Pushing the Frontiers of Unconstrained Crowd Counting: New Dataset and Benchmark Method
Vishwanath A. Sindagi, Rajeev Yasarla, Vishal M. Patel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1221-1231
In this work, we propose a novel crowd counting network that progressively generates crowd density maps via residual error estimation. The proposed method uses VGG16 as the backbone network and employs density map generated by the final layer as a coarse prediction to refine and generate finer density maps in a progressive fashion using residual learning. Additionally, the residual learning is guided by an uncertainty-based confidence weighting mechanism that permits the flow of only high-confidence residuals in the refinement path. The proposed Confidence Guided Deep Residual Counting Network (CG-DRCN) is evaluated on recent complex datasets, and it achieves significant improvements in errors. Furthermore, we introduce a new large scale unconstrained crowd counting dataset (JHU-CROWD) that is 2.8 larger than the most recent crowd counting datasets in terms of the number of images. It contains 4,250 images with 1.11 million annotations. In comparison to existing datasets, the proposed dataset is collected under a variety of diverse scenarios and environmental conditions. Specifically, the dataset includes several images with weather-based degradations and illumination variations in addition to many distractor images, making it a very challenging dataset. Additionally, the dataset consists of rich annotations at both image-level and head-level. Several recent methods are evaluated and compared on this dataset.
****
Spatial Correspondence With Generative Adversarial Network: Learning Depth From Monocular Videos
Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, Lili Ju; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7494-7504
Depth estimation from monocular videos has important applications in many areas such as autonomous driving and robot navigation. It is a very challenging problem without knowing the camera pose since errors in camera-pose estimation can significantly affect the video-based depth estimation accuracy. In this paper, we present a novel SC-GAN network with end-to-end adversarial training for depth estimation from monocular videos without estimating the camera pose and pose change over time. To exploit cross-frame relations, SC-GAN includes a spatial correspondence module which uses Smolyak sparse grids to efficiently match the features across adjacent frames, and an attention mechanism to learn the importance of features in different directions. Furthermore, the generator in SC-GAN learns to estimate depth from the input frames, while the discriminator learns to distinguish between the ground-truth and estimated depth map for the reference frame. Experiments on the KITTI and Cityscapes datasets show that the proposed SC-GAN can achieve much more accurate depth maps than many existing state-of-the-art methods on monocular videos.
****
Learning Single Camera Depth Estimation Using Dual-Pixels
Rahul Garg, Neal Wadhwa, Sameer Ansari, Jonathan T. Barron; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7628-7637
Deep learning techniques have enabled rapid progress in monocular depth estimation, but their quality is limited by the ill-posed nature of the problem and the scarcity of high quality datasets. We estimate depth from a single cam-era by leveraging the dual-pixel auto-focus hardware that is increasingly common on modern camera sensors. Classic stereo algorithms and prior learning-based depth estimation techniques underperform when applied on this dual-pixel data, the former due to too-strong assumptions about RGB image matching, and the latter due to not leveraging the understanding of optics of dual-pixel image formation. To allow learning based methods to work well on dual-pixel imagery, we identify an inherent ambiguity in the depth estimated from dual-pixel cues, and develop an approach to estimate depth up to this ambiguity. Using our approach, existing monocular depth estimation techniques can be effectively applied to dual-pixel data, and

much smaller models can be constructed that still infer high quality depth. To d
emonstrate this, we capture a large dataset of in-the-wild 5-viewpoint RGB image
s paired with corresponding dual-pixel data, and show how view supervision with
this data can be used to learn depth up to the unknown ambiguities. On our new t
ask, our model is 30% more accurate than any prior work on learning-based monocu
lar or stereoscopic depth estimation.
****

Domain-Adaptive Single-View 3D Reconstruction
Pedro O. Pinheiro, Negar Rostamzadeh, Sungjin Ahn; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2019, pp. 7638-7647
Single-view 3D shape reconstruction is an important but challenging problem, mai
nly for two reasons. First, as shape annotation is very expensive to acquire, cu
rrent methods rely on synthetic data, in which ground-truth 3D annotation is eas
y to obtain. However, this results in domain adaptation problem when applied to
natural images. The second challenge is that there are multiple shapes that can
explain a given 2D image. In this paper, we propose a framework to improve over
these challenges using adversarial training. On one hand, we impose domain confu
sion between natural and synthetic image representations to reduce the distribut
ion gap. On the other hand, we impose the reconstruction to be `realistic' by fo
rcing it to lie on a (learned) manifold of realistic object shapes. Our experime
nts show that these constraints improve performance by a large margin over basel
ine reconstruction models. We achieve results competitive with the state of the
art with a much simpler architecture.
****

Transformable Bottleneck Networks
Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, Linjie Luo; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
pp. 7648-7657
We propose a novel approach to performing fine-grained 3D manipulation of image
content via a convolutional neural network, which we call the Transformable Bott
leneck Network (TBN). It applies given spatial transformations directly to a vol
umetric bottleneck within our encoder-bottleneck-decoder architecture. Multi-vie
w supervision encourages the network to learn to spatially disentangle the featu
re space within the bottleneck. The resulting spatial structure can be manipulat
ed with arbitrary spatial transformations. We demonstrate the efficacy of TBNs f
or novel view synthesis, achieving state-of-the-art results on a challenging ben
chmark. We demonstrate that the bottlenecks produced by networks trained for thi
s task contain meaningful spatial structure that allows us to intuitively perfor
m a variety of image manipulations in 3D, well beyond the rigid transformations
seen during training. These manipulations include non-uniform scaling, non-rigid
warping, and combining content from different images. Finally, we extract expli
cit 3D structure from the bottleneck, performing impressive 3D reconstruction fr
om a single input image.
****

RIO: 3D Object Instance Re-Localization in Changing Indoor Environments
Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, Matthias Nies
sner; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2019, pp. 7658-7667
In this work, we introduce the task of 3D object instance re-localization (RIO):
given one or multiple objects in an RGB-D scan, we want to estimate their corre
sponding 6DoF poses in another 3D scan of the same environment taken at a later
point in time. We consider RIO a particularly important task in 3D vision since
it enables a wide range of practical applications, including AI-assistants or ro
bots that are asked to find a specific object in a 3D scene. To address this pro
blem, we first introduce 3RScan, a novel dataset and benchmark, which features 1
482 RGB-D scans of 478 environments across multiple time steps. Each scene inclu
des several objects whose positions change over time, together with ground truth
annotations of object instances and their respective 6DoF mappings among re-sca
ns. Automatically finding 6DoF object poses leads to a particular challenging fe
ature matching task due to varying partial observations and changes in the surro

unding context. To this end, we introduce a new data-driven approach that effici ently finds matching features using a fully-convolutional 3D correspondence netw ork operating on multiple spatial scales. Combined with a 6DoF pose optimization , our method outperforms state-of-the-art baselines on our newly-established ben chmark, achieving an accuracy of 30.58%.
****

Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation
Kiru Park, Timothy Patten, Markus Vincze; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2019, pp. 7668-7677
Estimating the 6D pose of objects using only RGB images remains challenging beca use of problems such as occlusion and symmetries. It is also difficult to constr uct 3D models with precise texture without expert knowledge or specialized scann ing devices. To address these problems, we propose a novel pose estimation metho d, Pix2Pose, that predicts the 3D coordinates of each object pixel without textu red models. An auto-encoder architecture is designed to estimate the 3D coordina tes and expected errors per pixel. These pixel-wise predictions are then used in multiple stages to form 2D-3D correspondences to directly compute poses with th e PnP algorithm with RANSAC iterations. Our method is robust to occlusion by lev eraging recent achievements in generative adversarial training to precisely reco ver occluded parts. Furthermore, a novel loss function, the transformer loss, is proposed to handle symmetric objects by guiding predictions to the closest symm etric pose. Evaluations on three different benchmark datasets containing symmetr ic and occluded objects show our method outperforms the state of the art using o nly RGB images.
****

CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation
Zhigang Li, Gu Wang, Xiangyang Ji; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2019, pp. 7678-7687
6-DoF object pose estimation from a single RGB image is a fundamental and long-s tanding problem in computer vision. Current leading approaches solve it by train ing deep networks to either regress both rotation and translation from image dir ectly or to construct 2D-3D correspondences and further solve them via PnP indir ectly. We argue that rotation and translation should be treated differently for their significant difference. In this work, we propose a novel 6-DoF pose estima tion approach: Coordinates-based Disentangled Pose Network (CDPN), which disenta ngles the pose to predict rotation and translation separately to achieve highly accurate and robust pose estimation. Our method is flexible, efficient, highly a ccurate and can deal with texture-less and occluded objects. Extensive experimen ts on LINEMOD and Occlusion datasets are conducted and demonstrate the superiori ty of our approach. Concretely, our approach significantly exceeds the state-of- the- art RGB-based methods on commonly used metrics.
****

C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion
David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, Andrea Vedal di; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2019, pp. 7688-7697
We propose C3DPO, a method for extracting 3D models of deformable objects from 2 D keypoint annotations in unconstrained images. We do so by learning a deep netw ork that reconstructs a 3D object from a single view at a time, accounting for p artial occlusions, and explicitly factoring the effects of viewpoint changes and object deformations. In order to achieve this factorization, we introduce a nov el regularization technique. We first show that the factorization is successful if, and only if, there exists a certain canonicalization function of the reconst ructed shapes. Then, we learn the canonicalization function together with the re construction one, which constrains the result to be consistent. We demonstrate s tate-of-the-art reconstruction results for methods that do not use ground-truth 3D supervision for a number of benchmarks, including Up3D and PASCAL3D+.
****

Learning to Reconstruct 3D Manhattan Wireframes From a Single Image

Yichao Zhou, Haozhi Qi, Yuexiang Zhai, Qi Sun, Zhili Chen, Li-Yi Wei, Yi Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7698-7707

From a single view of an urban environment, we propose a method to effectively exploit the global structural regularities for obtaining a compact, accurate, and intuitive 3D wireframe representation. Our method trains a single convolutional neural network to simultaneously detect salient junctions and straight lines, as well as predict their 3D depth and vanishing points. Compared with state-of-the-art learning-based wireframe detection methods, our network is much simpler and more unified, leading to better 2D wireframe detection. With a global structural prior (such as Manhattan assumption), our method further reconstructs a full 3D wireframe model, a compact vector representation suitable for a variety of high-level vision tasks such as AR and CAD. We conduct extensive evaluations of our method on a large new synthetic dataset of urban scenes as well as real images. Our code and datasets will be published along with the paper.

****

Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning

Shichen Liu, Tianye Li, Weikai Chen, Hao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7708-7717

Rendering bridges the gap between 2D vision and 3D scenes by simulating the physical process of image formation. By inverting such renderer, one can think of a learning approach to infer 3D information from 2D images. However, standard graphics renderers involve a fundamental discretization step called rasterization, which prevents the rendering process to be differentiable, hence able to be learned. Unlike the state-of-the-art differentiable renderers, which only approximate the rendering gradient in the back propagation, we propose a truly differentiable rendering framework that is able to (1) directly render colorized mesh using differentiable functions and (2) back-propagate efficient supervision signals to mesh vertices and their attributes from various forms of image representations, including silhouette, shading and color images. The key to our framework is a novel formulation that views rendering as an aggregation function that fuses the probabilistic contributions of all mesh triangles with respect to the rendered pixels. Such formulation enables our framework to flow gradients to the occluded and far-range vertices, which cannot be achieved by the previous state-of-the-arts. We show that by using the proposed renderer, one can achieve significant improvement in 3D unsupervised single-view reconstruction both qualitatively and quantitatively. Experiments also demonstrate that our approach is able to handle the challenging tasks in image-based shape fitting, which remain nontrivial to existing differentiable renderers. Code is available at https://github.com/ShichenLiu/SoftRas.

****

Learnable Triangulation of Human Pose

Karim Iskakov, Egor Burkov, Victor Lempitsky, Yury Malkov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7718-7727

We present two novel solutions for multi-view 3D human pose estimation based on new learnable triangulation methods that combine 3D information from multiple 2D views. The first (baseline) solution is a basic differentiable algebraic triangulation with an addition of confidence weights estimated from the input images. The second, more complex, solution is based on volumetric aggregation of 2D feature maps from the 2D backbone followed by refinement via 3D convolutions that produce final 3D joint heatmaps. Crucially, both of the approaches are end-to-end differentiable, which allows us to directly optimize the target metric. We demonstrate transferability of the solutions across datasets and considerably improve the multi-view state of the art on the Human3.6M dataset.

****

xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera

Denis Tome, Patrick Peluse, Lourdes Agapito, Hernan Badino; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7728-7738

We present a new solution to egocentric 3D body pose estimation from monocular i
mages captured from a downward looking fish-eye camera installed on the rim of a
 head mounted virtual reality device. This unusual viewpoint, just 2 cm. away fr
om the user's face, leads to images with unique visual appearance, characterized
 by severe self-occlusions and strong perspective distortions that result in a d
rastic difference in resolution between lower and upper body. Our contribution i
s two-fold. Firstly, we propose a new encoder-decoder architecture with a novel
dual branch decoder designed specifically to account for the varying uncertainty
 in the 2D joint locations. Our quantitative evaluation, both on synthetic and r
eal-world datasets, shows that our strategy leads to substantial improvements in
 accuracy over state of the art egocentric pose estimation approaches. Our secon
d contribution is a new large-scale photorealistic synthetic dataset -- xR-EgoPo
se -- offering 383K frames of high quality renderings of people with a diversity
 of skin tones, body shapes, clothing, in a variety of backgrounds and lighting
conditions, performing a range of actions. Our experiments show that the high va
riability in our new synthetic training corpus leads to good generalization to r
eal world footage and to state of the art results on real world datasets with gr
ound truth. Moreover, an evaluation on the Human3.6M benchmark shows that the pe
rformance of our method is on par with top performing approaches on the more cla
ssic problem of 3D human pose from a third person viewpoint.
****

DeepHuman: 3D Human Reconstruction From a Single Image
Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, Yebin Liu; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7739-77
49
We propose DeepHuman, an image-guided volume-to-volume translation CNN for 3D hu
man reconstruction from a single RGB image. To reduce the ambiguities associated
 with the reconstruction of invisible areas, our method leverages a dense semant
ic representation generated from SMPL model as an additional input. One key feat
ure of our network is that it fuses different scales of image features into the
3D space through volumetric feature transformation, which helps to recover accur
ate surface geometry. The surface details are further refined through a normal r
efinement network, which can be concatenated with the volume generation network
using our proposed volumetric normal projection layer. We also contribute THuman
, a 3D real-world human model dataset containing approximately 7000 models. The
network is trained using training data generated from the dataset. Overall, due
to the specific design of our network and the diversity in our dataset, our meth
od enables 3D human model estimation given only a single image and outperforms s
tate-of-the-art approaches.
****

A Neural Network for Detailed Human Depth Estimation From a Single Image
Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, Ping Tan; Pr
oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20
19, pp. 7750-7759
This paper presents a neural network to estimate a detailed depth map of the for
eground human in a single RGB image. The result captures geometry details such a
s cloth wrinkles, which are important in visualization applications. To achieve
this goal, we separate the depth map into a smooth base shape and a residual det
ail shape and design a network with two branches to regress them respectively. W
e design a training strategy to ensure both base and detail shapes can be faithf
ully learned by the corresponding network branches. Furthermore, we introduce a
novel network layer to fuse a rough depth map and surface normals to further imp
rove the final result. Quantitative comparison with fused `ground truth' capture
d by real depth cameras and qualitative examples on unconstrained Internet image
s demonstrate the strength of the proposed method.
****

DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare
Yuanlu Xu, Song-Chun Zhu, Tony Tung; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2019, pp. 7760-7770
We present DenseRaC, a novel end-to-end framework for jointly estimating 3D huma

n pose and body shape from a monocular RGB image. Our two-step framework takes t
he body pixel-to-surface correspondence map (i.e., IUV map) as proxy representat
ion and then performs estimation of parameterized human pose and shape. Specific
ally, given an estimated IUV map, we develop a deep neural network optimizing 3D
 body reconstruction losses and further integrating a render-and-compare scheme
to minimize differences between the input and the rendered output, i.e., dense b
ody landmarks, body part masks, and adversarial priors. To boost learning, we fu
rther construct a large-scale synthetic dataset (MOCA) utilizing web-crawled Moc
ap sequences, 3D scans and animations. The generated data covers diversified cam
era views, human actions and body shapes, and is paired with full ground truth.
Our model jointly learns to represent the 3D human body from hybrid datasets, mi
tigating the problem of unpaired training data. Our experiments show that DenseR
aC obtains superior performance against state of the art on public benchmarks of
 various human-related tasks.
****

Not All Parts Are Created Equal: 3D Pose Estimation by Modeling Bi-Directional D
ependencies of Body Parts
Jue Wang, Shaoli Huang, Xinchao Wang, Dacheng Tao; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2019, pp. 7771-7780
Not all the human body parts have the same degree of freedom (DOF) due to the ph
ysiological structure. For example, the limbs may move more flexibly and freely
than the torso does. Most of the existing 3D pose estimation methods, despite th
e very promising results achieved, treat the body joints equally and consequentl
y often lead to larger reconstruction errors on the limbs. In this paper, we pro
pose a progressive approach that explicitly accounts for the distinct DOFs among
 the body parts. We model parts with higher DOFs like the elbows, as dependent c
omponents of the corresponding parts with lower DOFs like the torso, of which th
e 3D locations can be more reliably estimated. Meanwhile, the high-DOF parts may
, in turn, impose a constraint on where the low-DOF ones lie. As a result, parts
 with different DOFs supervise one another, yielding physically constrained and
plausible pose-estimation results. To further facilitate the prediction of the h
igh-DOF parts, we introduce a pose-attribution estimation, where the relative lo
cation of a limb joint with respect to the torso, which has the least DOF of a h
uman body, is explicitly estimated and further fed to the joint-estimation modul
e. The proposed approach achieves very promising results, outperforming the stat
e of the art on several benchmarks.
****

Extreme View Synthesis
Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H. Kim, Jan Kautz; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
 pp. 7781-7790
We present Extreme View Synthesis, a solution for novel view extrapolation that
works even when the number of input images is small---as few as two. In this con
text, occlusions and depth uncertainty are two of the most pressing issues, and
worsen as the degree of extrapolation increases. We follow the traditional parad
igm of performing depth-based warping and refinement, with a few key improvement
s. First, we estimate a depth probability volume, rather than just a single dept
h value for each pixel of the novel view. This allows us to leverage depth uncer
tainty in challenging regions, such as depth discontinuities. After using it to
get an initial estimate of the novel view, we explicitly combine learned image p
riors and the depth uncertainty to synthesize a refined image with less artifact
s. Our method is the first to show visually pleasing results for baseline magnif
ications of up to 30x.
****

View Independent Generative Adversarial Network for Novel View Synthesis
Xiaogang Xu, Ying-Cong Chen, Jiaya Jia; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2019, pp. 7791-7800
Synthesizing novel views from a 2D image requires to infer 3D structure and proj
ect it back to 2D from a new viewpoint. In this paper, we propose an encoder-dec
oder based generative adversarial network VI-GAN to tackle this problem. Our met

hod is to let the network, after seeing many images of objects belonging to the same category in different views, obtain essential knowledge of intrinsic properties of the objects. To this end, an encoder is designed to extract view-independent feature that characterizes intrinsic properties of the input image, which includes 3D structure, color, texture etc. We also make the decoder hallucinate the image of a novel view based on the extracted feature and an arbitrary user-specific camera pose. Extensive experiments demonstrate that our model can synthesize high-quality images in different views with continuous camera poses, and is general for various applications.
****

Cascaded Context Pyramid for Full-Resolution 3D Semantic Scene Completion
Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, Xiaoyun Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7801-7810
Semantic Scene Completion (SSC) aims to simultaneously predict the volumetric occupancy and semantic category of a 3D scene. It helps intelligent devices to understand and interact with the surrounding scenes. Due to the high-memory requirement, current methods only produce low-resolution completion predictions, and generally lose the object details. Furthermore, they also ignore the multi-scale spatial contexts, which play a vital role for the 3D inference. To address these issues, in this work we propose a novel deep learning framework, named Cascaded Context Pyramid Network (CCPNet), to jointly infer the occupancy and semantic labels of a volumetric 3D scene from a single depth image. The proposed CCPNet improves the labeling coherence with a cascaded context pyramid. Meanwhile, based on the low-level features, it progressively restores the fine-structures of objects with Guided Residual Refinement (GRR) modules. Our proposed framework has three outstanding advantages: (1) it explicitly models the 3D spatial context for performance improvement; (2) full-resolution 3D volumes are produced with structure-preserving details; (3) light-weight models with low-memory requirements are captured with a good extensibility. Extensive experiments demonstrate that in spite of taking a single-view depth map, our proposed framework can generate high-quality SSC results, and outperforms state-of-the-art approaches on both the synthetic SUNCG and real NYU datasets.
****

View-Consistent 4D Light Field Superpixel Segmentation
Numair Khan, Qian Zhang, Lucas Kasser, Henry Stone, Min H. Kim, James Tompkin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7811-7819
Many 4D light field processing applications rely on superpixel segmentations, for which occlusion-aware view consistency is important. Yet, existing methods often enforce consistency by propagating clusters from a central view only, which can lead to inconsistent superpixels for non-central views. Our proposed approach combines an occlusion-aware angular segmentation in horizontal and vertical EPI spaces with an occlusion-aware clustering and propagation step across all views. Qualitative video demonstrations show that this helps to remove flickering and inconsistent boundary shapes versus the state-of-the-art approach, and quantitative metrics reflect these findings with improved boundary accuracy and view consistency scores.
****

GLoSH: Global-Local Spherical Harmonics for Intrinsic Image Decomposition
Hao Zhou, Xiang Yu, David W. Jacobs; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7820-7829
Traditional intrinsic image decomposition focuses on decomposing images into reflectance and shading, leaving surfaces normals and lighting entangled in shading. In this work, we propose a Global-Local Spherical Harmonics (GLoSH) lighting model to improve the lighting component, and jointly predict reflectance and surface normals. The global SH models the holistic lighting while local SH account for the spatial variation of lighting. Also, a novel non-negative lighting constraint is proposed to encourage the estimated SH to be physically meaningful. To seamlessly reflect the GLoSH model, we design a coarse-to-fine network structure.

The coarse network predicts global SH, reflectance and normals, and the fine network predicts their local residuals. Lacking labels for reflectance and lighting, we apply synthetic data for model pre-training and fine-tune the model with real data in a self-supervised way. Compared to the state-of-the-art methods only targeting normals or reflectance and shading, our method recovers all components and achieves consistently better results on three real datasets, IIW, SAW and NYUv2.

****

Surface Normals and Shape From Water

Satoshi Murai, Meng-Yu Jennifer Kuo, Ryo Kawahara, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7830-7838

In this paper, we introduce a novel method for reconstructing surface normals and depth of dynamic objects in water. Past shape recovery methods have leveraged various visual cues for estimating shape (e.g., depth) or surface normals. Methods that estimate both compute one from the other. We show that these two geometric surface properties can be simultaneously recovered for each pixel when the object is observed underwater. Our key idea is to leverage multi-wavelength near-infrared light absorption along different underwater light paths in conjunction with surface shading. We derive a principled theory for this surface normals and shape from water method and a practical calibration method for determining its imaging parameters values. By construction, the method can be implemented as a one-shot imaging system. We prototype both an off-line and a video-rate imaging system and demonstrate the effectiveness of the method on a number of real-world static and dynamic objects. The results show that the method can recover intricate surface features that are otherwise inaccessible.

****

Restoration of Non-Rigidly Distorted Underwater Images Using a Combination of Compressive Sensing and Local Polynomial Image Representations

Jerin Geo James, Pranay Agrawal, Ajit Rajwade; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7839-7848

Images of static scenes submerged beneath a wavy water surface exhibit severe non-rigid distortions. The physics of water flow suggests that water surfaces possess spatio-temporal smoothness and temporal periodicity. Hence they possess a sparse representation in the 3D discrete Fourier (DFT) basis. Motivated by this, we pose the task of restoration of such video sequences as a compressed sensing (CS) problem. We begin by tracking a few salient feature points across the frames of a video sequence of the submerged scene. Using these point trajectories, we show that the motion fields at all other (non-tracked) points can be effectively estimated using a typical CS solver. This by itself is a novel contribution in the field of non-rigid motion estimation. We show that this method outperforms state of the art algorithms for underwater image restoration. We further consider a simple optical flow algorithm based on local polynomial expansion of the image frames (PEOF). Surprisingly, we demonstrate that PEOF is more efficient and often outperforms all the state of the art methods in terms of numerical measures. Finally, we demonstrate that a two-stage approach consisting of the CS step followed by PEOF much more accurately preserves the image structure and improves the (visual as well as numerical) video quality as compared to just the PEOF stage. The source code, datasets and supplemental material can be accessed at [??], [??].

****

Learning Perspective Undistortion of Portraits

Yajie Zhao, Zeng Huang, Tianye Li, Weikai Chen, Chloe LeGendre, Xinglei Ren, Ari Shapiro, Hao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7849-7859

Near-range portrait photographs often contain perspective distortion artifacts that bias human perception and challenge both facial recognition and reconstruction techniques. We present the first deep learning based approach to remove such artifacts from unconstrained portraits. In contrast to the previous state-of-the-art approach [??], our method handles even portraits with extreme perspective d

istortion, as we avoid the inaccurate and error-prone step of first fitting a 3D face model. Instead, we predict a distortion correction flow map that encodes a per-pixel displacement that removes distortion artifacts when applied to the input image. Our method also automatically infers missing facial features, i.e. occluded ears caused by strong perspective distortion, with coherent details. We demonstrate that our approach significantly outperforms the previous state-of-the-art both qualitatively and quantitatively, particularly for portraits with extreme perspective distortion or facial expressions. We further show that our technique benefits a number of fundamental tasks, significantly improving the accuracy of both face recognition and 3D reconstruction and enables a novel camera calibration technique from a single portrait. Moreover, we also build the first perspective portrait database with a large diversity in identities, expression and poses.

****

Towards Photorealistic Reconstruction of Highly Multiplexed Lensless Images
Salman S. Khan, Adarsh V. R., Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, Kaushik Mitra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7860-7869

Recent advancements in fields like Internet of Things (IoT), augmented reality, etc. have led to an unprecedented demand for miniature cameras with low cost that can be integrated anywhere and can be used for distributed monitoring. Mask-based lensless imaging systems make such inexpensive and compact models realizable. However, reduction in the size and cost of these imagers comes at the expense of their image quality due to the high degree of multiplexing inherent in their design. In this paper, we present a method to obtain image reconstructions from mask-based lensless measurements that are more photorealistic than those currently available in the literature. We particularly focus on FlatCam, a lensless imager consisting of a coded mask placed over a bare CMOS sensor. Existing techniques for reconstructing FlatCam measurements suffer from several drawbacks including lower resolution and dynamic range than lens-based cameras. Our approach overcomes these drawbacks using a fully trainable non-iterative deep learning based model. Our approach is based on two stages: an inversion stage that maps the measurement into the space of intermediate reconstruction and a perceptual enhancement stage that improves this intermediate reconstruction based on perceptual and signal distortion metrics. Our proposed method is fast and produces photo-realistic reconstruction as demonstrated on many real and challenging scenes.

****

Unconstrained Motion Deblurring for Dual-Lens Cameras
M. R. Mahesh Mohan, Sharath Girish, A. N. Rajagopalan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7870-7879

Recently, there has been a renewed interest in leveraging multiple cameras, but under unconstrained settings. They have been quite successfully deployed in smartphones, which have become de facto choice for many photographic applications. However, akin to normal cameras, the functionality of multi-camera systems can be marred by motion blur which is a ubiquitous phenomenon in hand-held cameras. Despite the far-reaching potential of unconstrained camera arrays, there is not a single deblurring method for such systems. In this paper, we propose a generalized blur model that elegantly explains the intrinsically coupled image formation model for dual-lens set-up, which are by far most predominant in smartphones. While image aesthetics is the main objective in normal camera deblurring, any method conceived for our problem is additionally tasked with ascertaining consistent scene-depth in the deblurred images. We reveal an intriguing challenge that stems from an inherent ambiguity unique to this problem which naturally disrupts this coherence. We address this issue by devising a judicious prior, and based on our model and prior propose a practical blind deblurring method for dual-lens cameras, that achieves state-of-the-art performance.

