In this work, we introduce BnB, a large-scale and diverse in-domain VLN dataset. We first collect image-caption (IC) pairs from hundreds of thousands of listings from online rental marketplaces. Using IC pairs we next propose automatic strategies to generate millions of VLN path-instruction (PI) pairs. We further propose a shuffling loss that improves the learning of temporal order inside PI pairs. We use BnB to pretrain our Airbert model that can be adapted to discriminative and generative settings and show that it outperforms state of the art for Room-to-Room (R2R) navigation and Remote Referring Expression (REVERIE) benchmarks. Moreover, our in-domain pretraining significantly increases performance on a challenging few-shot VLN evaluation, where we train the model only on VLN instructions from a few houses.

视觉和语言导航 (VLN) 的目标是使嵌入式代理能够使用自然语言指令在现实环境中导航。由于缺乏特定领域的训练数据以及图像和语言输入的高度多样性，VLN代理在未知环境中的推广仍然具有挑战性。最近的方法探索预训练以提高泛化能力，然而，使用通用图像字幕数据集或现有的小规模VLN环境是次优的，并且导致有限的改进。在这项工作中，我们介绍了BnB，一个大规模的和多样化的域内VLN数据集。我们首先从在线租赁市场的数十万个列表中收集图像标题 (IC) 对。使用IC对，我们接下来提出自动生成数百万VLN路径指令 (PI) 对的策略。我们进一步提出了一种洗牌损失，它改进了PI对内时序的学习。我们使用BnB对我们的Airbert模型进行预训练，该模型可以适应区分性和生成性设置，并表明它在房间到房间 (R2R) 导航和远程引用表达 (幻想) 基准方面优于最先进的水平。此外，我们的域内预训练显著提高了具有挑战性的少数镜头VLN评估的性能，我们仅根据少数几家公司的VLN指令训练模型。