****

Stochastic Exposure Coding for Handling Multi-ToF-Camera Interference
Jongho Lee, Mohit Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7880-7888

As continuous-wave time-of-flight (C-ToF) cameras become popular in 3D imaging a
pplications, they need to contend with the problem of multi-camera interference
(MCI). In a multi-camera environment, a ToF camera may receive light from the so
urces of other cameras, resulting in large depth errors. In this paper, we propo
se stochastic exposure coding (SEC), a novel approach for mitigating. SEC involv
es dividing a camera's integration time into multiple slots, and switching the c
amera off and on stochastically during each slot. This approach has two benefits
. First, by appropriately choosing the on probability for each slot, the camera
can effectively filter out both the AC and DC components of interfering signals,
 thereby mitigating depth errors while also maintaining high signal-to-noise rat
io. This enables high accuracy depth recovery with low power consumption. Second
, this approach can be implemented without modifying the C-ToF camera's coding f
unctions, and thus, can be used with a wide range of cameras with minimal change
s. We demonstrate the performance benefits of SEC with theoretical analysis, sim
ulations and real experiments, across a wide range of imaging scenarios.
****

Convolutional Approximations to the General Non-Line-of-Sight Imaging Operator
Byeongjoo Ahn, Akshat Dave, Ashok Veeraraghavan, Ioannis Gkioulekas, Aswin C
. Sankaranarayanan; Proceedings of the IEEE/CVF International Conference on Comp
uter Vision (ICCV), 2019, pp. 7889-7899
Non-line-of-sight (NLOS) imaging aims to reconstruct scenes outside the field of
 view of an imaging system. A common approach is to measure the so-called light
transients, which facilitates reconstructions through ellipsoidal tomography tha
t involves solving a linear least-squares. Unfortunately, the corresponding line
ar operator is very high-dimensional and lacks structures that facilitate fast s
olvers, and so, the ensuing optimization is a computationally daunting task. We
introduce a computationally tractable framework for solving the ellipsoidal tomo
graphy problem. Our main observation is that the Gram of the ellipsoidal tomogra
phy operator is convolutional, either exactly under certain idealized imaging co
nditions, or approximately in practice. This, in turn, allows us to obtain the e
llipsoidal tomography solution by using efficient deconvolution procedures to so
lve a linear least-squares problem involving the Gram operator. The computationa
l tractability of our approach also facilitates the use of various regularizers
during the deconvolution procedure. We demonstrate the advantages of our framewo
rk in a variety of simulated and real experiments.
****

Agile Depth Sensing Using Triangulation Light Curtains
Joseph R. Bartels, Jian Wang, William "Red" Whittaker, Srinivasa G. Narasimha
n; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2019, pp. 7900-7908
Depth sensors like LIDARs and Kinect use a fixed depth acquisition strategy that
 is independent of the scene of interest. Due to the low spatial and temporal re
solution of these sensors, this strategy can undersample parts of the scene that
 are important (small or fast moving objects), or oversample areas that are not
informative for the task at hand (a fixed planar wall). In this paper, we presen
t an approach and system to dynamically and adaptively sample the depths of a sc
ene using the principle of triangulation light curtains. The approach directly d
etects the presence or absence of objects at specified 3D lines. These 3D lines
can be sampled sparsely, non-uniformly, or densely only at specified regions. Th
e depth sampling can be varied in real-time, enabling quick object discovery or
detailed exploration of areas of interest. These results are achieved using a no
vel prototype light curtain system that is based on a 2D rolling shutter camera
with higher light efficiency, working range, and faster adaptation than previous
 work, making it useful broadly for autonomous navigation and exploration.
****

Asynchronous Single-Photon 3D Imaging
Anant Gupta, Atul Ingle, Mohit Gupta; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2019, pp. 7909-7918
Single-photon avalanche diodes (SPADs) are becoming popular in time-of-flight de
pth-ranging due to their unique ability to capture individual photons with picos

econd timing resolution. However, ambient light (e.g., sunlight) incident on a SPAD-based 3D camera leads to severe non-linear distortions (pileup) in the measured waveform, resulting in large depth errors. We propose asynchronous single-photon 3D imaging, a family of acquisition schemes to mitigate pileup during data acquisition itself. Asynchronous acquisition temporally misaligns SPAD measurement windows and the laser cycles through deterministically predefined or randomized offsets. Our key insight is that pileup distortions can be "averaged out" by choosing a sequence of offsets that span the entire depth range. We develop a generalized image formation model and perform theoretical analysis to explore the space of asynchronous acquisition schemes and design high-performance schemes. Our simulations and experiments demonstrate an improvement in depth accuracy of up to an order of magnitude as compared to the state-of-the-art, across a wide range of imaging scenarios, including those with high ambient flux.

****

Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation

Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7919-7929

Person re-identification (re-ID) aims at recognizing the same person from images taken across different cameras. To address this challenging task, existing re-ID models typically rely on a large amount of labeled training data, which is not practical for real-world applications. To alleviate this limitation, researchers now targets at cross-dataset re-ID which focuses on generalizing the discriminative ability to the unlabeled target domain when given a labeled source domain dataset. To achieve this goal, our proposed Pose Disentanglement and Adaptation Network (PDA-Net) aims at learning deep image representation with pose and domain information properly disentangled. With the learned cross-domain pose invariant feature space, our proposed PDA-Net is able to perform pose disentanglement across domains without supervision in identities, and the resulting features can be applied to cross-dataset re-ID. Both of our qualitative and quantitative results on two benchmark datasets confirm the effectiveness of our approach and its superiority over the state-of-the-art cross-dataset Re-ID approaches.

****

A Learned Representation for Scalable Vector Graphics

Raphael Gontijo Lopes, David Ha, Douglas Eck, Jonathon Shlens; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7930-7939

Dramatic advances in generative models have resulted in near photographic quality for artificially rendered faces, animals and other objects in the natural world. In spite of such advances, a higher level understanding of vision and imagery does not arise from exhaustively modeling an object, but instead identifying higher-level attributes that best summarize the aspects of an object. In this work we attempt to model the drawing process of fonts by building sequential generative models of vector graphics. This model has the benefit of providing a scale-invariant representation for imagery whose latent representation may be systematically manipulated and exploited to perform style propagation. We demonstrate the se results on a large dataset of fonts crawled from the web and highlight how such a model captures the statistical dependencies and richness of this dataset. We envision that our model can find use as a tool for graphic designers to facilitate font design.

****

ELF: Embedded Localisation of Features in Pre-Trained CNN

Assia Benbihi, Matthieu Geist, Cedric Pradalier; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7940-7949

This paper introduces a novel feature detector based only on information embedded inside a CNN trained on standard tasks (e.g. classification). While previous works already show that the features of a trained CNN are suitable descriptors, we show here how to extract the feature locations from the network to build a detector. This information is computed from the gradient of the feature map with re

spect to the input image. This provides a saliency map with local maxima on rele
vant keypoint locations. Contrary to recent CNN-based detectors, this method req
uires neither supervised training nor finetuning. We evaluate how repeatable and
 how 'matchable' the detected keypoints are with the repeatability and matching
scores. Matchability is measured with a simple descriptor introduced for the sak
e of the evaluation. This novel detector reaches similar performances on the sta
ndard evaluation HPatches dataset, as well as comparable robustness against illu
mination and viewpoint changes on Webcam and photo-tourism images. These results
 show that a CNN trained on a standard task embeds feature location information
that is as relevant as when the CNN is specifically trained for feature detectio
n.
****

Joint Group Feature Selection and Discriminative Filter Learning for Robust Visu
al Object Tracking

Tianyang Xu, Zhen-Hua Feng, Xiao-Jun Wu, Josef Kittler; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7950-7960

We propose a new Group Feature Selection method for Discriminative Correlation F
ilters (GFS-DCF) based visual object tracking. The key innovation of the propose
d method is to perform group feature selection across both channel and spatial d
imensions, thus to pinpoint the structural relevance of multi-channel features t
o the filtering system. In contrast to the widely used spatial regularisation or
 feature selection methods, to the best of our knowledge, this is the first time
 that channel selection has been advocated for DCF-based tracking. We demonstrat
e that our GFS-DCF method is able to significantly improve the performance of a
DCF tracker equipped with deep neural network features. In addition, our GFS-DCF
 enables joint feature selection and filter learning, achieving enhanced discrim
ination and interpretability of the learned filters. To further improve the perf
ormance, we adaptively integrate historical information by constraining filters
to be smooth across temporal frames, using an efficient low-rank approximation.
By design, specific temporal-spatial-channel configurations are dynamically lear
ned in the tracking process, highlighting the relevant features, and alleviating
 the performance degrading impact of less discriminative representations and red
ucing information redundancy. The experimental results obtained on OTB2013, OTB2
015, VOT2017, VOT2018 and TrackingNet demonstrate the merits of our GFS-DCF and
its superiority over the state-of-the-art trackers. The code is publicly availab
le at https://github.com/XU-TIANYANG/GFS-DCF.
****

Sampling Wisely: Deep Image Embedding by Top-K Precision Optimization

Jing Lu, Chaofan Xu, Wei Zhang, Ling-Yu Duan, Tao Mei; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7961-7970

Deep image embedding aims at learning a convolutional neural network (CNN) based
 mapping function that maps an image to a feature vector. The embedding quality
is usually evaluated by the performance in image search tasks. Since very few us
ers bother to open the second page search results, top-k precision mostly domina
tes the user experience and thus is one of the crucial evaluation metrics for th
e embedding quality. Despite being extensively studied, existing algorithms are
usually based on heuristic observation without theoretical guarantee. Consequent
ly, gradient descent direction on the training loss is mostly inconsistent with
the direction of optimizing the concerned evaluation metric. This inconsistency
certainly misleads the training direction and degrades the performance. In contr
ast to existing works, in this paper, we propose a novel deep image embedding al
gorithm with end-to-end optimization to top-k precision, the evaluation metric t
hat is closely related to user experience. Specially, our loss function is const
ructed with wisely selected "misplaced" images along the top k nearest neighbor
decision boundary, so that the gradient descent update directly promotes the con
cerned metric, top-k precision. Further more, our theoretical analysis on the up
per bounding and consistency properties of the proposed loss supports that minim
izing our proposed loss is equivalent to maximizing top-k precision. Experiments
 show that our proposed algorithm outperforms all compared state-of-the-art deep
 image embedding algorithms on three benchmark datasets.

****
On the Global Optima of Kernelized Adversarial Representation Learning
Bashir Sadeghi, Runyi Yu, Vishnu Boddeti; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7971-7979
Adversarial representation learning is a promising paradigm for obtaining data representations that are invariant to certain sensitive attributes while retaining the information necessary for predicting target attributes. Existing approaches solve this problem through iterative adversarial minimax optimization and lack theoretical guarantees. In this paper, we first study the "linear" form of this problem i.e., the setting where all the players are linear functions. We show that the resulting optimization problem is both non-convex and non-differentiable. We obtain an exact closed-form expression for its global optima through spectral learning and provide performance guarantees in terms of analytical bounds on the achievable utility and invariance. We then extend this solution and analysis to non-linear functions through kernel representation. Numerical experiments on UCI, Extended Yale B and CIFAR-100 datasets indicate that, (a) practically, our solution is ideal for "imparting" provable invariance to any biased pre-trained data representation, and (b) the global optima of the "kernel" form can provide a comparable trade-off between utility and invariance in comparison to iterative minimax optimization of existing deep neural network based approaches, but with provable guarantees.
****
Addressing Model Vulnerability to Distributional Shifts Over Image Transformation Sets
Riccardo Volpi, Vittorio Murino; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7980-7989
We are concerned with the vulnerability of computer vision models to distributional shifts. We formulate a combinatorial optimization problem that allows evaluating the regions in the image space where a given model is more vulnerable, in terms of image transformations applied to the input, and face it with standard search algorithms. We further embed this idea in a training procedure, where we define new data augmentation rules according to the image transformations that the current model is most vulnerable to, over iterations. An empirical evaluation on classification and semantic segmentation problems suggests that the devised algorithm allows to train models that are more robust against content-preserving image manipulations and, in general, against distributional shifts.
****
Attract or Distract: Exploit the Margin of Open Set
Qianyu Feng, Guoliang Kang, Hehe Fan, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7990-7999
Open set domain adaptation aims to diminish the domain shift across domains, with partially shared classes. There exist unknown target samples out of the knowledge of source domain. Compared to the close set setting, how to separate the unknown (unshared) class from the known (shared) ones plays the key role. Whereas, previous methods did not emphasize the semantic structure of the open set data, which may introduce bias into the domain alignment and confuse the classifier around the decision boundary. In this paper, we exploit the semantic structure of open set data from two aspects: 1) Semantic Categorical Alignment, which aims to achieve good separability of target known classes by categorically aligning the centroid of target with the source. 2) Semantic Contrastive Mapping, which aims to push the unknown class away from the decision boundary. Empirically, we demonstrate that our method performs favourably against the state-of-the-art methods on representative benchmarks, e.g. Digits and Office-31 datasets.
****
MIC: Mining Interclass Characteristics for Improved Metric Learning
Karsten Roth, Biagio Brattoli, Bjorn Ommer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8000-8009
Metric learning seeks to embed images of objects such that class-defined relations are captured by the embedding space. However, variability in images is not just due to different depicted object classes, but also depends on other latent ch

aracteristics such as viewpoint or illumination. In addition to these structured properties, random noise further obstructs the visual relations of interest. The common approach to metric learning is to enforce a representation that is invariant under all factors but the ones of interest. In contrast, we propose to explicitly learn the latent characteristics that are shared by and go across object classes. We can then directly explain away structured visual variability, rather than assuming it to be unknown random noise. We propose a novel surrogate task to learn visual characteristics shared across classes with a separate encoder. This encoder is trained jointly with the encoder for class information by reducing their mutual information. On five standard image retrieval benchmarks the approach significantly improves upon the state-of-the-art. Code is available at https://github.com/Confusezius/metric-learning-mining-interclass-characteristics.
****

Self-Supervised Representation Learning via Neighborhood-Relational Encoding
Mohammad Sabokrou, Mohammad Khalooei, Ehsan Adeli; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8010-8019
In this paper, we propose a novel self-supervised representation learning by taking advantage of a neighborhood-relational encoding (NRE) among the training data. Conventional unsupervised learning methods only focused on training deep networks to understand the primitive characteristics of the visual data, mainly to be able to reconstruct the data from a latent space. They often neglected the relation among the samples, which can serve as an important metric for self-supervision. Different from the previous work, NRE aims at preserving the local neighborhood structure on the data manifold. Therefore, it is less sensitive to outliers. We integrate our NRE component with an encoder-decoder structure for learning to represent samples considering their local neighborhood information. Such discriminative and unsupervised representation learning scheme is adaptable to different computer vision tasks due to its independence from intense annotation requirements. We evaluate our proposed method for different tasks, including classification, detection, and segmentation based on the learned latent representations. In addition, we adopt the auto-encoding capability of our proposed method for applications like defense against adversarial example attacks and video anomaly detection. Results confirm the performance of our method is better or at least comparable with the state-of-the-art for each specific application, but with a generic and self-supervised approach.
****

AWSD: Adaptive Weighted Spatiotemporal Distillation for Video Representation
Mohammad Tavakolian, Hamed R. Tavakoli, Abdenour Hadid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8020-8029
We propose an Adaptive Weighted Spatiotemporal Distillation (AWSD) technique for video representation by encoding the appearance and dynamics of the videos into a single RGB image map. This is obtained by adaptively dividing the videos into small segments and comparing two consecutive segments. This allows using pre-trained models on still images for video classification while successfully capturing the spatiotemporal variations in the videos. The adaptive segment selection enables effective encoding of the essential discriminative information of untrimmed videos. Based on Gaussian Scale Mixture, we compute the weights by extracting the mutual information between two consecutive segments. Unlike pooling-based methods, our AWSD gives more importance to the frames that characterize actions or events thanks to its adaptive segment length selection. We conducted extensive experimental analysis to evaluate the effectiveness of our proposed method and compared our results against those of recent state-of-the-art methods on four benchmark datatsets, including UCF101, HMDB51, ActivityNet v1.3, and Maryland. The obtained results on these benchmark datatsets showed that our method significantly outperforms earlier works and sets the new state-of-the-art performance in video classification. Code is available at the project webpage: https://mohammadt68.github.io/AWSD/
****

Bilinear Attention Networks for Person Retrieval
Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, Mehrtash Hara

ndi; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2019, pp. 8030-8039
This paper investigates a novel Bilinear attention (Bi-attention) block, which d
iscovers and uses second order statistical information in an input feature map,
for the purpose of person retrieval. The Bi-attention block uses bilinear poolin
g to model the local pairwise feature interactions along each channel, while pre
serving the spatial structural information. We propose an Attention in Attention
 (AiA) mechanism to build inter-dependency among the second order local and glob
al features with the intent to make better use of, or pay more attention to, suc
h higher order statistical relationships. The proposed network, equipped with th
e proposed Bi-attention is referred to as Bilinear ATtention network (BAT-net).
Our approach outperforms current state-of-the-art by a considerable margin acros
s the standard benchmark datasets (e.g., CUHK03, Market-1501, DukeMTMC-reID and
MSMT17).
****

Discriminative Feature Learning With Consistent Attention Regularization for Per
son Re-Identification
Sanping Zhou, Fei Wang, Zeyi Huang, Jinjun Wang; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2019, pp. 8040-8049
Person re-identification (Re-ID) has undergone a rapid development with the bloo
ming of deep neural network. Most methods are very easily affected by target mis
alignment and background clutter in the training process. In this paper, we prop
ose a simple yet effective feedforward attention network to address the two ment
ioned problems, in which a novel consistent attention regularizer and an improve
d triplet loss are designed to learn foreground attentive features for person Re
-ID. Specifically, the consistent attention regularizer aims to keep the deduced
 foreground masks similar from the low-level, mid-level and high-level feature m
aps. As a result, the network will focus on the foreground regions at the lower
layers, which is benefit to learn discriminative features from the foreground re
gions at the higher layers. Last but not least, the improved triplet loss is int
roduced to enhance the feature learning capability, which can jointly minimize t
he intra-class distance and maximize the inter-class distance in each triplet un
it. Experimental results on the Market1501, DukeMTMC-reID and CUHK03 datasets ha
ve shown that our method outperforms most of the state-of-the-art approaches.
****

Semi-Supervised Domain Adaptation via Minimax Entropy
Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, Kate Saenko; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 8050-8058
Contemporary domain adaptation methods are very effective at aligning feature di
stributions of source and target domains without any target supervision. However
, we show that these techniques perform poorly when even a few labeled examples
are available in the target domain. To address this semi-supervised domain adapt
ation (SSDA) setting, we propose a novel Minimax Entropy (MME) approach that adv
ersarially optimizes an adaptive few-shot model. Our base model consists of a fe
ature encoding network, followed by a classification layer that computes the fea
tures' similarity to estimated prototypes (representatives of each class). Adapt
ation is achieved by alternately maximizing the conditional entropy of unlabeled
 target data with respect to the classifier and minimizing it with respect to th
e feature encoder. We empirically demonstrate the superiority of our method over
 many baselines, including conventional feature alignment and few-shot methods,
setting a new state of the art for SSDA. Our code is available at http://cs-peop
le.bu.edu/keisaito/research/MME.html.
****

Boosting Few-Shot Visual Learning With Self-Supervision
Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Perez, Matthieu Cord
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2019, pp. 8059-8068
Few-shot learning and self-supervised learning address different facets of the s
ame problem: how to train a model with little or no labeled data. Few-shot learn

ing aims for optimization methods and models that can learn efficiently to recog nize patterns in the low data regime. Self-supervised learning focuses instead o n unlabeled data and looks into it for the supervisory signal to feed high capac ity deep neural networks. In this work we exploit the complementarity of these t wo domains and propose an approach for improving few-shot learning through self- supervision. We use self-supervision as an auxiliary task in a few-shot learning pipeline, enabling feature extractors to learn richer and more transferable vis ual representations while still using few annotated samples. Through self-superv ision, our approach can be naturally extended towards using diverse unlabeled da ta from other datasets in the few-shot setting. We report consistent improvement s across an array of architectures, datasets and self-supervision techniques. We provide the implementation code at: https://github.com/valeoai/BF3S

****

FDA: Feature Disruptive Attack

Aditya Ganeshan, Vivek B.S., R. Venkatesh Babu; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2019, pp. 8069-8079

Though Deep Neural Networks (DNN) show excellent performance across various comp uter vision tasks, several works show their vulnerability to adversarial samples , i.e., image samples with imperceptible noise engineered to manipulate the netw ork's prediction. Adversarial sample generation methods range from simple to com plex optimization techniques. Majority of these methods generate adversaries thr ough optimization objectives that are tied to the pre-softmax or softmax output of the network. In this work we, (i) show the drawbacks of such attacks, (ii) pr opose two new evaluation metrics: Old Label New Rank (OLNR) and New Label Old Ra nk (NLOR) in order to quantify the extent of damage made by an attack, and (iii) propose a new attack FDA: Feature Disruptive attack, to address the drawbacks o f existing attacks. FDA works by generating image perturbation that disrupts fea tures at each layer of the network and causes deep-features to be highly corrupt . This allows FDA adversaries to severely reduce the performance of deep network s. We experimentally validate that FDA generates stronger adversaries than other state-of-the-art methods for Image classification, even in the presence of vari ous defense measures. More importantly, we show that FDA disrupts feature-repres entation based tasks even without access to the task-specific network or methodo logy.

****

A Novel Unsupervised Camera-Aware Domain Adaptation Framework for Person Re-Iden tification

Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, Yang Gao; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8080-8089

Unsupervised cross-domain person re-identification (Re-ID) faces two key issues. One is the data distribution discrepancy between source and target domains, and the other is the lack of discriminative information in target domain. From the perspective of representation learning, this paper proposes a novel end-to-end d eep domain adaptation framework to address them. For the first issue, we highlig ht the presence of camera-level sub-domains as a unique characteristic in person Re-ID, and develop a "camera-aware" domain adaptation method via adversarial le arning. With this method, the learned representation reduces distribution discre pancy not only between source and target domains but also across all cameras. Fo r the second issue, we exploit the temporal continuity in each camera of target domain to create discriminative information. This is implemented by dynamically generating online triplets within each batch, in order to maximally take advanta ge of the steadily improved representation in training process. Together, the ab ove two methods give rise to a new unsupervised domain adaptation framework for person Re-ID. Extensive experiments and ablation studies conducted on benchmark datasets demonstrate its superiority and interesting properties.

****

Cross-View Policy Learning for Street Navigation

Ang Li, Huiyi Hu, Piotr Mirowski, Mehrdad Farajtabar; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2019, pp. 8100-8109

The ability to navigate from visual observations in unfamiliar environments is a core component of intelligent agents and an ongoing challenge for Deep Reinforcement Learning (RL). Street View can be a sensible testbed for such RL agents, because it provides real-world photographic imagery at ground level, with diverse street appearances; it has been made into an interactive environment called StreetLearn and used for research on navigation. However, goal-driven street navigation agents have not so far been able to transfer to unseen areas without extensive retraining, and relying on simulation is not a scalable solution. Since aerial images are easily and globally accessible, we propose instead to transfer a ground view policy, from training areas to unseen (target) parts of the city, by utilizing aerial view observations. Our core idea is to pair the ground view with an aerial view and to learn a joint policy that is transferable across views. We achieve this by learning a similar embedding space for both views, distilling the policy across views and dropping out visual modalities. We further reformulate the transfer learning paradigm into three stages: 1) cross-modal training, when the agent is initially trained on multiple city regions, 2) aerial view-only adaptation to a new area, when the agent is adapted to a held-out region using only the easily obtainable aerial view, and 3) ground view-only transfer, when the agent is tested on navigation tasks on unseen ground views, without aerial imagery. Our experimental results suggest that the proposed cross-view policy learning enables better generalization of the agent and allows for more effective transfer to unseen environments.The ability to navigate from visual observations in unfamiliar environments is a core component of intelligent agents and an ongoing challenge for Deep Reinforcement Learning (RL). Street View can be a sensible testbed for such RL agents, because it provides real-world photographic imagery at ground level, with diverse street appearances; it has been made into an interactive environment called StreetLearn and used for research on navigation. However, goal-driven street navigation agents have not so far been able to transfer to unseen areas without extensive retraining, and relying on simulation is not a scalable solution. Since aerial images are easily and globally accessible, we propose instead to train a multi-modal policy on ground and aerial views, then transfer the ground view policy to unseen (target) parts of the city by utilizing aerial view observations. Our core idea is to pair the ground view with an aerial view and to learn a joint policy that is transferable across views. We achieve this by learning a similar embedding space for both views, distilling the policy across views and dropping out visual modalities. We further reformulate the transfer learning paradigm into three stages: 1) cross-modal training, when the agent is initially trained on multiple city regions, 2) aerial view-only adaptation to a new area, when the agent is adapted to a held-out region using only the easily obtainable aerial view, and 3) ground view-only transfer, when the agent is tested on navigation tasks on unseen ground views, without aerial imagery. Experimental results suggest that the proposed cross-view policy learning enables better generalization of the agent and allows for more effective transfer to unseen environments.

****

Learning Across Tasks and Domains

Pierluigi Zama Ramirez, Alessio Tonioni, Samuele Salti, Luigi Di Stefano; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8110-8119

Recent works have proven that many relevant visual tasks are closely related one to another. Yet, this connection is seldom deployed in practice due to the lack of practical methodologies to transfer learned concepts across different training processes. In this work, we introduce a novel adaptation framework that can operate across both task and domains. Our framework learns to transfer knowledge across tasks in a fully supervised domain (e.g., synthetic data) and use this knowledge on a different domain where we have only partial supervision (e.g., real data). Our proposal is complementary to existing domain adaptation techniques and extends them to cross tasks scenarios providing additional performance gains. We prove the effectiveness of our framework across two challenging tasks (i.e., monocular depth estimation and semantic segmentation) and four different domain

s (Synthia, Carla, Kitti, and Cityscapes).
****

EMPNet: Neural Localisation and Mapping Using Embedded Memory Points
Gil Avraham, Yan Zuo, Thanuja Dharmasiri, Tom Drummond; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8120-8129
Continuously estimating an agent's state space and a representation of its surroundings has proven vital towards full autonomy. A shared common ground among systems which successfully achieve this feat is the integration of previously encountered observations into the current state being estimated. This necessitates the use of a memory module for incorporating previously visited states whilst simultaneously offering an internal representation of the observed environment. In this work we develop a memory module which contains rigidly aligned point-embeddings that represent a coherent scene structure acquired from an RGB-D sequence of observations. The point-embeddings are extracted using modern convolutional neural network architectures, and alignment is performed by computing a dense correspondence matrix between a new observation and the current embeddings residing in the memory module. The whole framework is end-to-end trainable, resulting in a recurrent joint optimisation of the point-embeddings contained in the memory. This process amplifies the shared information across states, providing increased robustness and accuracy. We show significant improvement of our method across a set of experiments performed on the synthetic VIZDoom environment and a real world Active Vision Dataset.
****

AVT: Unsupervised Learning of Transformation Equivariant Representations by Autoencoding Variational Transformations
Guo-Jun Qi, Liheng Zhang, Chang Wen Chen, Qi Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8130-8139
The learning of Transformation-Equivariant Representations (TERs), which is introduced by Hinton et al. [??], has been considered as a principle to reveal visual structures under various transformations. It contains the celebrated Convolutional Neural Networks (CNNs) as a special case that only equivary to the translations. In contrast, we seek to train TERs for a generic class of transformations and train them in an unsupervised fashion. To this end, we present a novel principled method by Autoencoding Variational Transformations (AVT), compared with the conventional approach to autoencoding data. Formally, given transformed images, the AVT seeks to train the networks by maximizing the mutual information between the transformations and representations. This ensures the resultant TERs of individual images contain the intrinsic information about their visual structures that would equivary extricably under various transformations in a generalized nonlinear case. Technically, we show that the resultant optimization problem can be efficiently solved by maximizing a variational lower-bound of the mutual information. This variational approach introduces a transformation decoder to approximate the intractable posterior of transformations, resulting in an autoencoding architecture with a pair of the representation encoder and the transformation decoder. Experiments demonstrate the proposed AVT model sets a new record for the performances on unsupervised tasks, greatly closing the performance gap to the supervised models.
****

Composite Shape Modeling via Latent Space Factorization
Anastasia Dubrovina, Fei Xia, Panos Achlioptas, Mira Shalah, Raphael Groscot, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8140-8149
We present a novel neural network architecture, termed Decomposer-Composer, for semantic structure-aware 3D shape modeling. Our method utilizes an auto-encoder-based pipeline, and produces a novel factorized shape embedding space, where the semantic structure of the shape collection translates into a data-dependent sub-space factorization, and where shape composition and decomposition become simple linear operations on the embedding coordinates. We further propose to model shape assembly using an explicit learned part deformation module, which utilizes a 3D spatial transformer network to perform an in-network volumetric grid deforma

tion, and which allows us to train the whole system end-to-end. The resulting ne
twork allows us to perform part-level shape manipulation, unattainable by existi
ng approaches. Our extensive ablation study, comparison to baseline methods and
qualitative analysis demonstrate the improved performance of the proposed method
.
****

Deep Comprehensive Correlation Mining for Image Clustering
Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, Hongb
in Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2019, pp. 8150-8159
Recent developed deep unsupervised methods allow us to jointly learn representat
ion and cluster unlabelled data. These deep clustering methods %like DAC start w
ith mainly focus on the correlation among samples, e.g., selecting high precisio
n pairs to gradually tune the feature representation, which neglects other usefu
l correlations. In this paper, we propose a novel clustering framework, named de
ep comprehensive correlation mining (DCCM), for exploring and taking full advant
age of various kinds of correlations behind the unlabeled data from three aspect
s: 1) Instead of only using pair-wise information, pseudo-label supervision is p
roposed to investigate category information and learn discriminative features. 2
) The features' robustness to image transformation of input space is fully explo
red, which benefits the network learning and significantly improves the performa
nce. 3) The triplet mutual information among features is presented for clusterin
g problem to lift the recently discovered instance-level deep mutual information
to a triplet-level formation, which further helps to learn more discriminative
features. Extensive experiments on several challenging datasets show that our me
thod achieves good performance, e.g., attaining 62.3% clustering accuracy on CIF
AR-10, which is 10.1% higher than the state-of-the-art results.
****

Unsupervised Multi-Task Feature Learning on Point Clouds
Kaveh Hassani, Mike Haley; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2019, pp. 8160-8171
We introduce an unsupervised multi-task model to jointly learn point and shape f
eatures on point clouds. We define three unsupervised tasks including clustering
, reconstruction, and self-supervised classification to train a multi-scale grap
h-based encoder. We evaluate our model on shape classification and segmentation
benchmarks. The results suggest that it outperforms prior state-of-the-art unsup
ervised models: In the ModelNet40 classification task, it achieves an accuracy o
f 89.1% and in ShapeNet segmentation task, it achieves an mIoU of 68.2 and accur
acy of 88.6%.
****

Reciprocal Multi-Layer Subspace Learning for Multi-View Clustering
Ruihuang Li, Changqing Zhang, Huazhu Fu, Xi Peng, Tianyi Zhou, Qinghua Hu;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2019, pp. 8172-8180
Multi-view clustering is a long-standing important research topic, however, rema
ins challenging when handling high-dimensional data and simultaneously exploring
the consistency and complementarity of different views. In this work, we presen
t a novel Reciprocal Multi-layer Subspace Learning (RMSL) algorithm for multi-vi
ew clustering, which is composed of two main components: Hierarchical Self-Repre
sentative Layers (HSRL), and Backward Encoding Networks (BEN). Specifically, HSR
L constructs reciprocal multi-layer subspace representations linked with a laten
t representation to hierarchically recover the underlying low-dimensional subspa
ces in which the high-dimensional data lie; BEN explores complex relationships a
mong different views and implicitly enforces the subspaces of all views to be co
nsistent with each other and more separable. The latent representation flexibly
encodes complementary information from multiple views and depicts data more comp
rehensively. Our model can be efficiently optimized by an alternating optimizati
on scheme. Extensive experiments on benchmark datasets show the superiority of R
MSL over other state-of-the-art clustering methods.
****

Geometric Disentanglement for Generative Latent Shape Models
Tristan Aumentado-Armstrong, Stavros Tsogkas, Allan Jepson, Sven Dickinson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8181-8190
Representing 3D shapes is a fundamental problem in artificial intelligence, which has numerous applications within computer vision and graphics. One avenue that has recently begun to be explored is the use of latent representations of generative models. However, it remains an open problem to learn a generative model of shapes that is interpretable and easily manipulated, particularly in the absence of supervised labels. In this paper, we propose an unsupervised approach to partitioning the latent space of a variational autoencoder for 3D point clouds in a natural way, using only geometric information, that builds upon prior work utilizing generative adversarial models of point sets. Our method makes use of tools from spectral geometry to separate intrinsic and extrinsic shape information, and then considers several hierarchical disentanglement penalties for dividing the latent space in this manner. We also propose a novel disentanglement penalty that penalizes the predicted change in the latent representation of the output,with respect to the latent variables of the initial shape. We show that the resulting latent representation exhibits intuitive and interpretable behaviour, enabling tasks such as pose transfer that cannot easily be performed by models with an entangled representation.
****
GAN-Tree: An Incrementally Learned Hierarchical Generative Framework for Multi-Modal Data Distributions
Jogendra Nath Kundu, Maharshi Gor, Dakshit Agrawal, R. Venkatesh Babu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8191-8200
Despite the remarkable success of generative adversarial networks, their performance seems less impressive for diverse training sets, requiring learning of discontinuous mapping functions. Though multi-mode prior or multi-generator models have been proposed to alleviate this problem, such approaches may fail depending on the empirically chosen initial mode components. In contrast to such bottom-up approaches, we present GAN-Tree, which follows a hierarchical divisive strategy to address such discontinuous multi-modal data. Devoid of any assumption on the number of modes, GAN-Tree utilizes a novel mode-splitting algorithm to effectively split the parent mode to semantically cohesive children modes, facilitating unsupervised clustering. Further, it also enables incremental addition of new data modes to an already trained GAN-Tree, by updating only a single branch of the tree structure. As compared to prior approaches, the proposed framework offers a higher degree of flexibility in choosing a large variety of mutually exclusive and exhaustive tree nodes called GAN-Set. Extensive experiments on synthetic and natural image datasets including ImageNet demonstrate the superiority of GAN-Tree against the prior state-of-the-art.
****
GODS: Generalized One-Class Discriminative Subspaces for Anomaly Detection
Jue Wang, Anoop Cherian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8201-8211
One-class learning is the classic problem of fitting a model to data for which annotations are available only for a single class. In this paper, we propose a novel objective for one-class learning. Our key idea is to use a pair of orthonormal frames -- as subspaces -- to "sandwich" the labeled data via optimizing for two objectives jointly: i) minimize the distance between the origins of the two subspaces, and ii) to maximize the margin between the hyperplanes and the data, either subspace demanding the data to be in its positive and negative orthant respectively. Our proposed objective however leads to a non-convex optimization problem, to which we resort to Riemannian optimization schemes and derive an efficient conjugate gradient scheme on the Stiefel manifold. To study the effectiveness of our scheme, we propose a new dataset Dash-Cam-Pose, consisting of clips with skeleton poses of humans seated in a car, the task being to classify the clips as normal or abnormal; the latter is when any human pose is out-of-position wit

h regard to say an airbag deployment. Our experiments on the proposed Dash-Cam-P ose dataset, as well as several other standard anomaly/novelty detection benchma rks demonstrate the benefits of our scheme, achieving state-of-the-art one-class accuracy.
****

Neighborhood Preserving Hashing for Scalable Video Retrieval
Shuyan Li, Zhixiang Chen, Jiwen Lu, Xiu Li, Jie Zhou; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8212-8221
In this paper, we propose a Neighborhood Preserving Hashing (NPH) method for sca lable video retrieval in an unsupervised manner. Unlike most existing deep video hashing methods which indiscriminately compress an entire video into a binary c ode, we embed the spatial-temporal neighborhood information into the encoding ne twork such that the neighborhood-relevant visual content of a video can be prefe rentially encoded into a binary code under the guidance of the neighborhood info rmation. Specifically, we propose a neighborhood attention mechanism which focus es on partial useful content of each input frame conditioned on the neighborhood information. We then integrate the neighborhood attention mechanism into an RNN -based reconstruction scheme to encourage the binary codes to capture the spatia l-temporal structure in a video which is consistent with that in the neighborhoo d. As a consequence, the learned hashing functions can map similar videos to sim ilar binary codes. Extensive experiments on three widely-used benchmark datasets validate the effectiveness of our proposed approach.
****

Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification
Xinyu Zhang, Jiewei Cao, Chunhua Shen, Mingyu You; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2019, pp. 8222-8231
Person re-identification (Re-ID) has achieved great improvement with deep learni ng and a large amount of labelled training data. However, it remains a challengi ng task for adapting a model trained in a source domain of labelled data to a ta rget domain of only unlabelled data available. In this work, we develop a self-t raining method with progressive augmentation framework (PAST) to promote the mod el performance progressively on the target dataset. Specially, our PAST framewor k consists of two stages, namely, conservative stage and promoting stage. The co nservative stage captures the local structure of target-domain data points with triplet-based loss functions, leading to improved feature representations. The p romoting stage continuously optimizes the network by appending a changeable clas sification layer to the last layer of the model, enabling the use of global info rmation about the data distribution. Importantly, we propose a new self-training strategy that progressively augments the model capability by adopting conservat ive and promoting stages alternately. Furthermore, to improve the reliability of selected triplet samples, we introduce a ranking-based triplet loss in the cons ervative stage, which is a label-free objective function based on the similariti es between data pairs. Experiments demonstrate that the proposed method achieves state-of-the-art person Re-ID performance under the unsupervised cross-domain s etting. Code is available at: tinyurl.com/PASTReID
****

SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects
Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, Kun Fu; Proceedings of the IEEE/CVF International Conference on Computer V ision (ICCV), 2019, pp. 8232-8241
Object detection has been a building block in computer vision. Though considerab le progress has been made, there still exist challenges for objects with small s ize, arbitrary direction, and dense distribution. Apart from natural images, suc h issues are especially pronounced for aerial images of great importance. This p aper presents a novel multi-category rotation detector for small, cluttered and rotated objects, namely SCRDet. Specifically, a sampling fusion network is devis ed which fuses multi-layer feature with effective anchor sampling, to improve th e sensitivity to small objects. Meanwhile, the supervised pixel attention networ k and the channel attention network are jointly explored for small and cluttered

object detection by suppressing the noise and highlighting the objects feature. For more accurate rotation estimation, the IoU constant factor is added to the smooth L1 loss to address the boundary problem for the rotating bounding box. Extensive experiments on two remote sensing public datasets DOTA, NWPU VHR-10 as well as natural image datasets COCO, VOC2007 and scene text data ICDAR2015 show the state-of-the-art performance of our detector. The code and models will be available at https://github.com/DetectionTeamUCAS.
****

Cross-X Learning for Fine-Grained Visual Categorization
Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S. Davis, Jun Li, Jian Yang, Ser-Nam Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8242-8251
Recognizing objects from subcategories with very subtle differences remains a challenging task due to the large intra-class and small inter-class variation. Recent work tackles this problem in a weakly-supervised manner: object parts are first detected and the corresponding part-specific features are extracted for fine-grained classification. However, these methods typically treat the part-specific features of each image in isolation while neglecting their relationships between different images. In this paper, we propose Cross-X learning, a simple yet effective approach that exploits the relationships between different images and between different network layers for robust multi-scale feature learning. Our approach involves two novel components: (i) a cross-category cross-semantic regularizer that guides the extracted features to represent semantic parts and, (ii) a cross-layer regularizer that improves the robustness of multi-scale features by matching the prediction distribution across multiple layers. Our approach can be easily trained end-to-end and is scalable to large datasets like NABirds. We empirically analyze the contributions of different components of our approach and demonstrate its robustness, effectiveness and state-of-the-art performance on five benchmark datasets. Code is available at https://github.com/cswluo/CrossX.
****

Maximum-Margin Hamming Hashing
Rong Kang, Yue Cao, Mingsheng Long, Jianmin Wang, Philip S. Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8252-8261
Deep hashing enables computation and memory efficient image search through end-to-end learning of feature representations and binary codes. While linear scan over binary hash codes is more efficient than over the high-dimensional representations, its linear-time complexity is still unacceptable for very large databases. Hamming space retrieval enables constant-time search through hash lookups, where for each query, there is a Hamming ball centered at the query and the data points within the ball are returned as relevant. Since inside the Hamming ball implies retrievable while outside irretrievable, it is crucial to explicitly characterize the Hamming ball. The main idea of this work is to directly embody the Hamming radius into the loss functions, leading to Maximum-Margin Hamming Hashing (MMHH), a new model specifically optimized for Hamming space retrieval. We introduce a max-margin t-distribution loss, where the t-distribution concentrates more similar data points to be within the Hamming ball, and the margin characterizes the Hamming radius such that less penalization is applied to similar data points within the Hamming ball. The loss function also introduces robustness to data noise, where the similarity supervision may be inaccurate in practical problems. The model is trained end-to-end using a new semi-batch optimization algorithm tailored to extremely imbalanced data. Our method yields state-of-the-art results on four datasets and shows superior performance on noisy data.
****

Conservative Wasserstein Training for Pose Estimation
Xiaofeng Liu, Yang Zou, Tong Che, Peng Ding, Ping Jia, Jane You, B.V.K. Vijaya Kumar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8262-8272
This paper targets the task with discrete and periodic class labels (e.g., pose/orientation estimation) in the context of deep learning. The commonly used cross

-entropy or regression loss is not well matched to this problem as they ignore t
he periodic nature of the labels and the class similarity, or assume labels are
continuous value. We propose to incorporate inter-class correlations in a Wasser
stein training framework by pre-defining (i.e., using arc length of a circle) or
 adaptively learning the ground metric. We extend the ground metric as a linear,
 convex or concave increasing function w.r.t. arc length from an optimization pe
rspective. We also propose to construct the conservative target labels which mod
el the inlier and outlier noises using a wrapped unimodal-uniform mixture distri
bution. Unlike the one-hot setting, the conservative label makes the computation
 of Wasserstein distance more challenging. We systematically conclude the practi
cal closed-form solution of Wasserstein distance for pose data with either one-h
ot or conservative target label. We evaluate our method on head, body, vehicle a
nd 3D object pose benchmarks with exhaustive ablation studies. The Wasserstein l
oss obtaining superior performance over the current methods, especially using co
nvex mapping function for ground metric, conservative label, and closed-form sol
ution.
****

Learning to Rank Proposals for Object Detection
Zhiyu Tan, Xuecheng Nie, Qi Qian, Nan Li, Hao Li; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2019, pp. 8273-8281
Non-Maximum Suppression (NMS) is an essential step of modern object detection mo
dels for removing duplicated candidates. The efficacy of NMS heavily affects the
 final detection results. Prior works exploit suppression criterions relying on
either the objectiveness derived from classification or the overlapness produced
 by regression, both of which are heuristically designed and fail to explicitly
link with the suppression rank. To address this issue, in this paper, we propose
 a novel Learning-to-Rank (LTR) model to produce the suppression rank via a lear
ning procedure, thus facilitating the candidate generation and lifting the detec
tion performance. In particular, we define a ranking score based on IoU to indic
ate the ranks of candidates during the NMS step, where candidates with high rank
ing score will be reserved and the ones with low ranking score will be eliminate
d. We design a lightweight network to predict the ranking score. We introduce a
ranking loss to supervise the generation of these ranking scores, which encourag
es candidates with IoU to the ground-truth to rank higher. To facilitate the tra
ining procedure, we design a novel sampling strategy via dividing candidates int
o different levels and select hard pairs to adopt in the training. During the in
ference phase, this module can be exploited as a plugin to the current object de
tector. The training and inference of the overall framework is end-to-end. Compr
ehensive experiments on benchmarks PASCAL VOC and MS COCO demonstrate the genera
lity and effectiveness of our model for facilitating existing object detectors t
o state-of-the-art accuracy.
****

Vehicle Re-Identification With Viewpoint-Aware Metric Learning
Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, Yichen Wei; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
pp. 8282-8291
This paper considers vehicle re-identification (re-ID) problem. The extreme view
point variation (up to 180 degrees) poses great challenges for existing approach
es. Inspired by the behavior in human's recognition process, we propose a novel
viewpoint-aware metric learning approach. It learns two metrics for similar view
points and different viewpoints in two feature spaces, respectively, giving rise
 to viewpoint-aware network (VANet). During training, two types of constraints a
re applied jointly. During inference, viewpoint is firstly estimated and the cor
responding metric is used. Experimental results confirm that VANet significantly
 improves re-ID accuracy, especially when the pair is observed from different vi
ewpoints. Our method establishes the new state-of-the-art on two benchmarks.
****

WSOD2: Learning Bottom-Up and Top-Down Objectness Distillation for Weakly-Superv
ised Object Detection
Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, Lei Zhang; Proceedings o

f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 829
2-8300
We study on weakly-supervised object detection (WSOD) which plays a vital role i
n relieving human involvement from object-level annotations. Predominant works i
ntegrate region proposal mechanisms with convolutional neural networks (CNN). Al
though CNN is proficient in extracting discriminative local features, grand chal
lenges still exist to measure the likelihood of a bounding box containing a comp
lete object (i.e., "objectness"). In this paper, we propose a novel WSOD framewo
rk with Objectness Distillation (i.e., WSOD2) by designing a tailored training m
echanism for weakly-supervised object detection. Multiple regression targets are
 specifically determined by jointly considering bottom-up (BU) and top-down (TD)
 objectness from low-level measurement and CNN confidences with an adaptive line
ar combination. As bounding box regression can facilitate a region proposal lear
ning to approach its regression target with high objectness during training, dee
p objectness representation learned from bottom-up evidences can be gradually di
stilled into CNN by optimization. We explore different adaptive training curves
for BU/TD objectness, and show that the proposed WSOD2 can achieve state-of-the-
art results.
****

Localization of Deep Inpainting Using High-Pass Fully Convolutional Network
Haodong Li, Jiwu Huang; Proceedings of the IEEE/CVF International Conference on
 Computer Vision (ICCV), 2019, pp. 8301-8310
Image inpainting has been substantially improved with deep learning in the past
years. Deep inpainting can fill image regions with plausible contents, which are
 not visually apparent. Although inpainting is originally designed to repair ima
ges, it can even be used for malicious manipulations, e.g., removal of specific
objects. Therefore, it is necessary to identify the presence of inpainting in an
 image. This paper presents a method to locate the regions manipulated by deep i
npainting. The proposed method employs a fully convolutional network that is bas
ed on high-pass filtered image residuals. Firstly, we analyze and observe that t
he inpainted regions are more distinguishable from the untouched ones in the res
idual domain. Hence, a high-pass pre-filtering module is designed to get image r
esiduals for enhancing inpainting traces. Then, a feature extraction module, whi
ch learns discriminative features from image residuals, is built with four conca
tenated ResNet blocks. The learned feature maps are finally enlarged by an up-sa
mpling module, so that a pixel-wise inpainting localization map is obtained. The
 whole network is trained end-to-end with a loss addressing the class imbalance.
 Extensive experimental results evaluated on both synthetic and realistic images
 subjected to deep inpainting have shown the effectiveness of the proposed metho
d.
****

Clustered Object Detection in Aerial Images
Fan Yang, Heng Fan, Peng Chu, Erik Blasch, Haibin Ling; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8311-8320
Detecting objects in aerial images is challenging for at least two reasons: (1)
target objects like pedestrians are very small in pixels, making them hardly dis
tinguished from surrounding background; and (2) targets are in general sparsely
and non-uniformly distributed, making the detection very inefficient. In this pa
per, we address both issues inspired by observing that these targets are often c
lustered. In particular, we propose a Clustered Detection (ClusDet) network that
 unifies object clustering and detection in an end-to-end framework. The key com
ponents in ClusDet include a cluster proposal sub-network (CPNet), a scale estim
ation sub-network (ScaleNet), and a dedicated detection network (DetecNet). Give
n an input image, CPNet produces object cluster regions and ScaleNet estimates o
bject scales for these regions. Then, each scale-normalized cluster region is fe
d into DetecNet for object detection. ClusDet has several advantages over previo
us solutions: (1) it greatly reduces the number of chips for final object detect
ion and hence achieves high running time efficiency, (2) the cluster-based scale
 estimation is more accurate than previously used single-object based ones, henc
e effectively improves the detection for small objects, and (3) the final DetecN

et is dedicated for clustered regions and implicitly models the prior context in formation so as to boost detection accuracy. The proposed method is tested on th ree popular aerial image datasets including VisDrone, UAVDT and DOTA. In all exp eriments, ClusDet achieves promising performance in comparison with state-of-the -art detectors.
****
Unsupervised Graph Association for Person Re-Identification
Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, Stan Z. Li; Proceed ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, p p. 8321-8330
In this paper, we propose an unsupervised graph association (UGA) framework to l earn the underlying viewinvariant representations from the video pedestrian trac klets. The core points of UGA are mining the underlying cross-view associations and reducing the damage of noise associations. To this end, UGA is adopts a two- stage training strategy: (1) intra-camera learning stage and (2) intercamera lea rning stage. The former learns the intra-camera representation for each camera. While the latter builds a cross-view graph (CVG) to associate different cameras. By doing this, we can learn view-invariant representation for all person. Exten sive experiments and ablation studies on seven re-id datasets demonstrate the su periority of the proposed UGA over most state-of-the-art unsupervised and domain adaptation re-id methods.
****
Learning a Mixture of Granularity-Specific Experts for Fine-Grained Categorizati on
Lianbo Zhang, Shaoli Huang, Wei Liu, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8331-8340
We aim to divide the problem space of fine-grained recognition into some specifi c regions. To achieve this, we develop a unified framework based on a mixture of experts. Due to limited data available for the fine-grained recognition problem , it is not feasible to learn diverse experts by using a data division strategy. To tackle the problem, we promote diversity among experts by combing an expert gradually-enhanced learning strategy and a Kullback-Leibler divergence based con straint. The strategy learns new experts on the dataset with the prior knowledge from former experts and adds them to the model sequentially, while the introduc ed constraint forces the experts to produce diverse prediction distribution. The se drive the experts to learn the task from different aspects, making them speci alized in different subspace problems. Experiments show that the resulting model improves the classification performance and achieves the state-of-the-art perfo rmance on several fine-grained benchmark datasets.
****
advPattern: Physical-World Attacks on Deep Person Re-Identification via Adversar ially Transformable Patterns
Zhibo Wang, Siyan Zheng, Mengkai Song, Qian Wang, Alireza Rahimpour, Hairon g Qi; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2019, pp. 8341-8350
Person re-identification (re-ID) is the task of matching person images across ca mera views, which plays an important role in surveillance and security applicati ons. Inspired by great progress of deep learning, deep re-ID models began to be popular and gained state-of-the-art performance. However, recent works found tha t deep neural networks (DNNs) are vulnerable to adversarial examples, posing pot ential threats to DNNs based applications. This phenomenon throws a serious ques tion about whether deep re-ID based systems are vulnerable to adversarial attack s. In this paper, we take the first attempt to implement robust physical-world a ttacks against deep re-ID. We propose a novel attack algorithm, called advPatter n, for generating adversarial patterns on clothes, which learns the variations o f image pairs across cameras to pull closer the image features from the same cam era, while pushing features from different cameras farther. By wearing our craft ed "invisible cloak", an adversary can evade person search, or impersonate a tar get person to fool deep re-ID models in physical world. We evaluate the effectiv eness of our transformable patterns on adversaries' clothes with Market1501 and

our established PRCS dataset. The experimental results show that the rank-1 accu racy of re-ID models for matching the adversary decreases from 87.9% to 27.1% un der Evading Attack. Furthermore, the adversary can impersonate a target person w ith 47.1% rank-1 accuracy and 67.9% mAP under Impersonation Attack. The results demonstrate that deep re-ID systems are vulnerable to our physical attacks.
****

ABD-Net: Attentive but Diverse Person Re-Identification
Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8351-8361
Attention mechanisms have been found effective for person re-identification (Re-ID). However, the learned "attentive" features are often not naturally uncorrela ted or "diverse", which compromises the retrieval performance based on the Eucli dean distance. We advocate the complementary powers of attention and diversity f or Re-ID, by proposing an Attentive but Diverse Network (ABD-Net). ABD-Net seaml essly integrates attention modules and diversity regularizations throughout the entire network to learn features that are representative, robust, and more discr iminative. Specifically, we introduce a pair of complementary attention modules, focusing on channel aggregation and position awareness, respectively. Then, we plug in a novel orthogonality constraint that efficiently enforces diversity on both hidden activations and weights. Through an extensive set of ablation study, we verify that the attentive and diverse terms each contributes to the performa nce boosts of ABD-Net. It consistently outperforms existing state-of-the-art met hods on there popular person Re-ID benchmarks.
****

From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer
Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, Chunhua Shen; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 019, pp. 8362-8371
Visual counting, a task that predicts the number of objects from an image/video, is an open-set problem by nature, i.e., the number of population can vary in [0 ,+[?]) in theory. However, the collected images and labeled count values are lim ited in reality, which means only a small closed set is observed. Existing metho ds typically model this task in a regression manner, while they are likely to su ffer from an unseen scene with counts out of the scope of the closed set. In fac t, counting is decomposable. A dense region can always be divided until the coun t values of sub-regions are within the previously observed closed set. Inspired by this idea, we propose a simple but effective approach, Spatial Divide-and-Con quer Network (S-DCNet). S-DCNet learns to classify closed-set counts and can gen eralize to open-set counts via S-DC. S-DCNet is also efficient. To avoid repeate dly computing sub-region convolutional features, S-DC is executed on the feature map instead of on the input image. S-DCNet achieves the state-of-the-art perfor mance on three crowd counting datasets (ShanghaiTech, UCF_CC_50 and UCF-QNRF), a vehicle counting dataset (TRANCOS) and a plant counting dataset (MTC). Compared to the previous best methods, S-DCNet brings a 20.2% relative improvement on th e ShanghaiTechPart B, 20.9% on the UCF-QNRF, 22.5% on the TRANCOS and 15.1% on t he MTC. Code has been made available at: https://github.com/xhp-hust-2018-2011/S -DCNet.
****

Towards Precise End-to-End Weakly Supervised Object Detection Network
Ke Yang, Dongsheng Li, Yong Dou; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2019, pp. 8372-8381
It is challenging for weakly supervised object detection network to precisely pr edict the positions of the objects, since there are no instance-level category a nnotations. Most existing methods tend to solve this problem by using a two-phas e learning procedure, i.e., multiple instance learning detector followed by a fu lly supervised learning detector with bounding-box regression. Based on our obse rvation, this procedure may lead to local minima for some object categories. In this paper, we propose to jointly train the two phases in an end-to-end manner t o tackle this problem. Specifically, we design a single network with both multip

le instance learning and bounding-box regression branches that share the same backbone. Meanwhile, a guided attention module using classification loss is added to the backbone for effectively extracting the implicit location information in the features. Experimental results on public datasets show that our method achieves state-of-the-art performance.
****

Learn to Scale: Generating Multipolar Normalized Density Maps for Crowd Counting
Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, Xiang Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8382-8390
Dense crowd counting aims to predict thousands of human instances from an image, by calculating integrals of a density map over image pixels. Existing approaches mainly suffer from the extreme density variations. Such density pattern shift poses challenges even for multi-scale model ensembling. In this paper, we propose a simple yet effective approach to tackle this problem. First, a patch-level density map is extracted by a density estimation model and further grouped into several density levels which are determined over full datasets. Second, each patch density map is automatically normalized by an online center learning strategy with a multipolar center loss. Such a design can significantly condense the density distribution into several clusters, and enable that the density variance can be learned by a single model. Extensive experiments demonstrate the superiority of the proposed method. Our work outperforms the state-of-the-art by 4.2%, 14.3%, 27.1% and 20.1% in MAE, on the ShanghaiTech Part A, ShanghaiTech Part B, UCF_CC_50 and UCF-QNRF datasets, respectively.
****

Ground-to-Aerial Image Geo-Localization With a Hard Exemplar Reweighting Triplet Loss
Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, Gongjian Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8391-8400
The task of ground-to-aerial image geo-localization can be achieved by matching a ground view query image to a reference database of aerial/satellite images. It is highly challenging due to the dramatic viewpoint changes and unknown orientations. In this paper, we propose a novel in-batch reweighting triplet loss to emphasize the positive effect of hard exemplars during end-to-end training. We also integrate an attention mechanism into our model using feature-level contextual information. To analyze the difficulty level of each triplet, we first enforce a modified logistic regression to triplets with a distance rectifying factor. Then, the reference negative distances for corresponding anchors are set, and the relative weights of triplets are computed by comparing their difficulty to the corresponding references. To reduce the influence of extreme hard data and less useful simple exemplars, the final weights are pruned using upper and lower bound constraints. Experiments on two benchmark datasets show that the proposed approach significantly outperforms the state-of-the-art methods.
****

Learning to Discover Novel Visual Categories via Deep Transfer Clustering
Kai Han, Andrea Vedaldi, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8401-8409
We consider the problem of discovering novel object categories in an image collection. While these images are unlabelled, we also assume prior knowledge of related but different image classes. We use such prior knowledge to reduce the ambiguity of clustering, and improve the quality of the newly discovered classes. Our contributions are twofold. The first contribution is to extend Deep Embedded Clustering to a transfer learning setting; we also improve the algorithm by introducing a representation bottleneck, temporal ensembling, and consistency. The second contribution is a method to estimate the number of classes in the unlabelled data. This also transfers knowledge from the known classes, using them as probes to diagnose different choices for the number of classes in the unlabelled subset. We thoroughly evaluate our method, substantially outperforming state-of-the-art techniques in a large number of benchmarks, including ImageNet, OmniGlot, CI

FAR-100, CIFAR-10, and SVHN.
****
AM-LFS: AutoML for Loss Function Search
Chuming Li, Xin Yuan, Chen Lin, Minghao Guo, Wei Wu, Junjie Yan, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8410-8419
Designing an effective loss function plays an important role in visual analysis. Most existing loss function designs rely on hand-crafted heuristics that require domain experts to explore the large design space, which is usually sub-optimal and time-consuming. In this paper, we propose AutoML for Loss Function Search (AM-LFS) which leverages REINFORCE to search loss functions during the training process. The key contribution of this work is the design of search space which can guarantee the generalization and transferability on different vision tasks by including a bunch of existing prevailing loss functions in a unified formulation. We also propose an efficient optimization framework which can dynamically optimize the parameters of loss function's distribution during training. Extensive experimental results on four benchmark datasets show that, without any tricks, our method outperforms existing hand-crafted loss functions in various computer vision tasks.
****
Few-Shot Object Detection via Feature Reweighting
Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, Trevor Darrell; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8420-8429
Conventional training of a deep CNN based object detector demands a large number of bounding box annotations, which may be unavailable for rare categories. In this work we develop a few-shot object detector that can learn to detect novel objects from only a few annotated examples. Our proposed model leverages fully labeled base classes and quickly adapts to novel classes, using a meta feature learner and a reweighting module within a one-stage detection architecture. The feature learner extracts meta features that are generalizable to detect novel object classes, using training data from base classes with sufficient samples. The reweighting module transforms a few support examples from the novel classes to a global vector that indicates the importance or relevance of meta features for detecting the corresponding objects. These two modules, together with a detection prediction module, are trained end-to-end based on an episodic few-shot learning scheme and a carefully designed loss function. Through extensive experiments we demonstrate that our model outperforms well-established baselines by a large margin for few-shot object detection, on multiple datasets and settings. We also present analysis on various aspects of our proposed model, aiming to provide some inspiration for future few-shot detection works.
****
Objects365: A Large-Scale, High-Quality Dataset for Object Detection
Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, Jian Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8430-8439
In this paper, we introduce a new large-scale object detection dataset, Objects365, which has 365 object categories over 600K training images. More than 10 million, high-quality bounding boxes are manually labeled through a three-step, carefully designed annotation pipeline. It is the largest object detection dataset (with full annotation) so far and establishes a more challenging benchmark for the community. Objects365 can serve as a better feature learning dataset for localization-sensitive tasks like object detection and semantic segmentation. The Objects365 pre-trained models significantly outperform ImageNet pre-trained models with 5.6 points gain (42 vs 36.4) based on the standard setting of 90K iterations on COCO benchmark. Even compared with much long training time like 540K iterations, our Objects365 pretrained model with 90K iterations still have 2.7 points gain (42 vs 39.3). Meanwhile, the finetuning time can be greatly reduced (up to 10 times) when reaching the same accuracy. Better generalization ability of Object365 has also been verified on CityPersons, VOC segmentation, and ADE tasks. Th

e dataset as well as the pretrained-models have been released at www.objects365.org.
****

Efficient and Accurate Arbitrary-Shaped Text Detection With Pixel Aggregation Network
Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8440-8449
Scene text detection, an important step of scene text reading systems, has witnessed rapid development with convolutional neural networks. Nonetheless, two main challenges still exist and hamper its deployment to real-world applications. The first problem is the trade-off between speed and accuracy. The second one is to model the arbitrary-shaped text instance. Recently, some methods have been proposed to tackle arbitrary-shaped text detection, but they rarely take the speed of the entire pipeline into consideration, which may fall short in practical applications. In this paper, we propose an efficient and accurate arbitrary-shaped text detector, termed Pixel Aggregation Network (PAN), which is equipped with a low computational-cost segmentation head and a learnable post-processing. More specifically, the segmentation head is made up of Feature Pyramid Enhancement Module (FPEM) and Feature Fusion Module (FFM). FPEM is a cascadable U-shaped module, which can introduce multi-level information to guide the better segmentation. FFM can gather the features given by the FPEMs of different depths into a final feature for segmentation. The learnable post-processing is implemented by Pixel Aggregation (PA), which can precisely aggregate text pixels by predicted similarity vectors. Experiments on several standard benchmarks validate the superiority of the proposed PAN. It is worth noting that our method can achieve a competitive F-measure of 79.9% at 84.2 FPS on CTW1500.
****

Foreground-Aware Pyramid Reconstruction for Alignment-Free Occluded Person Re-Identification
Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8450-8459
Re-identifying a person across multiple disjoint camera views is important for intelligent video surveillance, smart retailing and many other applications. However, existing person re-identification methods are challenged by the ubiquitous occlusion over persons and suffer performance degradation. This paper proposes a novel occlusion-robust and alignment-free model for occluded person ReID and extends its application to realistic and crowded scenarios. The proposed model first leverages the fully convolution network (FCN) and pyramid pooling to extract spatial pyramid features. Then an alignment-free matching approach namely Foreground-aware Pyramid Reconstruction (FPR) is developed to accurately compute matching scores between occluded persons, regardless of their different scales and sizes. FPR uses the error from robust reconstruction over spatial pyramid features to measure similarities between two persons. More importantly, we design a occlusion-sensitive foreground probability generator that focuses more on clean human body parts to robustify the similarity computation with less contamination from occlusion. The FPR is easily embedded into any end-to-end person ReID models. The effectiveness of the proposed method is clearly demonstrated by the experimental results (Rank-1 accuracy) on three occluded person datasets: Partial REID (78.30%), Partial iLIDS (68.08%), Occluded REID (81.00%), and three benchmark person datasets: Market1501 (95.42%), DukeMTMC (88.64%), CUHK03 (76.08%).
****

Collect and Select: Semantic Alignment Metric Learning for Few-Shot Learning
Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8460-8469
Few-shot learning aims to learn latent patterns from few training examples and has shown promises in practice. However, directly calculating the distances between the query image and support image in existing methods may cause ambiguity bec

ause dominant objects can locate anywhere on images. To address this issue, this paper proposes a Semantic Alignment Metric Learning (SAML) method for few-shot learning that aligns the semantically relevant dominant objects through a "collect-and-select" strategy. Specifically, we first calculate a relation matrix (RM) to "collect" the distances of each local region pairs of the 3D tensor extracted from a query image and the mean tensor of the support images. Then, the attention technique is adapted to "select" the semantically relevant pairs and put more weights on them. Afterwards, a multi-layer perceptron (MLP) is utilized to map the reweighted RMs to their corresponding similarity scores. Theoretical analysis demonstrates the generalization ability of SAML and gives a theoretical guarantee. Empirical results demonstrate that semantic alignment is achieved. Extensive experiments on benchmark datasets validate the strengths of the proposed approach and demonstrate that SAML significantly outperforms the current state-of-the-art methods. The source code is available at https://github.com/haofusheng/SAML.

****

Bayesian Adaptive Superpixel Segmentation

Roy Uziel, Meitar Ronen, Oren Freifeld; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8470-8479

Superpixels provide a useful intermediate image representation. Existing superpixel methods, however, suffer from at least some of the following drawbacks: 1) topology is handled heuristically; 2) the number of superpixels is either predefined or estimated at a prohibitive cost; 3) lack of adaptiveness. As a remedy, we propose a novel probabilistic model, self-coined Bayesian Adaptive Superpixel Segmentation (BASS), together with an efficient inference. BASS is a Bayesian nonparametric mixture model that also respects topology and favors spatial coherence. The optimizationbased and topology-aware inference is parallelizable and implemented in GPU. Quantitatively, BASS achieves results that are either better than the state-of-the-art or close to it, depending on the performance index and/or dataset. Qualitatively, we argue it achieves the best results; we demonstrate this by not only subjective visual inspection but also objective quantitative performance evaluation of the downstream application of face detection. Our code is available at https://github.com/uzielroy/BASS.

****

CapsuleVOS: Semi-Supervised Video Object Segmentation Using Capsule Routing

Kevin Duarte, Yogesh S. Rawat, Mubarak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8480-8489

In this work we propose a capsule-based approach for semi-supervised video object segmentation. Current video object segmentation methods are frame-based and often require optical flow to capture temporal consistency across frames which can be difficult to compute. To this end, we propose a video based capsule network, CapsuleVOS, which can segment several frames at once conditioned on a reference frame and segmentation mask. This conditioning is performed through a novel routing algorithm for attention-based efficient capsule selection. We address two challenging issues in video object segmentation: 1) segmentation of small objects and 2) occlusion of objects across time. The issue of segmenting small objects is addressed with a zooming module which allows the network to process small spatial regions of the video. Apart from this, the framework utilizes a novel memory module based on recurrent networks which helps in tracking objects when they move out of frame or are occluded. The network is trained end-to-end and we demonstrate its effectiveness on two benchmark video object segmentation datasets; it outperforms current offline approaches on the Youtube-VOS dataset while having a run-time that is almost twice as fast as competing methods. The code is publicly available at https://github.com/KevinDuarte/CapsuleVOS.

****

BAE-NET: Branched Autoencoder for Shape Co-Segmentation

Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, Hao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8490-8499

We treat shape co-segmentation as a representation learning problem and introduc

e BAE-NET, a branched autoencoder network, for the task. The unsupervised BAE-NE T is trained with a collection of un-segmented shapes, using a shape reconstruct ion loss, without any ground-truth labels. Specifically, the network takes an in put shape and encodes it using a convolutional neural network, whereas the decod er concatenates the resulting feature code with a point coordinate and outputs a value indicating whether the point is inside/outside the shape. Importantly, th e decoder is branched: each branch learns a compact representation for one commo nly recurring part of the shape collection, e.g., airplane wings. By complementi ng the shape reconstruction loss with a label loss, BAE-NET is easily tuned for one-shot learning. We show unsupervised, weakly supervised, and one-shot learnin g results by BAE-NET, demonstrating that using only a couple of exemplars, our n etwork can generally outperform state-of-the-art supervised methods trained on h undreds of segmented shapes. Code is available at https://github.com/czq142857/B AE-NET.

****

VV-Net: Voxel VAE Net With Group Convolutions for Point Cloud Segmentation

Hsien-Yu Meng, Lin Gao, Yu-Kun Lai, Dinesh Manocha; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2019, pp. 8500-8508

We present a novel algorithm for point cloud segmentation.Our approach transform s unstructured point clouds into regular voxel grids, and further uses a kernel-based interpolated variational autoencoder (VAE) architecture to encode the loca l geometry within each voxel.Traditionally, the voxel representation only compri ses Boolean occupancy information, which fails to capture the sparsely distribut ed points within voxels in a compact manner. In order to handle sparse distribut ions of points, we further employ radial basis functions (RBF) to compute a loca l, continuous representation within each voxel. Our approach results in a good v olumetric representation that effectively tackles noisy point cloud datasets and is more robust for learning. Moreover, we further introduce group equivariant C NN to 3D, by defining the convolution operator on a symmetry group acting on Z ^3 and its isomorphic sets. This improves the expressive capacity without increa sing parameters, leading to more robust segmentation results.We highlight the pe rformance on standard benchmarks and show that our approach outperforms state-of -the-art segmentation algorithms on the ShapeNet and S3DIS datasets.

****

Miss Detection vs. False Alarm: Adversarial Learning for Small Object Segmentati on in Infrared Images

Huan Wang, Luping Zhou, Lei Wang; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2019, pp. 8509-8518

A key challenge of infrared small object segmentation (ISOS) is to balance miss detection (MD) and false alarm (FA). This usually needs "opposite" strategies to suppress the two terms, and has not been well resolved in the literature. In th is paper, we propose a deep adversarial learning framework to improve this situa tion. Departing from the tradition of jointly reducing MD and FA via a single ob jective, we decompose this difficult task into two sub-tasks handled by two mode ls trained adversarially, with each focusing on reducing either MD or FA. Such a new design brings forth at least three advantages. First, as each model focuses on a relatively simpler sub-task, the overall difficulty of ISOS is somehow dec reased. Second, the adversarial training of the two models naturally produces a delicate balance of MD and FA, and low rates for both MD and FA could be achieve d at Nash equilibrium. Third, this MD-FA detachment gives us more flexibility to develop specific models dedicated to each sub-task. To realize the above design , we propose a conditional Generative Adversarial Network comprising of two gene rators and one discriminator. Each generator strives for one sub-task, while the discriminator differentiates the three segmentation results from the two genera tors and the ground truth. Moreover, in order to better serve the sub-tasks, the two generators, based on context aggregation networks, utilzse different size o f receptive fields, providing both local and global views of objects for segment ation. As verified on multiple infrared image data sets, our method consistently achieves better segmentation than many state-of-the-art ISOS methods.

****

Group-Wise Deep Object Co-Segmentation With Co-Attention Recurrent Neural Network

Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, Anqi Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8519-8528

Effective feature representations which should not only express the images individual properties, but also reflect the interaction among group images are essentially crucial for real-world co-segmentation. This paper proposes a novel end-to-end deep learning approach for group-wise object co-segmentation with a recurrent network architecture. Specifically, the semantic features extracted from a pre-trained CNN of each image are first processed by single image representation branch to learn the unique properties. Meanwhile, a specially designed Co-Attention Recurrent Unit (CARU) recurrently explores all images to generate the final group representation by using the co-attention between images, and simultaneously suppresses noisy information. The group feature which contains synergetic information is broadcasted to each individual image and fused with multi-scale fine-resolution features to facilitate the inferring of co-segmentation. Moreover, we propose a groupwise training objective to utilize the co-object similarity and figure-ground distinctness as the additional supervision. The whole modules are collaboratively optimized in an end-to-end manner, further improving the robustness of the approach. Comprehensive experiments on three benchmarks can demonstrate the superiority of our approach in comparison with the state-of-the-art methods.

****

Human Attention in Image Captioning: Dataset and Analysis

Sen He, Hamed R. Tavakoli, Ali Borji, Nicolas Pugeault; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8529-8538

In this work, we present a novel dataset consisting of eye movements and verbal descriptions recorded synchronously over images. Using this data, we study the differences in human attention during free-viewing and image captioning tasks. We look into the relationship between human atten- tion and language constructs during perception and sen- tence articulation. We also analyse attention deployment mechanisms in the top-down soft attention approach that is argued to mimic human attention in captioning tasks, and investigate whether visual saliency can help image caption- ing. Our study reveals that (1) human attention behaviour differs in free-viewing and image description tasks. Hu- mans tend to fixate on a greater variety of regions under the latter task, (2) there is a strong relationship between de- scribed objects and attended objects (97% of the described objects are being attended), (3) a convolutional neural net- work as feature encoder accounts for human-attended re- gions during image captioning to a great extent (around 78%), (4) soft-attention mechanism differs from human at- tention, both spatially and temporally, and there is low correlation between caption scores and attention consis- tency scores. These indicate a large gap between humans and machines in regards to top-down attention, and (5) by integrating the soft attention model with image saliency, we can significantly improve the model's performance on Flickr30k and MSCOCO benchmarks. The dataset can be found at: https://github.com/SenHe/ Human-Attention-in-Image-Captioning.

****

Variational Uncalibrated Photometric Stereo Under General Lighting

Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Queau, Daniel Cremers; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8539-8548

Photometric stereo (PS) techniques nowadays remain constrained to an ideal laboratory setup where modeling and calibration of lighting is amenable. To eliminate such restrictions, we propose an efficient principled variational approach to uncalibrated PS under general illumination. To this end, the Lambertian reflectance model is approximated through a spherical harmonic expansion, which preserves the spatial invariance of the lighting. The joint recovery of shape, reflectance and illumination is then formulated as a single variational problem. There the shape estimation is carried out directly in terms of the underlying perspective depth map, thus implicitly ensuring integrability and bypassing the need for a

subsequent normal integration. To tackle the resulting nonconvex problem numeric ally, we undertake a two-phase procedure to initialize a balloon-like perspectiv e depth map, followed by a "lagged" block coordinate descent scheme. The experim ents validate efficiency and robustness of this approach. Across a variety of ev aluations, we are able to reduce the mean angular error consistently by a factor of 2-3 compared to the state-of-the-art.
****

SPLINE-Net: Sparse Photometric Stereo Through Lighting Interpolation and Normal Estimation Networks
Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, Alex C. Kot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8549-8558
This paper solves the Sparse Photometric stereo through Lighting Interpolation a nd Normal Estimation using a generative Network (SPLINE-Net). SPLINE-Net contain s a lighting interpolation network to generate dense lighting observations given a sparse set of lights as inputs followed by a normal estimation network to est imate surface normals. Both networks are jointly constrained by the proposed sym metric and asymmetric loss functions to enforce isotropic constrain and perform outlier rejection of global illumination effects. SPLINE-Net is verified to outp erform existing methods for photometric stereo of general BRDFs by using only te n images of different lights instead of using nearly one hundred images.
****

Hyperspectral Image Reconstruction Using Deep External and Internal Learning
Tao Zhang, Ying Fu, Lizhi Wang, Hua Huang; Proceedings of the IEEE/CVF Intern ational Conference on Computer Vision (ICCV), 2019, pp. 8559-8568
To solve the low spatial and/or temporal resolution problem which the convention al hypelrspectral cameras often suffer from, coded snapshot hyperspectral imagin g systems have attracted more attention recently. Recovering a hyperspectral ima ge (HSI) from its corresponding coded image is an ill-posed inverse problem, and learning accurate prior of HSI is essential to solve this inverse problem. In t his paper, we present an effective convolutional neural network (CNN) based meth od for coded HSI reconstruction, which learns the deep prior from the external d ataset as well as the internal information of input coded image with spatial-spe ctral constraint. Our method can effectively exploit spatial-spectral correlatio n and sufficiently represent the variety nature of HSIs. Experimental results sh ow our method outperforms the state-of-the-art methods under both comprehensive quantitative metrics and perceptive quality.
****

Gravity as a Reference for Estimating a Person's Height From Video
Didier Bieler, Semih Gunel, Pascal Fua, Helge Rhodin; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2019, pp. 8569-8577
Estimating the metric height of a person from monocular imagery without addition al assumptions is ill-posed. Existing solutions either require manual calibratio n of ground plane and camera geometry, special cameras, or reference objects of known size. We focus on motion cues and exploit gravity on earth as an omniprese nt reference 'object' to translate acceleration, and subsequently height, measur ed in image-pixels to values in meters. We require videos of motion as input, wh ere gravity is the only external force. This limitation is different to those of existing solutions that recover a person's height and, therefore, our method op ens up new application fields. We show theoretically and empirically that a simp le motion trajectory analysis suffices to translate from pixel measurements to t he person's metric height, reaching a MAE of up to 3.9 cm on jumping motions, an d that this works without camera and ground plane calibration.
****

Shadow Removal via Shadow Image Decomposition
Hieu Le, Dimitris Samaras; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8578-8587
We propose a novel deep learning method for shadow removal. Inspired by physical models of shadow formation, we use a linear illumination transformation to mode l the shadow effects in the image that allows the shadow image to be expressed a

s a combination of the shadow-free image, the shadow parameters, and a matte lay
er. We use two deep networks, namely SP-Net and M-Net, to predict the shadow par
ameters and the shadow matte respectively. This system allows us to remove the s
hadow effects on the images. We train and test our framework on the most challen
ging shadow removal dataset (ISTD). Compared to the state-of-the-art method, our
 model achieves a 40% error reduction in terms of root mean square error (RMSE)
for the shadow area, reducing RMSE from 13.3 to 7.9. Moreover, we create an augm
ented ISTD dataset based on an image decomposition system by modifying the shado
w parameters to generate new synthetic shadow images. Training our model on this
 new augmented ISTD dataset further lowers the RMSE on the shadow area to 7.4.
****

OperatorNet: Recovering 3D Shapes From Difference Operators
Ruqi Huang, Marie-Julie Rakotosaona, Panos Achlioptas, Leonidas J. Guibas, M
aks Ovsjanikov; Proceedings of the IEEE/CVF International Conference on Computer
 Vision (ICCV), 2019, pp. 8588-8597
This paper proposes a learning-based framework for reconstructing 3D shapes from
 functional operators, compactly encoded as small-sized matrices. To this end we
 introduce a novel neural architecture, called OperatorNet, which takes as input
 a set of linear operators representing a shape and produces its 3D embedding. W
e demonstrate that this approach significantly outperforms previous purely geome
tric methods for the same problem. Furthermore, we introduce a novel functional
operator, which encodes the extrinsic or pose-dependent shape information, and t
hus complements purely intrinsic pose-oblivious operators, such as the classical
 Laplacian. Coupled with this novel operator, our reconstruction network achieve
s very high reconstruction accuracy, even in the presence of incomplete informat
ion about a shape, given a soft or functional map expressed in a reduced basis.
Finally, we demonstrate that the multiplicative functional algebra enjoyed by th
ese operators can be used to synthesize entirely new unseen shapes, in the conte
xt of shape interpolation and shape analogy applications.
****

Neural Inverse Rendering of an Indoor Scene From a Single Image
Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, Jan
 Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2019, pp. 8598-8607
Inverse rendering aims to estimate physical attributes of a scene, e.g., reflect
ance, geometry, and lighting, from image(s). Inverse rendering has been studied
primarily for single objects or with methods that solve for only one of the scen
e attributes. We propose the first learning based approach that jointly estimate
s albedo, normals, and lighting of an indoor scene from a single image. Our key
contribution is the Residual Appearance Renderer (RAR), which can be trained to
synthesize complex appearance effects (e.g., inter-reflection, cast shadows, nea
r-field illumination, and realistic shading), which would be neglected otherwise
. This enables us to perform self-supervised learning on real data using a recon
struction loss, based on re-synthesizing the input image from the estimated comp
onents. We finetune with real data after pretraining with synthetic data. To thi
s end, we use physically-based rendering to create a large-scale synthetic datas
et, named SUNCG-PBR, which is a significant improvement over prior datasets. Exp
erimental results show that our approach outperforms state-of-the-art methods th
at estimate one or more scene attributes.
****

ForkNet: Multi-Branch Volumetric Semantic Completion From a Single Depth Image
Yida Wang, David Joseph Tan, Nassir Navab, Federico Tombari; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8608-8
617
We propose a novel model for 3D semantic completion from a single depth image, b
ased on a single encoder and three separate generators used to reconstruct diffe
rent geometric and semantic representations of the original and completed scene,
 all sharing the same latent space. To transfer information between the geometri
c and semantic branches of the network, we introduce paths between them concaten
ating features at corresponding network layers. Motivated by the limited amount

of training samples from real scenes, an interesting attribute of our architectu re is the capacity to supplement the existing dataset by generating a new traini ng dataset with high quality, realistic scenes that even includes occlusion and real noise. We build the new dataset by sampling the features directly from late nt space which generates a pair of partial volumetric surface and completed volu metric semantic surface. Moreover, we utilize multiple discriminators to increas e the accuracy and realism of the reconstructions. We demonstrate the benefits o f our approach on standard benchmarks for the two most common completion tasks: semantic 3D scene completion and 3D object completion.
****

Moving Indoor: Unsupervised Video Depth Learning in Challenging Environments
Junsheng Zhou, Yuwang Wang, Kaihuai Qin, Wenjun Zeng; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2019, pp. 8618-8627
Recently unsupervised learning of depth from videos has made remarkable progress and the results are comparable to fully supervised methods in outdoor scenes li ke KITTI. However, there still exist great challenges when directly applying thi s technology in indoor environments, e.g., large areas of non-texture regions li ke white wall, more complex ego-motion of handheld camera, transparent glasses a nd shiny objects. To overcome these problems, we propose a new optical-flow base d training paradigm which reduces the difficulty of unsupervised learning by pro viding a clearer training target and handles the non-texture regions. Our experi mental evaluation demonstrates that the result of our method is comparable to fu lly supervised methods on the NYU Depth V2 benchmark. To the best of our knowled ge, this is the first quantitative result of purely unsupervised learning method reported on indoor datasets.
****

GraphX-Convolution for Point Cloud Deformation in 2D-to-3D Conversion
Anh-Duc Nguyen, Seonghwa Choi, Woojae Kim, Sanghoon Lee; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8628-8637
In this paper, we present a novel deep method to reconstruct a point cloud of an object from a single still image. Prior arts in the field struggle to reconstru ct an accurate and scalable 3D model due to either the inefficient and expensive 3D representations, the dependency between the output and number of model param eters or the lack of a suitable computing operation. We propose to overcome thes e by deforming a random point cloud to the object shape through two steps: featu re blending and deformation. In the first step, the global and point-specific sh ape features extracted from a 2D object image are blended with the encoded featu re of a randomly generated point cloud, and then this mixture is sent to the def ormation step to produce the final representative point set of the object. In th e deformation process, we introduce a new layer termed as GraphX that considers the inter-relationship between points like common graph convolutions but operate s on unordered sets. Moreover, with a simple trick, the proposed model can gener ate an arbitrary-sized point cloud, which is the first deep method to do so. Ext ensive experiments verify that we outperform existing models and halve the state -of-the-art distance score in single image 3D reconstruction.
****

Holistic++ Scene Understanding: Single-View 3D Holistic Scene Parsing and Human Pose Estimation With Human-Object Interaction and Physical Commonsense
Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, Song-Chun Zhu; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 19, pp. 8648-8657
We propose a new 3D holistic++ scene understanding problem, which jointly tackle s two tasks from a single-view image: (i) holistic scene parsing and reconstruct ion---3D estimations of object bounding boxes, camera pose, and room layout, and (ii) 3D human pose estimation. The intuition behind is to leverage the coupled nature of these two tasks to improve the granularity and performance of scene un derstanding. We propose to exploit two critical and essential connections betwee n these two tasks: (i) human-object interaction (HOI) to model the fine-grained relations between agents and objects in the scene, and (ii) physical commonsense to model the physical plausibility of the reconstructed scene. The optimal conf

iguration of the 3D scene, represented by a parse graph, is inferred using Markov chain Monte Carlo (MCMC), which efficiently traverses through the non-differentiable joint solution space. Experimental results demonstrate that the proposed algorithm significantly improves the performance of the two tasks on three datasets, showing an improved generalization ability.
****

MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding
Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, Tomokazu Murakami; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8658-8667
Unlike vision modalities, body-worn sensors or passive sensing can avoid the failure of action understanding in vision related challenges, e.g. occlusion and appearance variation. However, a standard large-scale dataset does not exist, in which different types of modalities across vision and sensors are integrated. To address the disadvantage of vision-based modalities and push towards multi/cross modal action understanding, this paper introduces a new large-scale dataset recorded from 20 distinct subjects with seven different types of modalities: RGB videos, keypoints, acceleration, gyroscope, orientation, Wi-Fi and pressure signal. The dataset consists of more than 36k video clips for 37 action classes covering a wide range of daily life activities such as desktop-related and check-in-based ones in four different distinct scenarios. On the basis of our dataset, we propose a novel multi modality distillation model with attention mechanism to realize an adaptive knowledge transfer from sensor-based modalities to vision-based modalities. The proposed model significantly improves performance of action recognition compared to models trained with only RGB information. The experimental results confirm the effectiveness of our model on cross-subject, -view, -scene and -session evaluation criteria. We believe that this new large-scale multimodal dataset will contribute the community of multimodal based action understanding.
****

HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization
Hang Zhao, Antonio Torralba, Lorenzo Torresani, Zhicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8668-8678
This paper presents a new large-scale dataset for recognition and temporal localization of human actions collected from Web videos. We refer to it as HACS (Human Action Clips and Segments). We leverage consensus and disagreement among visual classifiers to automatically mine candidate short clips from unlabeled videos, which are subsequently validated by human annotators. The resulting dataset is dubbed HACS Clips. Through a separate process we also collect annotations defining action segment boundaries. This resulting dataset is called HACS Segments. Overall, HACS Clips consists of 1.5M annotated clips sampled from 504K untrimmed videos, and HACS Segments contains 139K action segments densely annotated in 50K untrimmed videos spanning 200 action categories. HACS Clips contains more labeled examples than any existing video benchmark. This renders our dataset both a large-scale action recognition benchmark and an excellent source for spatiotemporal feature learning. In our transfer learning experiments on three target datasets, HACS Clips outperforms Kinetics-600, Moments-In-Time and Sports1M as a pretraining source. On HACS Segments, we evaluate state-of-the-art methods of action proposal generation and action localization, and highlight the new challenges posed by our dense temporal annotations.
****

3C-Net: Category Count and Center Loss for Weakly-Supervised Action Localization
Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8679-8687
Temporal action localization is a challenging computer vision problem with numerous real-world applications. Most existing methods require laborious frame-level supervision to train action localization models. In this work, we propose a framework, called 3C-Net, which only requires video-level supervision (weak supervi

sion) in the form of action category labels and the corresponding count. We intr
oduce a novel formulation to learn discriminative action features with enhanced
localization capabilities. Our joint formulation has three terms: a classificati
on term to ensure the separability of learned action features, an adapted multi-
label center loss term to enhance the action feature discriminability and a coun
ting loss term to delineate adjacent action sequences, leading to improved local
ization. Comprehensive experiments are performed on two challenging benchmarks:
THUMOS14 and ActivityNet 1.2. Our approach sets a new state-of-the-art for weakl
y-supervised temporal action localization on both datasets. On the THUMOS14 data
set, the proposed method achieves an absolute gain of 4.6% in terms of mean aver
age precision (mAP), compared to the state-of-the-art. Source code is available
at https://github.com/naraysa/3c-net.
****

Grounded Human-Object Interaction Hotspots From Video
Tushar Nagarajan, Christoph Feichtenhofer, Kristen Grauman; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8688-869
7
Learning how to interact with objects is an important step towards embodied visu
al intelligence, but existing techniques suffer from heavy supervision or sensin
g requirements. We propose an approach to learn human-object interaction "hotspo
ts" directly from video. Rather than treat affordances as a manually supervised
semantic segmentation task, our approach learns about interactions by watching v
ideos of real human behavior and anticipating afforded actions. Given a novel im
age or video, our model infers a spatial hotspot map indicating where an object
would be manipulated in a potential interaction even if the object is currently
at rest. Through results with both first and third person video, we show the val
ue of grounding affordances in real human-object interactions. Not only are our
weakly supervised hotspots competitive with strongly supervised affordance metho
ds, but they can also anticipate object interaction for novel object categories.
Project page: http://vision.cs.utexas.edu/projects/interaction-hotspots/
****

Hallucinating IDT Descriptors and I3D Optical Flow Features for Action Recogniti
on With CNNs
Lei Wang, Piotr Koniusz, Du Q. Huynh; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2019, pp. 8698-8708
In this paper, we revive the use of old-fashioned handcrafted video representati
ons for action recognition and put new life into these techniques via a CNN-base
d hallucination step. Despite of the use of RGB and optical flow frames, the I3D
model (amongst others) thrives on combining its output with the Improved Dense
Trajectory (IDT) and extracted with its low-level video descriptors encoded via
Bag-of-Words (BoW) and Fisher Vectors (FV). Such a fusion of CNNs and handcrafte
d representations is time-consuming due to pre-processing, descriptor extraction
, encoding and tuning parameters. Thus, we propose an end-to-end trainable netwo
rk with streams which learn the IDT-based BoW/FV representations at the training
stage and are simple to integrate with the I3D model. Specifically, each stream
takes I3D feature maps ahead of the last 1D conv. layer and learns to `translat
e' these maps to BoW/FV representations. Thus, our model can hallucinate and use
such synthesized BoW/FV representations at the testing stage. We show that even
features of the entire I3D optical flow stream can be hallucinated thus simplif
ying the pipeline. Our model saves 20-55h of computations and yields state-of-th
e-art results on four publicly available datasets.
****

Learning to Paint With Model-Based Deep Reinforcement Learning
Zhewei Huang, Wen Heng, Shuchang Zhou; Proceedings of the IEEE/CVF Internation
al Conference on Computer Vision (ICCV), 2019, pp. 8709-8718
We show how to teach machines to paint like human painters, who can use a small
number of strokes to create fantastic paintings. By employing a neural renderer
in model-based Deep Reinforcement Learning (DRL), our agents learn to determine
the position and color of each stroke and make long-term plans to decompose text
ure-rich images into strokes. Experiments demonstrate that excellent visual effe

cts can be achieved using hundreds of strokes. The training process does not req
uire the experience of human painters or stroke tracking data. The code is avail
able at https://github.com/hzwer/ICCV2019-LearningToPaint.
****

Neural Re-Simulation for Generating Bounces in Single Images
Carlo Innamorati, Bryan Russell, Danny M. Kaufman, Niloy J. Mitra; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp.
8719-8728
We introduce a method to generate videos of dynamic virtual objects plausibly in
teracting via collisions with a still image's environment. Given a starting traj
ectory, physically simulated with the estimated geometry of a single, static inp
ut image, we learn to 'correct' this trajectory to a visually plausible one via
a neural network. The neural network can then be seen as learning to 'correct' t
raditional simulation output, generated with incomplete and imprecise world info
rmation, to obtain context-specific, visually plausible re-simulated output - a
process we call neural re-simulation. We train our system on a set of 50k synthe
tic scenes where a virtual moving object (ball) has been physically simulated. W
e demonstrate our approach on both our synthetic dataset and a collection of rea
l-life images depicting everyday scenes, obtaining consistent improvement over b
aseline alternatives throughout.
****

Deep Appearance Maps
Maxim Maximov, Laura Leal-Taixe, Mario Fritz, Tobias Ritschel; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8729
-8738
We propose a deep representation of appearance, i.e. the relation of color, surf
ace orientation, viewer position, material and illumination. Previous approaches
have used deep learning to extract classic appearance representations relating
to reflectance model parameters (e.g. Phong) or illumination (e.g. HDR environme
nt maps). We suggest to directly represent appearance itself as a network we cal
l a deep appearance map (DAM). This is a 4D generalization over 2D reflectance m
aps, which held the view direction fixed. First, we show how a DAM can be learne
d from images or video frames and later be used to synthesize appearance, given
new surface orientations and viewer positions. Second, we demonstrate how anothe
r network can be used to map from an image or video frames to a DAM network to r
eproduce this appearance, without using a lengthy optimization such as stochasti
c gradient descent (learning-to-learn). Finally, we show the example of an appea
rance estimation-and-segmentation task, mapping from an image showing multiple m
aterials to multiple deep appearance maps.
****

GarNet: A Two-Stream Network for Fast and Accurate 3D Cloth Draping
Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu S
alzmann, Pascal Fua; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2019, pp. 8739-8748
While Physics-Based Simulation (PBS) can accurately drape a 3D garment on a 3D b
ody, it remains too costly for real-time applications, such as virtual try-on. B
y contrast, inference in a deep network, requiring a single forward pass, is muc
h faster. Taking advantage of this, we propose a novel architecture to fit a 3D
garment template to a 3D body. Specifically, we build upon the recent progress i
n 3D point cloud processing with deep networks to extract garment features at va
rying levels of detail, including point-wise, patch-wise and global features. We
fuse these features with those extracted in parallel from the 3D body, so as to
model the cloth-body interactions. The resulting two-stream architecture, which
we call as GarNet, is trained using a loss function inspired by physics-based m
odeling, and delivers visually plausible garment shapes whose 3D points are, on
average, less than 1 cm away from those of a PBS method, while running 100 times
faster. Moreover, the proposed method can model various garment types with diff
erent cutting patterns when parameters of those patterns are given as input to t
he network.
****

Joint Embedding of 3D Scan and CAD Objects
Manuel Dahnert, Angela Dai, Leonidas J. Guibas, Matthias Niessner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8749-8758
3D scan geometry and CAD models often contain complementary information towards understanding environments, which could be leveraged through establishing a mapping between the two domains. However, this is a challenging task due to strong, lower-level differences between scan and CAD geometry. We propose a novel approach to learn a joint embedding space between scan and CAD geometry, where semantically similar objects from both domains lie close together. To achieve this, we introduce a new 3D CNN-based approach to learn a joint embedding space representing object similarities across these domains. To learn a shared space where scan objects and CAD models can interlace, we propose a stacked hourglass approach to separate foreground and background from a scan object, and transform it to a complete, CAD-like representation to produce a shared embedding space. This embedding space can then be used for CAD model retrieval; to further enable this task, we introduce a new dataset of ranked scan-CAD similarity annotations, enabling new, fine-grained evaluation of CAD model retrieval to cluttered, noisy, partial scans. Our learned joint embedding outperforms current state of the art for CAD model retrieval by 12% in instance retrieval accuracy.
****
CompoNet: Learning to Generate the Unseen by Part Synthesis and Composition
Nadav Schor, Oren Katzir, Hao Zhang, Daniel Cohen-Or; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8759-8768
Data-driven generative modeling has made remarkable progress by leveraging the power of deep neural networks. A reoccurring challenge is how to enable a model to generate a rich variety of samples from the entire target distribution, rather than only from a distribution confined to the training data. In other words, we would like the generative model to go beyond the observed samples and learn to generate "unseen", yet still plausible, data. In our work, we present CompoNet, a generative neural network for 2D or 3D shapes that is based on a part-based prior, where the key idea is for the network to synthesize shapes by varying both the shape parts and their compositions. Treating a shape not as an unstructured whole, but as a (re-)composable set of deformable parts, adds a combinatorial dimension to the generative process to enrich the diversity of the output, encouraging the generator to venture more into the "unseen". We show that our part-based model generates richer variety of plausible shapes compared with baseline generative models. To this end, we introduce two quantitative metrics to evaluate the diversity of a generative model and assess how well the generated data covers both the training data and unseen data from the same target distribution.
****
DDSL: Deep Differentiable Simplex Layer for Learning Geometric Signals
Chiyu "Max" Jiang, Dana Lansigan, Philip Marcus, Matthias Niessner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8769-8778
We present a Deep Differentiable Simplex Layer (DDSL) for neural networks for geometric deep learning. The DDSL is a differentiable layer compatible with deep neural networks for bridging simplex mesh-based geometry representations (point clouds, line mesh, triangular mesh, tetrahedral mesh) with raster images (e.g., 2D/3D grids). The DDSL uses Non-Uniform Fourier Transform (NUFT) to perform differentiable, efficient, anti- aliased rasterization of simplex-based signals. We present a complete theoretical framework for the process as well as an efficient backpropagation algorithm. Compared to previous differentiable renderers and rasterizers, the DDSL generalizes to arbitrary simplex degrees and dimensions. In particular, we explore its applications to 2D shapes and illustrate two applications of this method: (1) mesh editing and optimization guided by neural network outputs, and (2) using DDSL for a differentiable rasterization loss to facilitate end-to-end training of polygon generators. We are able to validate the effectiveness of gradient-based shape optimization with the example of airfoil optimization, and using the differentiable rasterization loss to facilitate end-to-end tr

aining, we surpass state of the art for polygonal image segmentation given groun d-truth bounding boxes.
****

EGNet: Edge Guidance Network for Salient Object Detection
Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, Ming-Ming Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2019, pp. 8779-8788
Fully convolutional neural networks (FCNs) have shown their advantages in the sa lient object detection task. However, most existing FCNs-based methods still suf fer from coarse object boundaries. In this paper, to solve this problem, we focu s on the complementarity between salient edge information and salient object inf ormation. Accordingly, we present an edge guidance network (EGNet) for salient o bject detection with three steps to simultaneously model these two kinds of comp lementary information in a single network. In the first step, we extract the sal ient object features by a progressive fusion way. In the second step, we integra te the local edge information and global location information to obtain the sali ent edge features. Finally, to sufficiently leverage these complementary feature s, we couple the same salient edge features with salient object features at vari ous resolutions. Benefiting from the rich edge information and location informat ion in salient edge features, the fused features can help locate salient objects , especially their boundaries more accurately. Experimental results demonstrate that the proposed method performs favorably against the state-of-the-art methods on six widely used datasets without any pre-processing and post-processing. The source code is available at http: //mmcheng.net/egnet/.
****

SID4VAM: A Benchmark Dataset With Synthetic Images for Visual Attention Modeling
David Berga, Xose R. Fdez-Vidal, Xavier Otazu, Xose M. Pardo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8789-8798
A benchmark of saliency models performance with a synthetic image dataset is pro vided. Model performance is evaluated through saliency metrics as well as the in fluence of model inspiration and consistency with human psychophysics. SID4VAM i s composed of 230 synthetic images, with known salient regions. Images were gene rated with 15 distinct types of low-level features (e.g. orientation, brightness , color, size...) with a target-distractor pop-out type of synthetic patterns. W e have used Free-Viewing and Visual Search task instructions and 7 feature contr asts for each feature category. Our study reveals that state-of-the-art Deep Lea rning saliency models do not perform well with synthetic pattern images, instead , models with Spectral/Fourier inspiration outperform others in saliency metrics and are more consistent with human psychophysical experimentation. This study p roposes a new way to evaluate saliency models in the forthcoming literature, acc ounting for synthetic images with uniquely low-level feature contexts, distinct from previous eye tracking image datasets.
****

Two-Stream Action Recognition-Oriented Video Super-Resolution
Haochen Zhang, Dong Liu, Zhiwei Xiong; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2019, pp. 8799-8808
We study the video super-resolution (SR) problem for facilitating video analytic s tasks, e.g. action recognition, instead of for visual quality. The popular act ion recognition methods based on convolutional networks, exemplified by two-stre am networks, are not directly applicable on video of low spatial resolution. Thi s can be remedied by performing video SR prior to recognition, which motivates u s to improve the SR procedure for recognition accuracy. Tailored for two-stream action recognition networks, we propose two video SR methods for the spatial and temporal streams respectively. On the one hand, we observe that regions with ac tion are more important to recognition, and we propose an optical-flow guided we ighted mean-squared-error loss for our spatial-oriented SR (SoSR) network to emp hasize the reconstruction of moving objects. On the other hand, we observe that existing video SR methods incur temporal discontinuity between frames, which als o worsens the recognition accuracy, and we propose a siamese network for our tem

poral-oriented SR (ToSR) training that emphasizes the temporal continuity betwee n consecutive frames. We perform experiments using two state-of-the-art action r ecognition networks and two well-known datasets--UCF101 and HMDB51. Results demo nstrate the effectiveness of our proposed SoSR and ToSR in improving recognition accuracy.

****

Where Is My Mirror?

Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, Rynson W.H. Lau; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 19, pp. 8809-8818

Mirrors are everywhere in our daily lives. Existing computer vision systems do n ot consider mirrors, and hence may get confused by the reflected content inside a mirror, resulting in a severe performance degradation. However, separating the real content outside a mirror from the reflected content inside it is non-trivi al. The key challenge is that mirrors typically reflect contents similar to thei r surroundings, making it very difficult to differentiate the two. In this paper , we present a novel method to segment mirrors from an input image. To the best of our knowledge, this is the first work to address the mirror segmentation prob lem with a computational approach. We make the following contributions. First, w e construct a large-scale mirror dataset that contains mirror images with corres ponding manually annotated masks. This dataset covers a variety of daily life sc enes, and will be made publicly available for future research. Second, we propos e a novel network, called MirrorNet, for mirror segmentation, by modeling both s emantical and low-level color/texture discontinuities between the contents insid e and outside of the mirrors. Third, we conduct extensive experiments to evaluat e the proposed method, and show that it outperforms the carefully chosen baselin es from the state-of-the-art detection and segmentation methods.

****

Disentangled Image Matting

Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiam ing Liu, Jiaying Liu, Jue Wang, Jian Sun; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2019, pp. 8819-8828

Most previous image matting methods require a roughly-specificed trimap as input , and estimate fractional alpha values for all pixels that are in the unknown re gion of the trimap. In this paper, we argue that directly estimating the alpha m atte from a coarse trimap is a major limitation of previous methods, as this pra ctice tries to address two difficult and inherently different problems at the sa me time: identifying true blending pixels inside the trimap region, and estimate accurate alpha values for them. We propose AdaMatting, a new end-to-end matting framework that disentangles this problem into two sub-tasks: trimap adaptation and alpha estimation. Trimap adaptation is a pixel-wise classification problem t hat infers the global structure of the input image by identifying definite foreg round, background, and semi-transparent image regions. Alpha estimation is a reg ression problem that calculates the opacity value of each blended pixel. Our met hod separately handles these two sub-tasks within a single deep convolutional ne ural network (CNN). Extensive experiments show that AdaMatting has additional st ructure awareness and trimap fault-tolerance. Our method achieves the state-of-t he-art performance on Adobe Composition-1k dataset both qualitatively and quanti tatively. It is also the current best-performing method on the alphamatting.com online evaluation for all commonly-used metrics.

****

Guided Super-Resolution As Pixel-to-Pixel Transformation

Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, Konrad Schindler; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8829-8837

Guided super-resolution is a unifying framework for several computer vision task s where the inputs are a low-resolution source image of some target quantity (e. g., perspective depth acquired with a time-of-flight camera) and a high-resoluti on guide image from a different domain (e.g., a grey-scale image from a conventi onal camera); and the target output is a high-resolution version of the source (

in our example, a high-res depth map). The standard way of looking at this probl em is to formulate it as a super-resolution task, i.e., the source image is upsa mpled to the target resolution, while transferring the missing high-frequency de tails from the guide. Here, we propose to turn that interpretation on its head a nd instead see it as a pixel-to-pixel mapping of the guide image to the domain o f the source image. The pixel-wise mapping is parametrised as a multi-layer perc eptron, whose weights are learned by minimising the discrepancies between the so urce image and the downsampled target image. Importantly, our formulation makes it possible to regularise only the mapping function, while avoiding regularisati on of the outputs; thus producing crisp, natural-looking images. The proposed me thod is unsupervised, using only the specific source and guide images to fit the mapping. We evaluate our method on two different tasks, super-resolution of dep th maps and of tree height maps. In both cases, we clearly outperform recent bas elines in quantitative comparisons, while delivering visually much sharper outpu ts.

****

Deep Learning for Light Field Saliency Detection

Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, Huchuan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8838-8848

Recent research in 4D saliency detection is limited by the deficiency of a large -scale 4D light field dataset. To address this, we introduce a new dataset to as sist the subsequent research in 4D light field saliency detection. To the best o f our knowledge, this is to date the largest light field dataset in which the da taset provides 1465 all-focus images with human-labeled ground truth masks and t he corresponding focal stacks for every light field image. To verify the effecti veness of the light field data, we first introduce a fusion framework which incl udes two CNN streams where the focal stacks and all-focus images serve as the in put. The focal stack stream utilizes a recurrent attention mechanism to adaptive ly learn to integrate every slice in the focal stack, which benefits from the ex tracted features of the good slices. Then it is incorporated with the output map generated by the all-focus stream to make the saliency prediction. In addition, we introduce adversarial examples by adding noise intentionally into images to help train the deep network, which can improve the robustness of the proposed ne twork. The noise is designed by users, which is imperceptible but can fool the C NNs to make the wrong prediction. Extensive experiments show the effectiveness a nd superiority of the proposed model on the popular evaluation metrics. The prop osed method performs favorably compared with the existing 2D, 3D and 4D saliency detection methods on the proposed dataset and existing LFSD light field dataset . The code and results can be found at https://github.com/OIPLab-DUT/ ICCV2019_D eeplightfield_Saliency. Moreover, to facilitate research in this field, all imag es we collected are shared in a ready-to-use manner.

****

Optimizing the F-Measure for Threshold-Free Salient Object Detection

Kai Zhao, Shanghua Gao, Wenguan Wang, Ming-Ming Cheng; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8849-8857

Current CNN-based solutions to salient object detection (SOD) mainly rely on the optimization of cross-entropy loss (CELoss). Then the quality of detected salie ncy maps is often evaluated in terms of F-measure. In this paper, we investigate an interesting issue: can we consistently use the F-measure formulation in both training and evaluation for SOD? By reformulating the standard F-measure we pro pose the relaxed F-measure which is differentiable w.r.t the posterior and can b e easily appended to the back of CNNs as the loss function. Compared to the conv entional cross-entropy loss of which the gradients decrease dramatically in the saturated area, our loss function, named FLoss, holds considerable gradients eve n when the activation approaches the target. Consequently, the FLoss can continu ously force the network to produce polarized activations. Comprehensive benchmar ks on several popular datasets show that FLoss outperforms the state-of-the-art with a considerable margin. More specifically, due to the polarized predictions, our method is able to obtain high-quality saliency maps without carefully tunin

g the optimal threshold, showing significant advantages in real-world applicatio
ns.
****

Image Inpainting With Learnable Bidirectional Attention Maps
Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu,
Shilei Wen, Errui Ding; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2019, pp. 8858-8867
Most convolutional network (CNN)-based inpainting methods adopt standard convolu
tion to indistinguishably treat valid pixels and holes, making them limited in h
andling irregular holes and more likely to generate inpainting results with colo
r discrepancy and blurriness. Partial convolution has been suggested to address
this issue, but it adopts handcrafted feature re-normalization, and only conside
rs forward mask-updating. In this paper, we present a learnable attention map mo
dule for learning feature re-normalization and mask-updating in an end-to-end ma
nner, which is effective in adapting to irregular holes and propagation of convo
lution layers. Furthermore, learnable reverse attention maps are introduced to a
llow the decoder of U-Net to concentrate on filling in irregular holes instead o
f reconstructing both holes and known regions, resulting in our learnable bidire
ctional attention maps. Qualitative and quantitative experiments show that our m
ethod performs favorably against state-of-the-arts in generating sharper, more c
oherent and visually plausible inpainting results. The source code and pre-train
ed models will be available at: https://github.com/Vious/LBAM_inpainting/.
****

Joint Demosaicking and Denoising by Fine-Tuning of Bursts of Raw Images
Thibaud Ehret, Axel Davy, Pablo Arias, Gabriele Facciolo; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8868-8877
Demosaicking and denoising are the first steps of any camera image processing pi
peline and are key for obtaining high quality RGB images. A promising current re
search trend aims at solving these two problems jointly using convolutional neur
al networks. Due to the unavailability of ground truth data these networks canno
t be currently trained using real RAW images. Instead, they resort to simulated
data. In this paper we present a method to learn demosaicking directly from mosa
icked images, without requiring ground truth RGB data. We apply this to learn jo
int demosaicking and denoising only from RAW images, thus enabling the use of re
al data. In addition we show that for this application fine-tuning a network to
a specific burst improves the quality of restoration for both demosaicking and d
enoising.
****

DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better
Orest Kupyn, Tetiana Martyniuk, Junru Wu, Zhangyang Wang; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8878-8887
We present a new end-to-end generative adversarial network (GAN) for single imag
e motion deblurring, named DeblurGAN-V2, which considerably boosts state-of-the-
art deblurring performance while being much more flexible and efficient. DeblurG
AN-V2 is based on a relativistic conditional GAN with a double-scale discriminat
or. For the first time, we introduce the Feature Pyramid Network into deblurring
, as a core building block in the generator of DeblurGAN-V2. It can flexibly wor
k with a wide range of backbones, to navigate the balance between performance an
d efficiency. The plug-in of sophisticated backbones (e.g. Inception ResNet v2)
can lead to solid state-of-the-art performance. Meanwhile, with light-weight bac
kbones (e.g. MobileNet and its variants), DeblurGAN-V2 becomes 10-100 times fast
er than the nearest competitors, while maintaining close to state-of-the-art res
ults, implying the option of real-time video deblurring. We demonstrate that Deb
lurGAN-V2 has very competitive performance on several popular benchmarks, in ter
ms of deblurring quality (both objective and subjective), as well as efficiency.
In addition, we show the architecture to be effective for general image restora
tion tasks too. Our models and codes will be made available upon acceptance.
****

Reflective Decoding Network for Image Captioning
Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, Yu-Wing Tai; Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8888-8897

State-of-the-art image captioning methods mostly focus on improving visual features, less attention has been paid to utilizing the inherent properties of language to boost captioning performance. In this paper, we show that vocabulary coherence between words and syntactic paradigm of sentences are also important to generate high-quality image caption. Following the conventional encoder-decoder framework, we propose the Reflective Decoding Network (RDN) for image captioning, which enhances both the long-sequence dependency and position perception of words in a caption decoder. Our model learns to collaboratively attend on both visual and textual features and meanwhile perceive each word's relative position in the sentence to maximize the information delivered in the generated caption. We evaluate the effectiveness of our RDN on the COCO image captioning datasets and achieve superior performance over the previous methods. Further experiments reveal that our approach is particularly advantageous for hard cases with complex scenes to describe by captions.

****

Joint Optimization for Cooperative Image Captioning

Gilad Vered, Gal Oren, Yuval Atzmon, Gal Chechik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8898-8907

When describing images with natural language, descriptions can be made more informative if tuned for downstream tasks. This can be achieved by training two networks: a "speaker" that generates sentences given an image and a "listener" that uses them to perform a task. Unfortunately, training multiple networks jointly to communicate, faces two major challenges. First, the descriptions generated by a speaker network are discrete and stochastic, making optimization very hard and inefficient. Second, joint training usually causes the vocabulary used during communication to drift and diverge from natural language. To address these challenges, we present an effective optimization technique based on partial-sampling from a multinomial distribution combined with straight-through gradient updates, which we name PSST for Partial-Sampling Straight-Through. We then show that the generated descriptions can be kept close to natural by constraining them to be similar to human descriptions. Together, this approach creates descriptions that are both more discriminative and more natural than previous approaches. Evaluations on the COCO benchmark show that PSST improve the recall@10 from 60% to 86% maintaining comparable language naturalness. Human evaluations show that it also increases naturalness while keeping the discriminative power of generated captions.

****

Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning

Tanzila Rahman, Bicheng Xu, Leonid Sigal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8908-8917

Multi-modal learning, particularly among imaging and linguistic modalities, has made amazing strides in many high-level fundamental visual understanding problems, ranging from language grounding to dense event captioning. However, much of the research has been limited to approaches that either do not take audio corresponding to video into account at all, or those that model the audio-visual correlations in service of sound or sound source localization. In this paper, we present the evidence, that audio signals can carry surprising amount of information when it comes to high-level visual-lingual tasks. Specifically, we focus on the problem of weakly-supervised dense event captioning in videos and show that audio on its own can nearly rival performance of a state-of-the-art visual model and, combined with video, can improve on the state-of-the-art performance. Extensive experiments on the ActivityNet Captions dataset show that our proposed multi-modal approach outperforms state-of-the-art unimodal methods, as well as validate specific feature representation and architecture design choices.

****

Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning

Jingyi Hou, Xinxiao Wu, Wentian Zhao, Jiebo Luo, Yunde Jia; Proceedings of t

he IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8918-8927
Video captioning is a challenging task that involves not only visual perception but also syntax representation learning. Recent progress in video captioning has been achieved through visual perception, but syntax representation learning is still under-explored. We propose a novel video captioning approach that takes into account both visual perception and syntax representation learning to generate accurate descriptions of videos. Specifically, we use sentence templates composed of Part-of-Speech (POS) tags to represent the syntax structure of captions, and accordingly, syntax representation learning is performed by directly inferring POS tags from videos. The visual perception is implemented by a mixture model which translates visual cues into lexical words that are conditional on the learned syntactic structure of sentences. Thus, a video captioning task consists of two sub-tasks: video POS tagging and visual cue translation, which are jointly modeled and trained in an end-to-end fashion. Evaluations on three public benchmark datasets demonstrate that our proposed method achieves substantially better performance than the state-of-the-art methods, which validates the superiority of joint modeling of syntax representation learning and visual perception for video captioning.
****
Entangled Transformer for Image Captioning
Guang Li, Linchao Zhu, Ping Liu, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8928-8937
In image captioning, the typical attention mechanisms are arduous to identify the equivalent visual signals especially when predicting highly abstract words. This phenomenon is known as the semantic gap between vision and language. This problem can be overcome by providing semantic attributes that are homologous to language. Thanks to the inherent recurrent nature and gated operating mechanism, Recurrent Neural Network (RNN) and its variants are the dominating architectures in image captioning. However, when designing elaborate attention mechanisms to integrate visual inputs and semantic attributes, RNN-like variants become unflexible due to their complexities. In this paper, we investigate a Transformer-based sequence modeling framework, built only with attention layers and feedforward layers. To bridge the semantic gap, we introduce EnTangled Attention (ETA) that enables the Transformer to exploit semantic and visual information simultaneously. Furthermore, Gated Bilateral Controller (GBC) is proposed to guide the interactions between the multimodal information. We name our model as ETA-Transformer. Remarkably, ETA-Transformer achieves state-of-the-art performance on the MSCOCO image captioning dataset. The ablation studies validate the improvements of our proposed modules.
****
Shapeglot: Learning Language for Shape Differentiation
Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8938-8947
In this work we explore how fine-grained differences between the shapes of common objects are expressed in language, grounded on 2D and/or 3D object representations. We first build a large scale, carefully controlled dataset of human utterances each of which refers to a 2D rendering of a 3D CAD model so as to distinguish it from a set of shape-wise similar alternatives. Using this dataset, we develop neural language understanding (listening) and production (speaking) models that vary in their grounding (pure 3D forms via point-clouds vs. rendered 2D images), the degree of pragmatic reasoning captured (e.g. speakers that reason about a listener or not), and the neural architecture (e.g. with or without attention). We find models that perform well with both synthetic and human partners, and with held out utterances and objects. We also find that these models are capable of zero-shot transfer learning to novel object classes (e.g. transfer from training on chairs to testing on lamps), as well as to real-world images drawn from furniture catalogs. Lesion studies indicate that the neural listeners depend heavily on part-related words and associate these words correctly with visual parts

of objects (without any explicit supervision on such parts), and that transfer to novel classes is most successful when known part-related words are available. This work illustrates a practical approach to language grounding, and provides a novel case study in the relationship between object shape and linguistic structure when it comes to object differentiation.
****

nocaps: novel object captioning at scale
Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, Peter Anderson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8948-8957
Image captioning models have achieved impressive results on datasets containing limited visual concepts and large amounts of paired image-caption training data. However, if these models are to ever function in the wild, a much larger variety of visual concepts must be learned, ideally from less supervision. To encourage the development of image captioning models that can learn visual concepts from alternative data sources, such as object detection datasets, we present the first large-scale benchmark for this task. Dubbed 'nocaps', for novel object captioning at scale, our benchmark consists of 166,100 human-generated captions describing 15,100 images from the Open Images validation and test sets. The associated training data consists of COCO image-caption pairs, plus Open Images image-level labels and object bounding boxes. Since Open Images contains many more classes than COCO, nearly 400 object classes seen in test images have no or very few associated training captions (hence, nocaps). We extend existing novel object captioning models to establish strong baselines for this benchmark and provide analysis to guide future work.
****

Fully Convolutional Geometric Features
Christopher Choy, Jaesik Park, Vladlen Koltun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8958-8966
Extracting geometric features from 3D scans or point clouds is the first step in applications such as registration, reconstruction, and tracking. State-of-the-art methods require computing low-level features as input or extracting patch-based features with limited receptive field. In this work, we present fully-convolutional geometric features, computed in a single pass by a 3D fully-convolutional network. We also present new metric learning losses that dramatically improve performance. Fully-convolutional geometric features are compact, capture broad spatial context, and scale to large scenes. We experimentally validate our approach on both indoor and outdoor datasets. Fully-convolutional geometric features achieve state-of-the-art accuracy without requiring prepossessing, are compact (32 dimensions), and are 290 times faster than the most accurate prior method.
****

Learning Local RGB-to-CAD Correspondences for Object Pose Estimation
Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, Jana Kosecka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8967-8976
We consider the problem of 3D object pose estimation. While much recent work has focused on the RGB domain, the reliance on accurately annotated images limits generalizability and scalability. On the other hand, the easily available object CAD models are rich sources of data, providing a large number of synthetically rendered images. In this paper, we solve this key problem of existing methods requiring expensive 3D pose annotations by proposing a new method that matches RGB images to CAD models for object pose estimation. Our key innovations compared to existing work include removing the need for either real-world textures for CAD models or explicit 3D pose annotations for RGB images. We achieve this through a series of objectives that learn how to select keypoints and enforce viewpoint and modality invariance across RGB images and CAD model renderings. Our experiments demonstrate that the proposed method can reliably estimate object pose in RGB images and generalize to object instances not seen during training.
****

Depth From Videos in the Wild: Unsupervised Monocular Depth Learning From Unknown Cameras
Ariel Gordon, Hanhan Li, Rico Jonschkowski, Anelia Angelova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8977-8986
We present a novel method for simultaneous learning of depth, egomotion, object motion, and camera intrinsics from monocular videos, using only consistency across neighboring video frames as supervision signal. Similarly to prior work, our method learns by applying differentiable warping to frames and comparing the result to adjacent ones, but it provides several improvements: We address occlusions geometrically and differentiably, directly using the depth maps as predicted during training. We introduce randomized layer normalization, a novel powerful regularizer, and we account for object motion relative to the scene. To the best of our knowledge, our work is the first to learn the camera intrinsic parameters, including lens distortion, from video in an unsupervised manner, thereby allowing us to extract accurate depth and motion from arbitrary videos of unknown origin at scale. We evaluate our results on the Cityscapes, KITTI and EuRoC datasets, establishing new state of the art on depth prediction and odometry, and demonstrate qualitatively that depth prediction can be learned from a collection of YouTube videos. The code will be open sourced once anonymity is lifted.
****
OmniMVS: End-to-End Learning for Omnidirectional Stereo Matching
Changhee Won, Jongbin Ryu, Jongwoo Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8987-8996
In this paper, we propose a novel end-to-end deep neural network model for omnidirectional depth estimation from a wide-baseline multi-view stereo setup. The images captured with ultra wide field-of-view (FOV) cameras on an omnidirectional rig are processed by the feature extraction module, and then the deep feature maps are warped onto the concentric spheres swept through all candidate depths using the calibrated camera parameters. The 3D encoder-decoder block takes the aligned feature volume to produce the omnidirectional depth estimate with regularization on uncertain regions utilizing the global context information. In addition, we present large-scale synthetic datasets for training and testing omnidirectional multi-view stereo algorithms. Our datasets consist of 11K ground-truth depth maps and 45K fisheye images in four orthogonal directions with various objects and environments. Experimental results show that the proposed method generates excellent results in both synthetic and real-world environments, and it outperforms the prior art and the omnidirectional versions of the state-of-the-art conventional stereo algorithms.
****
On the Over-Smoothing Problem of CNN Based Disparity Estimation
Chuangrong Chen, Xiaozhi Chen, Hui Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8997-9005
Currently, most deep learning based disparity estimation methods have the problem of over-smoothing at boundaries, which is unfavorable for some applications such as point cloud segmentation, mapping, etc. To address this problem, we first analyze the potential causes and observe that the estimated disparity at edge boundary pixels usually follows multimodal distributions, causing over-smoothing estimation. Based on this observation, we propose a single-modal weighted average operation on the probability distribution during inference, which can alleviate the problem effectively. To integrate the constraint of this inference method into training stage, we further analyze the characteristics of different loss functions and found that using cross entropy with gaussian distribution consistently further improves the performance. For quantitative evaluation, we propose a novel metric that measures the disparity error in the local structure of edge boundaries. Experiments on various datasets using various networks show our method's effectiveness and general applicability. Code will be available at https://github.com/chenchr/otosp.
****
Disentangling Propagation and Generation for Video Prediction

Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, Trevor Darrell; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9006-9015

A dynamic scene has two types of elements: those that move fluidly and can be predicted from previous frames, and those which are disoccluded (exposed) and cannot be extrapolated. Prior approaches to video prediction typically learn either to warp or to hallucinate future pixels, but not both. In this paper, we describe a computational model for high-fidelity video prediction which disentangles motion-specific propagation from motion-agnostic generation. We introduce a confidence-aware warping operator which gates the output of pixel predictions from a flow predictor for non-occluded regions and from a context encoder for occluded regions. Moreover, in contrast to prior works where confidence is jointly learned with flow and appearance using a single network, we compute confidence after a warping step, and employ a separate network to inpaint exposed regions. Empirical results on both synthetic and real datasets show that our disentangling approach provides better occlusion maps and produces both sharper and more realistic predictions compared to strong baselines.

****

Guided Image-to-Image Translation With Bi-Directional Feature Transformation

Badour AlBahar, Jia-Bin Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9016-9025

We address the problem of guided image-to-image translation where we translate an input image into another while respecting the constraints provided by an external, user-provided guidance image. Various types of conditioning mechanisms for leveraging the given guidance image have been explored, including input concatenation, feature concatenation, and conditional affine transformation of feature activations. All these conditioning mechanisms, however, are uni-directional, i.e., no information flow from the input image back to the guidance. To better utilize the constraints of the guidance image, we present a bi-directional feature transformation (bFT) scheme. We show that our novel bFT scheme outperforms other conditioning schemes and has comparable results to state-of-the-art methods on different tasks.

****

Towards Multi-Pose Guided Virtual Try-On Network

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, Jian Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9026-9035

Virtual try-on systems under arbitrary human poses have significant application potential, yet also raise extensive challenges, such as self-occlusions, heavy misalignment among different poses, and complex clothes textures. Existing virtual try-on methods can only transfer clothes given a fixed human pose, and still show unsatisfactory performances, often failing to preserve person identity or texture details, and with limited pose diversity. This paper makes the first attempt towards a multi-pose guided virtual try-on system, which enables clothes to transfer onto a person with diverse poses. Given an input person image, a desired clothes image, and a desired pose, the proposed Multi-pose Guided Virtual Try-On Network (MG-VTON) generates a new person image after fitting the desired clothes into the person and manipulating the pose. MG-VTON is constructed with three stages: 1) a conditional human parsing network is proposed that matches both the desired pose and the desired clothes shape; 2) a deep Warping Generative Adversarial Network (Warp-GAN) that warps the desired clothes appearance into the synthesized human parsing map and alleviates the misalignment problem between the input human pose and the desired one; 3) a refinement render network recovers the texture details of clothes and removes artifacts, based on multi-pose composition masks. Extensive experiments on commonly-used datasets and our newly-collected largest virtual try-on benchmark demonstrate that our MG-VTON significantly outperforms all state-of-the-art methods both qualitatively and quantitatively, showing promising virtual try-on performances.

****

Photorealistic Style Transfer via Wavelet Transforms

Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, Jung-Woo Ha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9036-9045

Recent style transfer models have provided promising artistic results. However, given a photograph as a reference style, existing methods are limited by spatial distortions or unrealistic artifacts, which should not happen in real photographs. We introduce a theoretically sound correction to the network architecture that remarkably enhances photorealism and faithfully transfers the style. The key ingredient of our method is wavelet transforms that naturally fits in deep networks. We propose a wavelet corrected transfer based on whitening and coloring transforms (WCT2) that allows features to preserve their structural information and statistical properties of VGG feature space during stylization. This is the first and the only end-to-end model that can stylize a 1024x1024 resolution image in 4.7 seconds, giving a pleasing and photorealistic quality without any post-processing. Last but not least, our model provides a stable video stylization without temporal constraints. Our code, generated images, pre-trained models and supplementary documents are all available at https://github.com/ClovaAI/WCT2.

****

Personalized Fashion Design

Cong Yu, Yang Hu, Yan Chen, Bing Zeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9046-9055

Fashion recommendation is the task of suggesting a fashion item that fits well with a given item. In this work, we propose to automatically synthesis new items for recommendation. We jointly consider the two key issues for the task, i.e., compatibility and personalization. We propose a personalized fashion design framework with the help of generative adversarial training. A convolutional network is first used to map the query image into a latent vector representation. This latent representation, together with another vector which characterizes user's style preference, are taken as the input to the generator network to generate the target item image. Two discriminator networks are built to guide the generation process. One is the classic real/fake discriminator. The other is a matching network which simultaneously models the compatibility between fashion items and learns users' preference representations. The performance of the proposed method is evaluated on thousands of outfits composited by online users. The experiments show that the items generated by our model are quite realistic. They have better visual quality and higher matching degree than those generated by alternative methods.

****

Tag2Pix: Line Art Colorization Using Text Tag With SECat and Changing Loss

Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, Sungjoo Yoo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9056-9065

Line art colorization is expensive and challenging to automate. A GAN approach is proposed, called Tag2Pix, of line art colorization which takes as input a gray scale line art and color tag information and produces a quality colored image. First, we present the Tag2Pix line art colorization dataset. A generator network is proposed which consists of convolutional layers to transform the input line art, a pre-trained semantic extraction network, and an encoder for input color information. The discriminator is based on an auxiliary classifier GAN to classify the tag information as well as genuineness. In addition, we propose a novel network structure called SECat, which makes the generator properly colorize even small features such as eyes, and also suggest a novel two-step training method where the generator and discriminator first learn the notion of object and shape and then, based on the learned notion, learn colorization, such as where and how to place which color. We present both quantitative and qualitative evaluations which prove the effectiveness of the proposed method.

****

Free-Form Video Inpainting With 3D Gated Convolution and Temporal PatchGAN

Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, Winston Hsu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9066-9075

Free-form video inpainting is a very challenging task that could be widely used

for video editing such as text removal. Existing patch-based methods could not h
andle non-repetitive structures such as faces, while directly applying image-bas
ed inpainting models to videos will result in temporal inconsistency (see this h
ttps://www.youtube.com/watch?v=BuTYfo4bO2I&list=PLnEeMdoBDCISRm0EZYFcQuaJ5ITUaaE
Ib&index=1). In this paper, we introduce a deep learning based free-form video i
npainting model, with proposed 3D gated convolutions to tackle the uncertainty o
f free-form masks and a novel Temporal PatchGAN loss to enhance temporal consist
ency. In addition, we collect videos and design a free-form mask generation algo
rithm to build the free-form video inpainting (FVI) dataset for training and eva
luation of video inpainting models. We demonstrate the benefits of these compone
nts and experiments on both the FaceForensics and our FVI dataset suggest that o
ur method is superior to existing ones. Related source code, full-resolution res
ult videos and the FVI dataset could be found on Github: https://github.com/amjl
tc295/Free-Form-Video-Inpainting
****

TextDragon: An End-to-End Framework for Arbitrary Shaped Text Spotting
Wei Feng, Wenhao He, Fei Yin, Xu-Yao Zhang, Cheng-Lin Liu; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9076-90
85
Most existing text spotting methods either focus on horizontal/oriented texts or
 perform arbitrary shaped text spotting with character-level annotations. In thi
s paper, we propose a novel text spotting framework to detect and recognize text
 of arbitrary shapes in an end-to-end manner, using only word/line-level annotat
ions for training. Motivated from the name of TextSnake, which is only a detecti
on model, we call the proposed text spotting framework TextDragon. In TextDragon
, a text detector is designed to describe the shape of text with a series of qua
drangles, which can handle text of arbitrary shapes. To extract arbitrary text r
egions from feature maps, we propose a new differentiable operator named RoISlid
e, which is the key to connect arbitrary shaped text detection and recognition.
Based on the extracted features through RoISlide, a CNN and CTC based text recog
nizer is introduced to make the framework free from labeling the location of cha
racters. The proposed method achieves state-of-the-art performance on two curved
 text benchmarks CTW1500 and Total-Text, and competitive results on the ICDAR 20
15 Dataset.
****

Chinese Street View Text: Large-Scale Chinese Text Reading With Partially Superv
ised Learning
Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, Jingtuo Liu; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019,
 pp. 9086-9095
Most existing text reading benchmarks make it difficult to evaluate the performa
nce of more advanced deep learning models in large vocabularies due to the limit
ed amount of training data. To address this issue, we introduce a new large-scal
e text reading benchmark dataset named Chinese Street View Text (C-SVT) with 430
,000 street view images, which is at least 14 times as large as the existing Chi
nese text reading benchmarks. To recognize Chinese text in the wild while keepin
g large-scale datasets labeling cost-effective, we propose to annotate one part
of the C-SVT dataset (30,000 images) in locations and text labels as full annota
tions and add 400,000 more images, where only the corresponding text-of-interest
 in the regions is given as weak annotations. To exploit the rich information fr
om the weakly annotated data, we design a text reading network in a partially su
pervised learning framework, which enables to localize and recognize text, learn
 from fully and weakly annotated data simultaneously. To localize the best match
ed text proposals from weakly labeled images, we propose an online proposal matc
hing module incorporated in the whole model, spotting the keyword regions by sha
ring parameters for end-to-end training. Compared with fully supervised training
 algorithms, this model can improve the end-to-end recognition performance remar
kably by 4.03% in F-score at the same labeling cost. The proposed model can also
 achieve state-of-the-art results on the ICDAR 2017-RCTW dataset, which demonstr
ates the effectiveness of the proposed partially supervised learning framework.

****
Deep Floor Plan Recognition Using a Multi-Task Network With Room-Boundary-Guided Attention
Zhiliang Zeng, Xianzhi Li, Ying Kin Yu, Chi-Wing Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9096-9104
This paper presents a new approach to recognize elements in floor plan layouts. Besides walls and rooms, we aim to recognize diverse floor plan elements, such as doors, windows and different types of rooms, in the floor layouts. To this end, we model a hierarchy of floor plan elements and design a deep multi-task neural network with two tasks: one to learn to predict room-boundary elements, and the other to predict rooms with types. More importantly, we formulate the room-boundary-guided attention mechanism in our spatial contextual module to carefully take room-boundary features into account to enhance the room-type predictions. Furthermore, we design a cross-and-within-task weighted loss to balance the multi-label tasks and prepare two new datasets for floor plan recognition. Experimental results demonstrate the superiority and effectiveness of our network over the state-of-the-art methods.
****
GA-DAN: Geometry-Aware Domain Adaptation Network for Scene Text Detection and Recognition
Fangneng Zhan, Chuhui Xue, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9105-9115
Recent adversarial learning research has achieved very impressive progress for modelling cross-domain data shifts in appearance space but its counterpart in modelling cross-domain shifts in geometry space lags far behind. This paper presents an innovative Geometry-Aware Domain Adaptation Network (GA-DAN) that is capable of modelling cross-domain shifts concurrently in both geometry space and appearance space and realistically converting images across domains with very different characteristics. In the proposed GA-DAN, a novel multi-modal spatial learning structure is designed which can convert a source-domain image into multiple images of different spatial views as in the target domain. A new disentangled cycle-consistency loss is introduced which balances the cycle consistency and greatly improves the concurrent learning in both appearance and geometry spaces. The proposed GA-DAN has been evaluated for the classic scene text detection and recognition tasks, and experiments show that the domain-adapted images achieve superior scene text detection and recognition performance while applied to network training.
****
Large-Scale Tag-Based Font Retrieval With Generative Feature Learning
Tianlang Chen, Zhaowen Wang, Ning Xu, Hailin Jin, Jiebo Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9116-9125
Font selection is one of the most important steps in a design workflow. Traditional methods rely on ordered lists which require significant domain knowledge and are often difficult to use even for trained professionals. In this paper, we address the problem of large-scale tag-based font retrieval which aims to bring semantics to the font selection process and enable people without expert knowledge to use fonts effectively. We collect a large-scale font tagging dataset of high-quality professional fonts. The dataset contains nearly 20,000 fonts, 2,000 tags, and hundreds of thousands of font-tag relations. We propose a novel generative feature learning algorithm that leverages the unique characteristics of fonts. The key idea is that font images are synthetic and can therefore be controlled by the learning algorithm. We design an integrated rendering and learning process so that the visual feature from one image can be used to reconstruct another image with different text. The resulting feature captures important font design details while is robust to nuisance factors such as text. We propose a novel attention mechanism to re-weight the visual feature for joint visual-text modeling. We combine the feature and the attention mechanism in a novel recognition-retrieval model. Experimental results show that our method significantly outperforms the state-of-the-art for the important problem of large-scale tag-based font retr

ieval.
****
Convolutional Character Networks
Linjie Xing, Zhi Tian, Weilin Huang, Matthew R. Scott; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9126-9136
Recent progress has been made on developing a unified framework for joint text detection and recognition in natural images, but existing joint models were mostly built on two-stage framework by involving ROI pooling, which can degrade the performance on recognition task. In this work, we propose convolutional character networks, referred as CharNet, which is an one-stage model that can process two tasks simultaneously in one pass. CharNet directly outputs bounding boxes of words and characters, with corresponding character labels. We utilize character as basic element, allowing us to overcome the main difficulty of existing approaches that attempted to optimize text detection jointly with a RNN-based recognition branch. In addition, we develop an iterative character detection approach able to transform the ability of character detection learned from synthetic data to real-world images. These technical improvements result in a simple, compact, yet powerful one-stage model that works reliably on multi-orientation and curved text. We evaluate CharNet on three standard benchmarks, where it consistently outperforms the state-of-the-art approaches [25, 24] by a large margin, e.g., with improvements of 65.33%->71.08% (with generic lexicon) on ICDAR 2015, and 54.0%->69.23% on Total-Text, on end-to-end text recognition. Code is available at: https://github.com/MalongTech/research-charnet.
****
Geometry Normalization Networks for Accurate Scene Text Detection
Youjiang Xu, Jiaqi Duan, Zhanghui Kuang, Xiaoyu Yue, Hongbin Sun, Yue Guan, Wayne Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9137-9146
Large geometry (e.g., orientation) variances are the key challenges in the scene text detection. In this work, we first conduct experiments to investigate the capacity of networks for learning geometry variances on detecting scene texts, and find that networks can handle only limited text geometry variances. Then, we put forward a novel Geometry Normalization Module (GNM) with multiple branches, each of which is composed of one Scale Normalization Unit and one Orientation Normalization Unit, to normalize each text instance to one desired canonical geometry range through at least one branch. The GNM is general and readily plugged into existing convolutional neural network based text detectors to construct end-to-end Geometry Normalization Networks (GNNets). Moreover, we propose a geometry-aware training scheme to effectively train the GNNets by sampling and augmenting text instances from a uniform geometry variance distribution. Finally, experiments on popular benchmarks of ICDAR 2015 and ICDAR 2017 MLT validate that our method outperforms all the state-of-the-art approaches remarkably by obtaining one-forward test F-scores of 88.52 and 74.54 respectively.
****
Symmetry-Constrained Rectification Network for Scene Text Recognition
Mingkun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui Bian, Song Bai, Cong Yao, Xiang Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9147-9156
Reading text in the wild is a very challenging task due to the diversity of text instances and the complexity of natural scenes. Recently, the community has paid increasing attention to the problem of recognizing text instances with irregular shapes. One intuitive and effective way to handle this problem is to rectify irregular text to a canonical form before recognition. However, these methods might struggle when dealing with highly curved or distorted text instances. To tackle this issue, we propose in this paper a Symmetry-constrained Rectification Network (ScRN) based on local attributes of text instances, such as center line, scale and orientation. Such constraints with an accurate description of text shape enable ScRN to generate better rectification results than existing methods and thus lead to higher recognition accuracy. Our method achieves state-of-the-art performance on text with both regular and irregular shapes. Specifically, the sy

stem outperforms existing algorithms by a large margin on datasets that contain quite a proportion of irregular text instances, e.g., ICDAR 2015, SVT-Perspective and CUTE80.
****

YOLACT: Real-Time Instance Segmentation
Daniel Bolya, Chong Zhou, Fanyi Xiao, Yong Jae Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9157-9166
We present a simple, fully-convolutional model for real-time instance segmentation that achieves 29.8 mAP on MS COCO at 33.5 fps evaluated on a single Titan Xp, which is significantly faster than any previous competitive approach. Moreover, we obtain this result after training on only one GPU. We accomplish this by breaking instance segmentation into two parallel subtasks: (1) generating a set of prototype masks and (2) predicting per-instance mask coefficients. Then we produce instance masks by linearly combining the prototypes with the mask coefficients. We find that because this process doesn't depend on repooling, this approach produces very high-quality masks and exhibits temporal stability for free. Furthermore, we analyze the emergent behavior of our prototypes and show they learn to localize instances on their own in a translation variant manner, despite being fully-convolutional. Finally, we also propose Fast NMS, a drop-in 12 ms faster replacement for standard NMS that only has a marginal performance penalty.
****

Expectation-Maximization Attention Networks for Semantic Segmentation
Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, Hong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9167-9176
Self-attention mechanism has been widely used for various tasks. It is designed to compute the representation of each position by a weighted sum of the features at all positions. Thus, it can capture long-range relations for computer vision tasks. However, it is computationally consuming. Since the attention maps are computed w.r.t all other positions. In this paper, we formulate the attention mechanism into an expectation-maximization manner and iteratively estimate a much more compact set of bases upon which the attention maps are computed. By a weighted summation upon these bases, the resulting representation is low-rank and deprecates noisy information from the input. The proposed Expectation-Maximization Attention (EMA) module is robust to the variance of input and is also friendly in memory and computation. Moreover, we set up the bases maintenance and normalization methods to stabilize its training procedure. We conduct extensive experiments on popular semantic segmentation benchmarks including PASCAL VOC, PASCAL Context, and COCO Stuff, on which we set new records.
****

Multi-Class Part Parsing With Joint Boundary-Semantic Awareness
Yifan Zhao, Jia Li, Yu Zhang, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9177-9186
Object part parsing in the wild, which requires to simultaneously detect multiple object classes in the scene and accurately segments semantic parts within each class, is challenging for the joint presence of class-level and part-level ambiguities. Despite its importance, however, this problem is not sufficiently explored in existing works. In this paper, we propose a joint parsing framework with boundary and semantic awareness to address this challenging problem. To handle part-level ambiguity, a boundary awareness module is proposed to make mid-level features at multiple scales attend to part boundaries for accurate part localization, which are then fused with high-level features for effective part recognition. For class-level ambiguity, we further present a semantic awareness module that selects discriminative part features relevant to a category to prevent irrelevant features being merged together. The proposed modules are lightweight and implementation friendly, improving the performance substantially when plugged into various baseline architectures. Without bells and whistles, the full model sets new state-of-the-art results on the Pascal-Part dataset, in both multi-class and the conventional single-class setting, while running substantially faster than recent high-performance approaches.

****
Explaining Neural Networks Semantically and Quantitatively
Runjin Chen, Hao Chen, Jie Ren, Ge Huang, Quanshi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9187-9196
This paper presents a method to pursue a semantic and quantitative explanation for the knowledge encoded in a convolutional neural network (CNN). The estimation of the specific rationale of each prediction made by the CNN presents a key issue of understanding neural networks, and it is of significant values in real applications. In this study, we propose to distill knowledge from the CNN into an explainable additive model, which explains the CNN prediction quantitatively. We discuss the problem of the biased interpretation of CNN predictions. To overcome the biased interpretation, we develop prior losses to guide the learning of the explainable additive model. Experimental results have demonstrated the effectiveness of our method.
****
PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment
Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9197-9206
Despite the great progress made by deep CNNs in image semantic segmentation, they typically require a large number of densely-annotated images for training and are difficult to generalize to unseen object categories. Few-shot segmentation has thus been developed to learn to perform segmentation from only a few annotated examples. In this paper, we tackle the challenging few-shot segmentation problem from a metric learning perspective and present PANet, a novel prototype alignment network to better utilize the information of the support set. Our PANet learns class-specific prototype representations from a few support images within an embedding space and then performs segmentation over the query images through matching each pixel to the learned prototypes. With non-parametric metric learning, PANet offers high-quality prototypes that are representative for each semantic class and meanwhile discriminative for different classes. Moreover, PANet introduces a prototype alignment regularization between support and query. With this, PANet fully exploits knowledge from the support and provides better generalization on few-shot segmentation. Significantly, our model achieves the mIoU score of 48.1% and 55.7% on PASCAL-5i for 1-shot and 5-shot settings respectively, surpassing the state-of-the-art method by 1.8% and 8.6%.
****
ShapeMask: Learning to Segment Novel Objects by Refining Shape Priors
Weicheng Kuo, Anelia Angelova, Jitendra Malik, Tsung-Yi Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9207-9216
Instance segmentation aims to detect and segment individual objects in a scene. Most existing methods rely on precise mask annotations of every category. However, it is difficult and costly to segment objects in novel categories because a large number of mask annotations is required. We introduce ShapeMask, which learns the intermediate concept of object shape to address the problem of generalization in instance segmentation to novel categories. ShapeMask starts with a bounding box detection and gradually refines it by first estimating the shape of the detected object through a collection of shape priors. Next, ShapeMask refines the coarse shape into an instance level mask by learning instance embeddings. The shape priors provide a strong cue for object-like prediction, and the instance embeddings model the instance specific appearance information. ShapeMask significantly outperforms the state-of-the-art by 6.4 and 3.8 AP when learning across categories, and obtains competitive performance in the fully supervised setting. It is also robust to inaccurate detections, decreased model capacity, and small training data. Moreover, it runs efficiently with 150ms inference time on a GPU and trains within 11 hours on TPUs. With a larger backbone model, ShapeMask increases the gap with state-of-the-art to 9.4 and 6.2 AP across categories. Code will be publicly available at: https://sites.google.com/view/shapemask/home.
****

Sequence Level Semantics Aggregation for Video Object Detection
Haiping Wu, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9217-9225
Video objection detection (VID) has been a rising research direction in recent years. A central issue of VID is the appearance degradation of video frames caused by fast motion. This problem is essentially ill-posed for a single frame. Therefore, aggregating features from other frames becomes a natural choice. Existing methods rely heavily on optical flow or recurrent neural networks for feature aggregation. However, these methods emphasize more on the temporally nearby frames. In this work, we argue that aggregating features in the full-sequence level will lead to more discriminative and robust features for video object detection. To achieve this goal, we devise a novel Sequence Level Semantics Aggregation (SELSA) module. We further demonstrate the close relationship between the proposed method and the classic spectral clustering method, providing a novel view for understanding the VID problem. We test the proposed method on the ImageNet VID and the EPIC KITCHENS dataset and achieve new state-of-the-art results. Our method does not need complicated postprocessing methods such as Seq-NMS or Tubelet rescoring, which keeps the pipeline simple and clean.
****

Video Object Segmentation Using Space-Time Memory Networks
Seoung Wug Oh, Joon-Young Lee, Ning Xu, Seon Joo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9226-9235
We propose a novel solution for semi-supervised video object segmentation. By the nature of the problem, available cues (e.g. video frame(s) with object masks) become richer with the intermediate predictions. However, the existing methods are unable to fully exploit this rich source of information. We resolve the issue by leveraging memory networks and learn to read relevant information from all available sources. In our framework, the past frames with object masks form an external memory, and the current frame as the query is segmented using the mask information in the memory. Specifically, the query and the memory are densely matched in the feature space, covering all the space-time pixel locations in a feed-forward fashion. Contrast to the previous approaches, the abundant use of the guidance information allows us to better handle the challenges such as appearance changes and occlussions. We validate our method on the latest benchmark sets and achieved the state-of-the-art performance (overall score of 79.4 on Youtube-VOS val set, J of 88.7 and 79.2 on DAVIS 2016/2017 val set respectively) while having a fast runtime (0.16 second/frame on DAVIS 2016 val set).
****

Zero-Shot Video Object Segmentation via Attentive Graph Neural Networks
Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, Ling Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9236-9245
This work proposes a novel attentive graph neural network (AGNN) for zero-shot video object segmentation (ZVOS). The suggested AGNN recasts this task as a process of iterative information fusion over video graphs. Specifically, AGNN builds a fully connected graph to efficiently represent frames as nodes, and relations between arbitrary frame pairs as edges. The underlying pair-wise relations are described by a differentiable attention mechanism. Through parametric message passing, AGNN is able to efficiently capture and mine much richer and higher-order relations between video frames, thus enabling a more complete understanding of video content and more accurate foreground estimation. Experimental results on three video segmentation datasets show that AGNN sets a new state-of-the-art in each case. To further demonstrate the generalizability of our framework, we extend AGNN to an additional task: image object co-segmentation (IOCS). We perform experiments on two famous IOCS datasets and observe again the superiority of our AGNN model. The extensive experiments verify that AGNN is able to learn the underlying semantic/appearance relationships among video frames or related images, and discover the common objects.
****

MeteorNet: Deep Learning on Dynamic 3D Point Cloud Sequences

Xingyu Liu, Mengyuan Yan, Jeannette Bohg; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9246-9255
Understanding dynamic 3D environment is crucial for robotic agents and many other applications. We propose a novel neural network architecture called MeteorNet for learning representations for dynamic 3D point cloud sequences. Different from previous work that adopts a grid-based representation and applies 3D or 4D convolutions, our network directly processes point clouds. We propose two ways to construct spatiotemporal neighborhoods for each point in the point cloud sequence. Information from these neighborhoods is aggregated to learn features per point. We benchmark our network on a variety of 3D recognition tasks including action recognition, semantic segmentation and scene flow estimation. MeteorNet shows stronger performance than previous grid-based methods while achieving state-of-the-art performance on Synthia. MeteorNet also outperforms previous baseline methods that are able to process at most two consecutive point clouds. To the best of our knowledge, this is the first work on deep learning for dynamic raw point cloud sequences.
****

3D Instance Segmentation via Multi-Task Metric Learning
Jean Lahoud, Bernard Ghanem, Marc Pollefeys, Martin R. Oswald; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9256-9266
We propose a novel method for instance label segmentation of dense 3D voxel grids. We target volumetric scene representations, which have been acquired with depth sensors or multi-view stereo methods and which have been processed with semantic 3D reconstruction or scene completion methods. The main task is to learn shape information about individual object instances in order to accurately separate them, including connected and incompletely scanned objects. We solve the 3D instance-labeling problem with a multi-task learning strategy. The first goal is to learn an abstract feature embedding, which groups voxels with the same instance label close to each other while separating clusters with different instance labels from each other. The second goal is to learn instance information by densely estimating directional information of the instance's center of mass for each voxel. This is particularly useful to find instance boundaries in the clustering post-processing step, as well as, for scoring the segmentation quality for the first goal. Both synthetic and real-world experiments demonstrate the viability and merits of our approach. In fact, it achieves state-of-the-art performance on the ScanNet 3D instance segmentation benchmark.
****

DeepGCNs: Can GCNs Go As Deep As CNNs?
Guohao Li, Matthias Muller, Ali Thabet, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9267-9276
Convolutional Neural Networks (CNNs) achieve impressive performance in a wide variety of fields. Their success benefited from a massive boost when very deep CNN models were able to be reliably trained. Despite their merits, CNNs fail to properly address problems with non-Euclidean data. To overcome this challenge, Graph Convolutional Networks (GCNs) build graphs to represent non-Euclidean data, borrow concepts from CNNs, and apply them in training. GCNs show promising results, but they are usually limited to very shallow models due to the vanishing gradient problem. As a result, most state-of-the-art GCN models are no deeper than 3 or 4 layers. In this work, we present new ways to successfully train very deep GCNs. We do this by borrowing concepts from CNNs, specifically residual/dense connections and dilated convolutions, and adapting them to GCN architectures. Extensive experiments show the positive effect of these deep GCN frameworks. Finally, we use these new concepts to build a very deep 56-layer GCN, and show how it significantly boosts performance (+3.7% mIoU over state-of-the-art) in the task of point cloud semantic segmentation. We believe that the community can greatly benefit from this work, as it opens up many opportunities for advancing GCN-based research.
****

Deep Hough Voting for 3D Object Detection in Point Clouds

Charles R. Qi, Or Litany, Kaiming He, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9277-9286
Current 3D object detection methods are heavily influenced by 2D detectors. In order to leverage architectures in 2D detectors, they often convert 3D point clouds to regular grids (i.e., to voxel grids or to bird's eye view images), or rely on detection in 2D images to propose 3D boxes. Few works have attempted to directly detect objects in point clouds. In this work, we return to first principles to construct a 3D detection pipeline for point cloud data and as generic as possible. However, due to the sparse nature of the data -- samples from 2D manifolds in 3D space -- we face a major challenge when directly predicting bounding box parameters from scene points: a 3D object centroid can be far from any surface point thus hard to regress accurately in one step. To address the challenge, we propose VoteNet, an end-to-end 3D object detection network based on a synergy of deep point set networks and Hough voting. Our model achieves state-of-the-art 3D detection on two large datasets of real 3D scans, ScanNet and SUN RGB-D with a simple design, compact model size and high efficiency. Remarkably, VoteNet outperforms previous methods by using purely geometric information without relying on color images.

****

M3D-RPN: Monocular 3D Region Proposal Network for Object Detection
Garrick Brazil, Xiaoming Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9287-9296
Understanding the world in 3D is a critical component of urban autonomous driving. Generally, the combination of expensive LiDAR sensors and stereo RGB imaging has been paramount for successful 3D object detection algorithms, whereas monocular image-only methods experience drastically reduced performance. We propose to reduce the gap by reformulating the monocular 3D detection problem as a standalone 3D region proposal network. We leverage the geometric relationship of 2D and 3D perspectives, allowing 3D boxes to utilize well-known and powerful convolutional features generated in the image-space. To help address the strenuous 3D parameter estimations, we further design depth-aware convolutional layers which enable location specific feature development and in consequence improved 3D scene understanding. Compared to prior work in monocular 3D detection, our method consists of only the proposed 3D region proposal network rather than relying on external networks, data, or multiple stages. M3D-RPN is able to significantly improve the performance of both monocular 3D Object Detection and Bird's Eye View tasks within the KITTI urban autonomous driving dataset, while efficiently using a shared multi-class model.

****

SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences
Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, Jurgen Gall; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9297-9307
Semantic scene understanding is important for various applications. In particular, self-driving cars need a fine-grained understanding of the surfaces and objects in their vicinity. Light detection and ranging (LiDAR) provides precise geometric information about the environment and is thus a part of the sensor suites of almost all self-driving cars. Despite the relevance of semantic scene understanding for this application, there is a lack of a large dataset for this task which is based on an automotive LiDAR. In this paper, we introduce a large dataset to propel research on laser-based semantic segmentation. We annotated all sequences of the KITTI Vision Odometry Benchmark and provide dense point-wise annotations for the complete 360-degree field-of-view of the employed automotive LiDAR. We propose three benchmark tasks based on this dataset: (i) semantic segmentation of point clouds using a single scan, (ii) semantic segmentation using multiple past scans, and (iii) semantic scene completion, which requires to anticipate the semantic scene in the future. We provide baseline experiments and show that there is a need for more sophisticated models to efficiently tackle these tasks. Our dataset opens the door for the development of more advanced methods, but also provides plentiful data to investigate new research directions.

****
WoodScape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving
Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier Perrotton, Patrick Perez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9308-9318
Fisheye cameras are commonly employed for obtaining a large field of view in surveillance, augmented reality and in particular automotive applications. In spite of their prevalence, there are few public datasets for detailed evaluation of computer vision algorithms on fisheye images. We release the first extensive fisheye automotive dataset, WoodScape, named after Robert Wood who invented the fisheye camera in 1906. WoodScape comprises of four surround view cameras and nine tasks including segmentation, depth estimation, 3D bounding box detection and soiling detection. Semantic annotation of 40 classes at the instance level is provided for over 10,000 images and annotation for other tasks are provided for over 100,000 images. With WoodScape, we would like to encourage the community to adapt computer vision models for fisheye camera instead of using naive rectification.
****
Scalable Place Recognition Under Appearance Change for Autonomous Driving
Anh-Dzung Doan, Yasir Latif, Tat-Jun Chin, Yu Liu, Thanh-Toan Do, Ian Reid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9319-9328
A major challenge in place recognition for autonomous driving is to be robust against appearance changes due to short-term (e.g., weather, lighting) and long-term (seasons, vegetation growth, etc.) environmental variations. A promising solution is to continuously accumulate images to maintain an adequate sample of the conditions and incorporate new changes into the place recognition decision. However, this demands a place recognition technique that is scalable on an ever growing dataset. To this end, we propose a novel place recognition technique that can be efficiently retrained and compressed, such that the recognition of new queries can exploit all available data (including recent changes) without suffering from visible growth in computational cost. Underpinning our method is a novel temporal image matching technique based on Hidden Markov Models. Our experiments show that, compared to state-of-the-art techniques, our method has much greater potential for large-scale place recognition for autonomous driving.
****
Exploring the Limitations of Behavior Cloning for Autonomous Driving
Felipe Codevilla, Eder Santana, Antonio M. Lopez, Adrien Gaidon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9329-9338
Driving requires reacting to a wide variety of complex environment conditions and agent behaviors. Explicitly modeling each possible scenario is unrealistic. In contrast, imitation learning can, in theory, leverage data from large fleets of human-driven cars. Behavior cloning in particular has been successfully used to learn simple visuomotor policies end-to-end, but scaling to the full spectrum of driving behaviors remains an unsolved problem. In this paper, we propose a new benchmark to experimentally investigate the scalability and limitations of behavior cloning. We show that behavior cloning leads to state-of-the-art results, executing complex lateral and longitudinal maneuvers, even in unseen environments, without being explicitly programmed to do so. However, we confirm some limitations of the behavior cloning approach: some well-known limitations (e.g., dataset bias and overfitting), new generalization issues (e.g., dynamic objects and the lack of a causal modeling), and training instabilities, all requiring further research before behavior cloning can graduate to real-world driving. The code, dataset, benchmark, and agent studied in this paper can be found at github.com/felipecode/coiltraine/blob/master/docs/exploring_limitations.md
****
Habitat: A Platform for Embodied AI Research

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, Dhruv Batra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9339-9347

We present Habitat, a platform for research in embodied artificial intelligence (AI). Habitat enables training embodied agents (virtual robots) in highly efficient photorealistic 3D simulation. Specifically, Habitat consists of: (i) Habitat-Sim: a flexible, high-performance 3D simulator with configurable agents, sensors, and generic 3D dataset handling. Habitat-Sim is fast -- when rendering a scene from Matterport3D, it achieves several thousand frames per second (fps) running single-threaded, and can reach over 10,000 fps multi-process on a single GPU. (ii) Habitat-API: a modular high-level library for end-to-end development of embodied AI algorithms -- defining tasks (e.g., navigation, instruction following, question answering), configuring, training, and benchmarking embodied agents. These large-scale engineering contributions enable us to answer scientific questions requiring experiments that were till now impracticable or 'merely' impractical. Specifically, in the context of point-goal navigation: (1) we revisit the comparison between learning and SLAM approaches from two recent works and find evidence for the opposite conclusion -- that learning outperforms SLAM if scaled to an order of magnitude more experience than previous investigations, and (2) we conduct the first cross-dataset generalization experiments train, test x Matterport3D, Gibson for multiple sensors blind, RGB, RGBD, D and find that only agents with depth (D) sensors generalize across datasets. We hope that our open-source platform and these findings will advance research in embodied AI.

****

Towards Interpretable Face Recognition

Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, Xiaoming Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9348-9357

Deep CNNs have been pushing the frontier of visual recognition over past years. Besides recognition accuracy, strong demands in understanding deep CNNs in the research community motivate developments of tools to dissect pre-trained models to visualize how they make predictions. Recent works further push the interpretability in the network learning stage to learn more meaningful representations. In this work, focusing on a specific area of visual recognition, we report our efforts towards interpretable face recognition. We propose a spatial activation diversity loss to learn more structured face representations. By leveraging the structure, we further design a feature activation diversity loss to push the interpretable representations to be discriminative and robust to occlusions. We demonstrate on three face recognition benchmarks that our proposed method is able to achieve the state-of-art face recognition accuracy with easily interpretable face representations.

****

Co-Mining: Deep Face Recognition With Noisy Labels

Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9358-9367

Face recognition has achieved significant progress with the growing scale of collected datasets, which empowers us to train strong convolutional neural networks (CNNs). While a variety of CNN architectures and loss functions have been devised recently, we still have a limited understanding of how to train the CNN models with the label noise inherent in existing face recognition datasets. To address this issue, this paper develops a novel co-mining strategy to effectively train on the datasets with noisy labels. Specifically, we simultaneously use the loss values as the cue to detect noisy labels, exchange the high-confidence clean faces to alleviate the errors accumulated issue caused by the sample-selection bias, and re-weight the predicted clean faces to make them dominate the discriminative model training in a mini-batch fashion. Extensive experiments by training on three popular datasets (i.e., CASIA-WebFace, MS-Celeb-1M and VggFace2) and testing on several benchmarks, including LFW, AgeDB, CFP, CALFW, CPLFW, RFW, and MegaFace, have demonstrated the effectiveness of our new approach over the state-o

f-the-art alternatives.
****

Few-Shot Adaptive Gaze Estimation
Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9368-9377
Inter-personal anatomical differences limit the accuracy of person-independent gaze estimation networks. Yet there is a need to lower gaze errors further to enable applications requiring higher quality. Further gains can be achieved by personalizing gaze networks, ideally with few calibration samples. However, over-parameterized neural networks are not amenable to learning from few examples as they can quickly over-fit. We embrace these challenges and propose a novel framework for Few-shot Adaptive GaZE Estimation (Faze) for learning person-specific gaze networks with very few (<= 9) calibration samples. Faze learns a rotation-aware latent representation of gaze via a disentangling encoder-decoder architecture along with a highly adaptable gaze estimator trained using meta-learning. It is capable of adapting to any new person to yield significant performance gains with as few as 3 samples, yielding state-of-the-art performance of 3.18-deg on Gaze Capture, a 19% improvement over prior art. We open-source our code at https://github.com/NVlabs/few_shot_gaze
****

Live Face De-Identification in Video
Oran Gafni, Lior Wolf, Yaniv Taigman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9378-9387
We propose a method for face de-identification that enables fully automatic video modification at high frame rates. The goal is to maximally decorrelate the identity, while having the perception (pose, illumination and expression) fixed. We achieve this by a novel feed-forward encoder-decoder network architecture that is conditioned on the high-level representation of a person's facial image. The network is global, in the sense that it does not need to be retrained for a given video or for a given identity, and it creates natural looking image sequences with little distortion in time.
****

Face Video Deblurring Using 3D Facial Priors
Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, Xin Tong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9388-9397
Existing face deblurring methods only consider single frames and do not account for facial structure and identity information. These methods struggle to deblur face videos that exhibit significant pose variations and misalignment. In this paper we propose a novel face video deblurring network capitalizing on 3D facial priors. The model consists of two main branches: i) a face video deblurring sub-network based on an encoder-decoder architecture, and ii) a 3D face reconstruction and rendering branch for predicting 3D priors of salient facial structures and identity knowledge. These structures encourage the deblurring branch to generate sharp faces with detailed structures. Our method not only uses low-level information (i.e., image intensity), but also middle-level information (i.e., 3D facial structure) and high-level knowledge (i.e., identity content) to further explore spatial constraints of facial components from blurry face frames. Extensive experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.
****

Semi-Supervised Monocular 3D Face Reconstruction With End-to-End Shape-Preserved Domain Transfer
Jingtan Piao, Chen Qian, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9398-9407
Monocular face reconstruction is a challenging task in computer vision, which aims to recover 3D face geometry from a single RGB face image. Recently, deep learning based methods have achieved great improvements on monocular face reconstruction. However, for deep learning-based methods to reach optimal performance, it

is paramount to have large-scale training images with ground-truth 3D face geome try, which is generally difficult for human to annotate. To tackle this problem, we propose a semi-supervised monocular reconstruction method, which jointly opt imizes a shape-preserved domain-transfer CycleGAN and a shape estimation network . The framework is semi-supervised trained with 3D rendered images with ground-t ruth shapes and in-the-wild face images without any extra annotation. The CycleG AN network transforms all realistic images to have the rendered style and is end -to-end trained within the overall framework. This is the key difference compare d with existing CycleGAN-based learning methods, which just used CycleGAN as a s eparate training sample generator. Novel landmark consistency loss and edge-awar e shape estimation loss are proposed for our two networks to jointly solve the c hallenging face reconstruction problem. Extensive experiments on public face rec onstruction datasets demonstrate the effectiveness of our overall method as well as the individual components.

****

3D Face Modeling From Diverse Raw Scan Data

Feng Liu, Luan Tran, Xiaoming Liu; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2019, pp. 9408-9418

Traditional 3D face models learn a latent representation of faces using linear s ubspaces from limited scans of a single database. The main roadblock of building a large-scale face model from diverse 3D databases lies in the lack of dense co rrespondence among raw scans. To address these problems, this paper proposes an innovative framework to jointly learn a nonlinear face model from a diverse set of raw 3D scan databases and establish dense point-to-point correspondence among their scans. Specifically, by treating input scans as unorganized point clouds, we explore the use of PointNet architectures for converting point clouds to ide ntity and expression feature representations, from which the decoder networks re cover their 3D face shapes. Further, we propose a weakly supervised learning app roach that does not require correspondence label for the scans. We demonstrate t he superior dense correspondence and representation power of our proposed method , and its contribution to single-image 3D face reconstruction.

****

A Decoupled 3D Facial Shape Model by Adversarial Training

Victoria Fernandez Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, Edmond Boyer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9419-9428

Data-driven generative 3D face models are used to compactly encode facial shape data into meaningful parametric representations. A desirable property of these m odels is their ability to effectively decouple natural sources of variation, in particular identity and expression. While factorized representations have been p roposed for that purpose, they are still limited in the variability they can cap ture and may present modeling artifacts when applied to tasks such as expression transfer. In this work, we explore a new direction with Generative Adversarial Networks and show that they contribute to better face modeling performances, esp ecially in decoupling natural factors, while also achieving more diverse samples . To train the model we introduce a novel architecture that combines a 3D genera tor with a 2D discriminator that leverages conventional CNNs, where the two comp onents are bridged by a geometry mapping layer. We further present a training sc heme, based on auxiliary classifiers, to explicitly disentangle identity and exp ression attributes. Through quantitative and qualitative results on standard fac e datasets, we illustrate the benefits of our model and demonstrate that it outp erforms competing state of the art methods in terms of decoupling and diversity.

****

Photo-Realistic Facial Details Synthesis From Single Image

Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, Jingyi Yu; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 942 9-9439

We present a single-image 3D face synthesis technique that can handle challengin g facial expressions while recovering fine geometric details. Our technique empl oys expression analysis for proxy face geometry generation and combines supervis

ed and unsupervised learning for facial detail synthesis. On proxy generation, w
e conduct emotion prediction to determine a new expression-informed proxy. On de
tail synthesis, we present a Deep Facial Detail Net (DFDN) based on Conditional
Generative Adversarial Net (CGAN) that employs both geometry and appearance loss
 functions. For geometry, we capture 366 high-quality 3D scans from 122 differen
t subjects under 3 facial expressions. For appearance, we use additional 163K in
-the-wild face images and apply image-based rendering to accommodate lighting va
riations. Comprehensive experiments demonstrate that our framework can produce h
igh-quality 3D faces with realistic details under challenging facial expressions
.
****

S2GAN: Share Aging Factors Across Ages and Share Aging Trends Among Individuals
Zhenliang He, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2019, pp. 9440-9449
Generally, we human follow the roughly common aging trends, e.g., the wrinkles o
nly tend to be more, longer or deeper. However, the aging process of each indivi
dual is more dominated by his/her personalized factors, including the invariant
factors such as identity and mole, as well as the personalized aging patterns, e
.g., one may age by graying hair while another may age by receding hairline. Fol
lowing this biological principle, in this work, we propose an effective and effi
cient method to simulate natural aging. Specifically, a personalized aging basis
 is established for each individual to depict his/her own aging factors. Then di
fferent ages share this basis, being derived through age-specific transforms. Th
e age-specific transforms represent the aging trends which are shared among all
individuals. The proposed method can achieve continuous face aging with favorabl
e aging accuracy, identity preservation, and fidelity. Furthermore, befitted fro
m the effective design, a unique model is capable of all ages and the prediction
 time is significantly saved.
****

PuppetGAN: Cross-Domain Image Manipulation by Demonstration
Ben Usman, Nick Dufour, Kate Saenko, Chris Bregler; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2019, pp. 9450-9458
In this work we propose a model that can manipulate individual visual attributes
 of objects in a real scene using examples of how respective attribute manipulat
ions affect the output of a simulation. As an example, we train our model to man
ipulate the expression of a human face using nonphotorealistic 3D renders of a f
ace with varied expression. Our model manages to preserve all other visual attri
butes of a real face, such as head orientation, even though this and other attri
butes are not labeled in either real or synthetic domain. Since our model learns
 to manipulate a specific property in isolation using only "synthetic demonstrat
ions" of such manipulations without explicitly provided labels, it can be applie
d to shape, texture, lighting, and other properties that are difficult to measur
e or represent as real-valued vectors. We measure the degree to which our model
preserves other attributes of a real image when a single specific attribute is m
anipulated. We use digit datasets to analyze how discrepancy in attribute distri
butions affects the performance of our model, and demonstrate results in a far m
ore difficult setting: learning to manipulate real human faces using nonphotorea
listic 3D renders.
****

Few-Shot Adversarial Learning of Realistic Neural Talking Head Models
Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, Victor Lempitsky; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp.
 9459-9468
Several recent works have shown how highly realistic human head images can be ob
tained by training convolutional neural networks to generate them. In order to c
reate a personalized talking head model, these works require training on a large
 dataset of images of a single person. However, in many practical scenarios, suc
h personalized talking head models need to be learned from a few image views of
a person, potentially even a single image. Here, we present a system with such f
ew-shot capability. It performs lengthy meta-learning on a large dataset of vide

os, and after that is able to frame few- and one-shot learning of neural talking head models of previously unseen people as adversarial training problems with high capacity generators and discriminators. Crucially, the system is able to initialize the parameters of both the generator and the discriminator in a person-specific way, so that training can be based on just a few images and done quickly, despite the need to tune tens of millions of parameters. We show that such an approach is able to learn highly realistic and personalized talking head models of new people and even portrait paintings.
****

Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection
Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, Xuming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9469-9478
Reasoning human object interactions is a core problem in human-centric scene understanding and detecting such relations poses a unique challenge to vision systems due to large variations in human-object configurations, multiple co-occurring relation instances and subtle visual difference between relation categories. To address those challenges, we propose a multi-level relation detection strategy that utilizes human pose cues to capture global spatial configurations of relations and as an attention mechanism to dynamically zoom into relevant regions at human part level. We develop a multi-branch deep network to learn a pose-augmented relation representation at three semantic levels, incorporating interaction context, object features and detailed semantic part cues. As a result, our approach is capable of generating robust predictions on fine-grained human object interactions with interpretable outputs. Extensive experimental evaluations on public benchmarks show that our model outperforms prior methods by a considerable margin, demonstrating its efficacy in handling complex scenes.
****

TRB: A Novel Triplet Representation for Understanding 2D Human Body
Haodong Duan, Kwan-Yee Lin, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9479-9488
Human pose and shape are two important components of 2D human body. However, how to efficiently represent both of them in images is still an open question. In this paper, we propose the Triplet Representation for Body (TRB) --- a compact 2D human body representation, with skeleton keypoints capturing human pose information and contour keypoints containing human shape information. TRB not only preserves the flexibility of skeleton keypoint representation, but also contains rich pose and human shape information. Therefore, it promises broader application areas, such as human shape editing and conditional image generation. We further introduce the challenging problem of TRB estimation, where joint learning of human pose and shape is required. We construct several large-scale TRB estimation datasets, based on the popular 2D pose datasets LSP, MPII and COCO. To effectively solve TRB estimation, we propose a two-branch network (TRB-net) with three novel techniques, namely X-structure (Xs), Directional Convolution (DC) and Pairwise mapping (PM), to enforce multi-level message passing for joint feature learning. We evaluate our proposed TRB-net and several leading approaches on our proposed TRB datasets, and demonstrate the superiority of our method through extensive evaluations.
****

Learning Trajectory Dependencies for Human Motion Prediction
Wei Mao, Miaomiao Liu, Mathieu Salzmann, Hongdong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9489-9497
Human motion prediction, i.e., forecasting future body poses given observed pose sequence, has typically been tackled with recurrent neural networks (RNNs). However, as evidenced by prior work, the resulted RNN models suffer from prediction errors accumulation, leading to undesired discontinuities in motion prediction. In this paper, we propose a simple feed-forward deep network for motion prediction, which takes into account both temporal smoothness and spatial dependencies among human body joints. In this context, we then propose to encode temporal information by working in trajectory space, instead of the traditionally-used pose

space. This alleviates us from manually defining the range of temporal dependencies (or temporal convolutional filter size, as done in previous work). Moreover, spatial dependency of human pose is encoded by treating a human pose as a generic graph (rather than a human skeletal kinematic tree) formed by links between every pair of body joints. Instead of using a pre-defined graph structure, we design a new graph convolutional network to learn graph connectivity automatically. This allows the network to capture long range dependencies beyond that of human kinematic tree. We evaluate our approach on several standard benchmark datasets for motion prediction, including Human3.6M, the CMU motion capture dataset and 3DPW. Our experiments clearly demonstrate that the proposed approach achieves state of the art performance, and is applicable to both angle-based and position-based pose representations. The code is available at https://github.com/wei-mao-2019/LearnTrajDep

****

Cross-Domain Adaptation for Animal Pose Estimation

Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, Yu-Wing Tai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9498-9507

In this paper, we are interested in pose estimation of animals. Animals usually exhibit a wide range of variations on poses and there is no available animal pose dataset for training and testing. To address this problem, we build an animal pose dataset to facilitate training and evaluation. Considering the heavy labor needed to label dataset and it is impossible to label data for all concerned animal species, we, therefore, proposed a novel cross-domain adaptation method to transform the animal pose knowledge from labeled animal classes to unlabeled animal classes. We use the modest animal pose dataset to adapt learned knowledge to multiple animals species. Moreover, humans also share skeleton similarities with some animals (especially four-footed mammals). Therefore, the easily available human pose dataset, which is of a much larger scale than our labeled animal dataset, provides important prior knowledge to boost up the performance on animal pose estimation. Experiments show that our proposed method leverages these pieces of prior knowledge well and achieves convincing results on animal pose estimation.

****

NOTE-RCNN: NOise Tolerant Ensemble RCNN for Semi-Supervised Object Detection

Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, Ram Nevatia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9508-9517

The labeling cost of large number of bounding boxes is one of the main challenges for training modern object detectors. To reduce the dependence on expensive bounding box annotations, we propose a new semi-supervised object detection formulation, in which a few seed box level annotations and a large scale of image level annotations are used to train the detector. We adopt a training-mining framework, which is widely used in weakly supervised object detection tasks. However, the mining process inherently introduces various kinds of labelling noises: false negatives, false positives and inaccurate boundaries, which can be harmful for training the standard object detectors (e.g. Faster RCNN). We propose a novel NOise Tolerant Ensemble RCNN (NOTE-RCNN) object detector to handle such noisy labels. Comparing to standard Faster RCNN, it contains three highlights: an ensemble of two classification heads and a distillation head to avoid overfitting on noisy labels and improve the mining precision, masking the negative sample loss in box predictor to avoid the harm of false negative labels, and training box regression head only on seed annotations to eliminate the harm from inaccurate boundaries of mined bounding boxes. We evaluate the methods on ILSVRC 2013 and MSCOCO 2017 dataset; we observe that the detection accuracy consistently improves as we iterate between mining and training steps, and state-of-the-art performance is achieved.

****

Unsupervised Out-of-Distribution Detection by Maximum Classifier Discrepancy

Qing Yu, Kiyoharu Aizawa; Proceedings of the IEEE/CVF International Conference

on Computer Vision (ICCV), 2019, pp. 9518-9526
Since deep learning models have been implemented in many commercial applications , it is important to detect out-of-distribution (OOD) inputs correctly to mainta in the performance of the models, ensure the quality of the collected data, and prevent the applications from being used for other-than-intended purposes. In th is work, we propose a two-head deep convolutional neural network (CNN) and maxim ize the discrepancy between the two classifiers to detect OOD inputs. We train a two-head CNN consisting of one common feature extractor and two classifiers whi ch have different decision boundaries but can classify in-distribution (ID) samp les correctly. Unlike previous methods, we also utilize unlabeled data for unsup ervised training and we use these unlabeled data to maximize the discrepancy bet ween the decision boundaries of two classifiers to push OOD samples outside the manifold of the in-distribution (ID) samples, which enables us to detect OOD sam ples that are far from the support of the ID samples. Overall, our approach sign ificantly outperforms other state-of-the-art methods on several OOD detection be nchmarks and two cases of real-world simulation.
****

SBSGAN: Suppression of Inter-Domain Background Shift for Person Re-Identificatio n
Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 9527-9536
Cross-domain person re-identification (re-ID) is challenging due to the bias bet ween training and testing domains. We observe that if backgrounds in the trainin g and testing datasets are very different, it dramatically introduces difficulti es to extract robust pedestrian features, and thus compromises the cross-domain person re-ID performance. In this paper, we formulate such problems as a backgro und shift problem. A Suppression of Background Shift Generative Adversarial Netw ork (SBSGAN) is proposed to generate images with suppressed backgrounds. Unlike simply removing backgrounds using binary masks, SBSGAN allows the generator to d ecide whether pixels should be preserved or suppressed to reduce segmentation er rors caused by noisy foreground masks. Additionally, we take ID-related cues, su ch as vehicles and companions into consideration. With high-quality generated im ages, a Densely Associated 2-Stream (DA-2S) network is introduced with Inter Str eam Densely Connection (ISDC) modules to strengthen the complementarity of the g enerated data and ID-related cues. The experiments show that the proposed method achieves competitive performance on three re-ID datasets, i.e., Market-1501, Du keMTMC-reID, and CUHK03, under the cross-domain person re-ID scenario.
****

Enriched Feature Guided Refinement Network for Object Detection
Jing Nie, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei P ang, Ling Shao; Proceedings of the IEEE/CVF International Conference on Compute r Vision (ICCV), 2019, pp. 9537-9546
We propose a single-stage detection framework that jointly tackles the problem o f multi-scale object detection and class imbalance. Rather than designing deeper networks, we introduce a simple yet effective feature enrichment scheme to prod uce multi-scale contextual features. We further introduce a cascaded refinement scheme which first instills multi-scale contextual features into the prediction layers of the single-stage detector in order to enrich their discriminative powe r for multi-scale detection. Second, the cascaded refinement scheme counters the class imbalance problem by refining the anchors and enriched features to improv e classification and regression. Experiments are performed on two benchmarks: PA SCAL VOC and MS COCO. For a 320x320 input on the MS COCO test-dev, our detector achieves state-of-the-art single-stage detection accuracy with a COCO AP of 33.2 in the case of single-scale inference, while operating at 21 milliseconds on a Titan XP GPU. For a 512x512 input on the MS COCO test-dev, our approach obtains an absolute gain of 1.6% in terms of COCO AP, compared to the best reported sing le-stage results[5]. Source code and models are available at: https://github.com /Ranchentx/EFGRNet.
****

Deep Meta Metric Learning

Guangyi Chen, Tianren Zhang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9547-9556
In this paper, we present a deep meta metric learning (DMML) approach for visual recognition. Unlike most existing deep metric learning methods formulating the learning process by an overall objective, our DMML formulates the metric learning in a meta way, and proves that softmax and triplet loss are consistent in the meta space. Specifically, we sample some subsets from the original training set and learn metrics across different subsets. In each sampled sub-task, we split the training data into a support set as well as a query set, and learn the set-based distance, instead of sample-based one, to verify the query cell from multiple support cells. In addition, we introduce hard sample mining for set-based distance to encourage the intra-class compactness. Experimental results on three visual recognition applications including person re-identification, vehicle re-identification and face verification show that the proposed DMML method outperforms most existing approaches.
****

Discriminative Feature Transformation for Occluded Pedestrian Detection
Chunluan Zhou, Ming Yang, Junsong Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9557-9566
Despite promising performance achieved by deep con- volutional neural networks for non-occluded pedestrian de- tection, it remains a great challenge to detect partially oc- cluded pedestrians. Compared with non-occluded pedes- trian examples, it is generally more difficult to distinguish occluded pedestrian examples from background in featue space due to the missing of occluded parts. In this paper, we propose a discriminative feature transformation which en- forces feature separability of pedestrian and non-pedestrian examples to handle occlusions for pedestrian detection. Specifically, in feature space it makes pedestrian exam- ples approach the centroid of easily classified non-occluded pedestrian examples and pushes non-pedestrian examples close to the centroid of easily classified non-pedestrian ex- amples. Such a feature transformation partially compen- sates the missing contribution of occluded parts in feature space, therefore improving the performance for occluded pedestrian detection. We implement our approach in the Fast R-CNN framework by adding one transformation net- work branch. We validate the proposed approach on two widely used pedestrian detection datasets: Caltech and CityPersons. Experimental results show that our approach achieves promising performance for both non-occluded and occluded pedestrian detection.
****

Contextual Attention for Hand Detection in the Wild
Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, Minh Hoai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9567-9576
We present Hand-CNN, a novel convolutional network architecture for detecting hand masks and predicting hand orientations in unconstrained images. Hand-CNN extends MaskRCNN with a novel attention mechanism to incorporate contextual cues in the detection process. This attention mechanism can be implemented as an efficient network module that captures non-local dependencies between features. This network module can be inserted at different stages of an object detection network, and the entire detector can be trained end-to-end. We also introduce large-scale annotated hand datasets containing hands in unconstrained images for training and evaluation. We show that Hand-CNN outperforms existing methods on the newly collected datasets and the publicly available PASCAL VOC human layout dataset. Data and code: https://www3.cs.stonybrook.edu/ cvl/projects/hand_det_attention/
****

Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning
Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, Liang Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9577-9586
Resembling the rapid learning capability of human, low-shot learning empowers vision systems to understand new concepts by training with few samples. Leading approaches derived from meta-learning on images with a single visual object. Obfus

cated by a complex background and multiple objects in one image, they are hard t
o promote the research of low-shot object detection/segmentation. In this work,
we present a flexible and general methodology to achieve these tasks. Our work e
xtends Faster /Mask R-CNN by proposing meta-learning over RoI (Region-of-Interes
t) features instead of a full image feature. This simple spirit disentangles mul
ti-object information merged with the background, without bells and whistles, en
abling Faster /Mask R-CNN turn into a meta-learner to achieve the tasks. Specifi
cally, we introduce a Predictor-head Remodeling Network (PRN) that shares its ma
in backbone with Faster /Mask R-CNN. PRN receives images containing low-shot obj
ects with their bounding boxes or masks to infer their class attentive vectors.
The vectors take channel-wise soft-attention on RoI features, remodeling those R
-CNN predictor heads to detect or segment the objects consistent with the classe
s these vectors represent. In our experiments, Meta R-CNN yields the new state o
f the art in low-shot object detection and improves low-shot object segmentation
 by Mask R-CNN. Code: https://yanxp.github.io/metarcnn.html.
****

Pyramid Graph Networks With Connection Attentions for Region-Based One-Shot Sema
ntic Segmentation
Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, Rui Yao; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 201
9, pp. 9587-9595
One-shot image segmentation aims to undertake the segmentation task of a novel c
lass with only one training image available. The difficulty lies in that image s
egmentation has structured data representations, which yields a many-to-many mes
sage passing problem. Previous methods often simplify it to a one-to-many proble
m by squeezing support data to a global descriptor. However, a mixed global repr
esentation drops the data structure and information of individual elements. In t
his paper, we propose to model structured segmentation data with graphs and appl
y attentive graph reasoning to propagate label information from support data to
query data. The graph attention mechanism could establish the element-to-element
 correspondence across structured data by learning attention weights between con
nected graph nodes. To capture correspondence at different semantic levels, we f
urther propose a pyramid-like structure that models different sizes of image reg
ions as graph nodes and undertakes graph reasoning at different levels. Experime
nts on PASCAL VOC 2012 dataset demonstrate that our proposed network significant
ly outperforms the baseline method and leads to new state-of-the-art performance
 on 1-shot and 5-shot segmentation benchmarks.
****

Presence-Only Geographical Priors for Fine-Grained Image Classification
Oisin Mac Aodha, Elijah Cole, Pietro Perona; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2019, pp. 9596-9606
Appearance information alone is often not sufficient to accurately differentiate
 between fine-grained visual categories. Human experts make use of additional cu
es such as where, and when, a given image was taken in order to inform their fin
al decision. This contextual information is readily available in many online ima
ge collections but has been underutilized by existing image classifiers that foc
us solely on making predictions based on the image contents. We propose an effic
ient spatio-temporal prior, that when conditioned on a geographical location and
 time, estimates the probability that a given object category occurs at that loc
ation. Our prior is trained from presence-only observation data and jointly mode
ls object categories, their spatio-temporal distributions, and photographer bias
es. Experiments performed on multiple challenging image classification datasets
show that combining our prior with the predictions from image classifiers result
s in a large improvement in final classification performance.
****

POD: Practical Object Detection With Scale-Sensitive Network
Junran Peng, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, Junjie Yan; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 96
07-9616
Scale-sensitive object detection remains a challenging task, where most of the e

xisting methods not learn it explicitly and not robust to scale variance. In addition, the most existing methods are less efficient during training or slow during inference, which are not friendly to real-time application. In this paper, we propose a practical object detection with scale-sensitive network.Our method first predicts a global continuous scale ,which shared by all position, for each convolution filter of each network stage. To effectively learn the scale, we average the spatial features and distill the scale from channels. For fast-deployment, we propose a scale decomposition method that transfers the robust fractional scale into combinations of fixed integral scales for each convolution filter, which exploit the dilated convolution. We demonstrate it on one-stage and two-stage algorithm under almost different configure. For practical application, training of our method is of efficiency and simplicity which gets rid of complex data sampling or optimize strategy. During testing, the proposed method requires no extra operation and is very friendly to hardware acceleration like TensorRT and TVM.On the COCO test-dev, our model could achieve a 41.5mAP on one-stage detector and 42.1 mAP on two-stage detectors based on ResNet-101, outperforming baselines by 2.4 and 2.1 respectively without extra FLOPS.
****

Human Uncertainty Makes Classification More Robust
Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, Olga Russakovsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9617-9626
The classification performance of deep neural networks has begun to asymptote at near-perfect levels. However, their ability to generalize outside the training set and their robustness to adversarial attacks have not. In this paper, we make progress on this problem by training with full label distributions that reflect human perceptual uncertainty. We first present a new benchmark dataset which we call CIFAR10H, containing a full distribution of human labels for each image of the CIFAR10 test set. We then show that, while contemporary classifiers fail to exhibit human-like uncertainty on their own, explicit training on our dataset closes this gap, supports improved generalization to increasingly out-of-training-distribution test datasets, and confers robustness to adversarial attacks.
****

FCOS: Fully Convolutional One-Stage Object Detection
Zhi Tian, Chunhua Shen, Hao Chen, Tong He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9627-9636
We propose a fully convolutional one-stage object detector (FCOS) to solve object detection in a per-pixel prediction fashion, analogue to semantic segmentation. Almost all state-of-the-art object detectors such as RetinaNet, SSD, YOLOv3, and Faster R-CNN rely on pre-defined anchor boxes. In contrast, our proposed detector FCOS is anchor box free, as well as proposal free. By eliminating the pre-defined set of anchor boxes, FCOS completely avoids the complicated computation related to anchor boxes such as calculating overlapping during training. More importantly, we also avoid all hyper-parameters related to anchor boxes, which are often very sensitive to the final detection performance. With the only post-processing non-maximum suppression (NMS), FCOS with ResNeXt-64x4d-101 achieves 44.7% in AP with single-model and single-scale testing, surpassing previous one-stage detectors with the advantage of being much simpler. For the first time, we demonstrate a much simpler and flexible detection framework achieving improved detection accuracy. We hope that the proposed FCOS framework can serve as a simple and strong alternative for many other instance-level tasks. Code is available at: https://tinyurl.com/FCOSv1
****

Self-Critical Attention Learning for Person Re-Identification
Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9637-9646
In this paper, we propose a self-critical attention learning method for person re-identification. Unlike most existing methods which train the attention mechanism in a weakly-supervised manner and ignore the attention confidence level, we l

earn the attention with a critic which measures the attention quality and provid
es a powerful supervisory signal to guide the learning process. Moreover, the cr
itic model facilitates the interpretation of the effectiveness of the attention
mechanism during the learning process, by estimating the quality of the attentio
n maps. Specifically, we jointly train our attention agent and critic in a reinf
orcement learning manner, where the agent produces the visual attention while th
e critic analyzes the gain from the attention and guides the agent to maximize t
his gain. We design spatial- and channel-wise attention models with our critic m
odule and evaluate them on three popular benchmarks including Market-1501, DukeM
TMC-ReID, and CUHK03. The experimental results demonstrate the superiority of ou
r method, which outperforms the state-of-the-art methods by a large margin of 5.
9%/2.1%, 6.3%/3.0%, and 10.5%/9.5% on mAP/Rank-1, respectively.
****

Temporal Knowledge Propagation for Image-to-Video Person Re-Identification
Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, Xilin Chen; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 96
47-9656
In many scenarios of Person Re-identification (Re-ID), the gallery set consists
of lots of surveillance videos and the query is just an image, thus Re-ID has to
be conducted between image and videos. Compared with videos, still person image
s lack temporal information. Besides, the information asymmetry between image an
d video features increases the difficulty in matching images and videos. To solv
e this problem, we propose a novel Temporal Knowledge Propagation (TKP) method w
hich propagates the temporal knowledge learned by the video representation netwo
rk to the image representation network. Specifically, given the input videos, we
enforce the image representation network to fit the outputs of video representa
tion network in a shared feature space. With back propagation, temporal knowledg
e can be transferred to enhance the image features and the information asymmetry
problem can be alleviated. With additional classification and integrated triple
t losses, our model can learn expressive and discriminative image and video feat
ures for image-to-video re-identification. Extensive experiments demonstrate the
effectiveness of our method and the overall results on two widely used datasets
surpass the state-of-the-art methods by a large margin.
****

RepPoints: Point Set Representation for Object Detection
Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, Stephen Lin; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9657-9666
Modern object detectors rely heavily on rectangular bounding boxes, such as anch
ors, proposals and the final predictions, to represent objects at various recogn
ition stages. The bounding box is convenient to use but provides only a coarse l
ocalization of objects and leads to a correspondingly coarse extraction of objec
t features. In this paper, we present RepPoints (representative points), a new f
iner representation of objects as a set of sample points useful for both localiz
ation and recognition. Given ground truth localization and recognition targets f
or training, RepPoints learn to automatically arrange themselves in a manner tha
t bounds the spatial extent of an object and indicates semantically significant
local areas. They furthermore do not require the use of anchors to sample a spac
e of bounding boxes. We show that an anchor-free object detector based on RepPoi
nts can be as effective as the state-of-the-art anchor-based detection methods,
with 46.5 AP and 67.4 AP_ 50 on the COCO test-dev detection benchmark, using Re
sNet-101 model. Code is available at https://github.com/microsoft/RepPoints \c
olor cyan https://github.com/microsoft/RepPoints .
****

SegEQA: Video Segmentation Based Visual Attention for Embodied Question Answerin
g
Haonan Luo, Guosheng Lin, Zichuan Liu, Fayao Liu, Zhenmin Tang, Yazhou Yao;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2019, pp. 9667-9676
Embodied Question Answering (EQA) is a newly defined research area where an agen
t is required to answer the user's questions by exploring the real world environ

ment. It has attracted increasing research interests due to its broad applicatio ns in automatic driving system, in-home robots, and personal assistants. Most of the existing methods perform poorly in terms of answering and navigation accura cy due to the absence of local details and vulnerability to the ambiguity caused by complicated vision conditions. To tackle these problems, we propose a segmen tation based visual attention mechanism for Embodied Question Answering. Firstly , We extract the local semantic features by introducing a novel high-speed video segmentation framework. Then by the guide of extracted semantic features, a bot tom-up visual attention mechanism is proposed for the Visual Question Answering (VQA) sub-task. Further, a feature fusion strategy is proposed to guide the trai ning of the navigator without much additional computational cost. The ablation e xperiments show that our method boosts the performance of VQA module by 4.2% (68 .99% vs 64.73%) and leads to 3.6% (48.59% vs 44.98%) overall improvement in EQA accuracy.

****

No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, a nd Training Techniques

Tanmay Gupta, Alexander Schwing, Derek Hoiem; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2019, pp. 9677-9685

We show that for human-object interaction detection a relatively simple factoriz ed model with appearance and layout encodings constructed from pre-trained objec t detectors outperforms more sophisticated approaches. Our model includes factor s for detection scores, human and object appearance, and coarse (box-pair config uration) and optionally fine-grained layout (human pose). We also develop traini ng techniques that improve learning efficiency by: (1) eliminating a train-infer ence mismatch; (2) rejecting easy negatives during mini-batch training; and (3) using a ratio of negatives to positives that is two orders of magnitude larger t han existing approaches. We conduct a thorough ablation study to understand the importance of different factors and training techniques using the challenging HI CO-Det dataset.

****

Cap2Det: Learning to Amplify Weak Caption Supervision for Object Detection

Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, Jesse Berent ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 9686-9695

Learning to localize and name object instances is a fundamental problem in visio n, but state-of-the-art approaches rely on expensive bounding box supervision. W hile weakly supervised detection (WSOD) methods relax the need for boxes to that of image-level annotations, even cheaper supervision is naturally available in the form of unstructured textual descriptions that users may freely provide when uploading image content. However, straightforward approaches to using such data for WSOD wastefully discard captions that do not exactly match object names. In stead, we show how to squeeze the most information out of these captions by trai ning a text-only classifier that generalizes beyond dataset boundaries. Our disc overy provides an opportunity for learning detection models from noisy but more abundant and freely-available caption data. We also validate our model on three classic object detection benchmarks and achieve state-of-the-art WSOD performanc e. Our code is available at https://github.com/yekeren/Cap2Det.

****

No Fear of the Dark: Image Retrieval Under Varying Illumination Conditions

Tomas Jenicek, Ondrej Chum; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2019, pp. 9696-9704

Image retrieval under varying illumination conditions, such as day and night ima ges, is addressed by image preprocessing, both hand-crafted and learned. Prior t o extracting image descriptors by a convolutional neural network, images are pho tometrically normalised in order to reduce the descriptor sensitivity to illumin ation changes. We propose a learnable normalisation based on the U-Net architect ure, which is trained on a combination of single-camera multi-exposure images an d a newly constructed collection of similar views of landmarks during day and ni ght. We experimentally show that both hand-crafted normalisation based on local

histogram equalisation and the learnable normalisation outperform standard appro
aches in varying illumination conditions, while staying on par with the state-of
-the-art methods on daylight illumination benchmarks, such as Oxford or Paris da
tasets.
****

Hierarchical Shot Detector
Jiale Cao, Yanwei Pang, Jungong Han, Xuelong Li; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2019, pp. 9705-9714
Single shot detector simultaneously predicts object categories and regression of
fsets of the default boxes. Despite of high efficiency, this structure has some
inappropriate designs: (1) The classification result of the default box is impro
perly assigned to that of the regressed box during inference, (2) Only regressio
n once is not good enough for accurate object detection. To solve the first prob
lem, a novel reg-offset-cls (ROC) module is proposed. It contains three hierarch
ical steps: box regression, the feature sampling location predication, and the r
egressed box classification with the features of offset locations. To further so
lve the second problem, a hierarchical shot detector (HSD) is proposed, which st
acks two ROC modules and one feature enhanced module. The second ROC treats the
regressed boxes and the feature sampling locations of features in the first ROC
as the inputs. Meanwhile, the feature enhanced module injected between two ROCs
aims to extract the local and non-local context. Experiments on the MS COCO and
PASCAL VOC datasets demonstrate the superiority of proposed HSD. Without the bel
ls or whistles, HSD outperforms all one-stage methods at real-time speed.
****

Few-Shot Learning With Global Class Representations
Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, Liwei Wang; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9715-97
24
In this paper, we propose to tackle the challenging few-shot learning (FSL) prob
lem by learning global class representations using both base and novel class tra
ining samples. In each training episode, an episodic class mean computed from a
support set is registered with the global representation via a registration modu
le. This produces a registered global class representation for computing the cla
ssification loss using a query set. Though following a similar episodic training
 pipeline as existing meta learning based approaches, our method differs signifi
cantly in that novel class training samples are involved in the training from th
e beginning. To compensate for the lack of novel class training samples, an effe
ctive sample synthesis strategy is developed to avoid overfitting. Importantly,
by joint base-novel class training, our approach can be easily extended to a mor
e practical yet challenging FSL setting, i.e., generalized FSL, where the label
space of test data is extended to both base and novel classes. Extensive experim
ents show that our approach is effective for both of the two FSL settings.
****

Better to Follow, Follow to Be Better: Towards Precise Supervision of Feature Su
per-Resolution for Small Object Detection
Junhyug Noh, Wonho Bae, Wonhee Lee, Jinhwan Seo, Gunhee Kim; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9725-
9734
In spite of recent success of proposal-based CNN models for object detection, it
 is still difficult to detect small objects due to the limited and distorted inf
ormation that small region of interests (RoI) contain. One way to alleviate this
 issue is to enhance the features of small RoIs using a super-resolution (SR) te
chnique. We investigate how to improve feature-level super-resolution especially
 for small object detection, and discover its performance can be significantly i
mproved by (i) utilizing proper high-resolution target features as supervision s
ignals for training of a SR model and (ii) matching the relative receptive field
s of training pairs of input low-resolution features and target high-resolution
features. We propose a novel feature-level super-resolution approach that not on
ly correctly addresses these two desiderata but also is integrable with any prop
osal-based detectors with feature pooling. In our experiments, our approach sign

ificantly improves the performance of Faster R-CNN on three benchmarks of Tsingh ua-Tencent 100K, PASCAL VOC and MS COCO. The improvement for small objects is re markably large, and encouragingly, those for medium and large objects are nontri vial too. As a result, we achieve new state-of-the-art performance on Tsinghua-T encent 100K and highly competitive results on both PASCAL VOC and MS COCO.
****

Weakly Supervised Object Detection With Segmentation Collaboration
Xiaoyan Li, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9735-9744
Weakly supervised object detection aims at learning precise object detectors, gi ven image category labels. In recent prevailing works, this problem is generally formulated as a multiple instance learning module guided by an image classifica tion loss. The object bounding box is assumed to be the one contributing most to the classification among all proposals. However, the region contributing most i s also likely to be a crucial part or the supporting context of an object. To ob tain a more accurate detector, in this work we propose a novel end-to-end weakly supervised detection approach, where a newly introduced generative adversarial segmentation module interacts with the conventional detection module in a collab orative loop. The collaboration mechanism takes full advantages of the complemen tary interpretations of the weakly supervised localization task, namely detectio n and segmentation tasks, forming a more comprehensive solution. Consequently, o ur method obtains more precise object bounding boxes, rather than parts or irrel evant surroundings. Expectedly, the proposed method achieves an accuracy of 53.7 % on the PASCAL VOC 2007 dataset, outperforming the state-of-the-arts and demons trating its superiority for weakly supervised object detection.
****

AutoFocus: Efficient Multi-Scale Inference
Mahyar Najibi, Bharat Singh, Larry S. Davis; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 9745-9755
This paper describes AutoFocus, an efficient multi-scale inference algorithm for deep-learning based object detectors. Instead of processing an entire image pyr amid, AutoFocus adopts a coarse to fine approach and only processes regions whic h are likely to contain small objects at finer scales. This is achieved by predi cting category agnostic segmentation maps for small objects at coarser scales, c alled FocusPixels. FocusPixels can be predicted with high recall, and in many ca ses, they only cover a small fraction of the entire image. To make efficient use of FocusPixels, an algorithm is proposed which generates compact rectangular Fo cusChips which enclose FocusPixels. The detector is only applied inside FocusChi ps, which reduces computation while processing finer scales. Different types of error can arise when detections from FocusChips of multiple scales are combined, hence techniques to correct them are proposed. AutoFocus obtains an mAP of 47.9 % (68.3% at 50% overlap) on the COCO test-dev set while processing 6.4 images pe r second on a Titan X (Pascal) GPU. This is 2.5X faster than our multi-scale bas eline detector and matches its mAP. The number of pixels processed in the pyrami d can be reduced by 5X with a 1% drop in mAP. AutoFocus obtains more than 10% mA P gain compared to RetinaNet but runs at the same speed with the same ResNet-101 backbone.
****

Leveraging Long-Range Temporal Relationships Between Proposals for Video Object Detection
Mykhailo Shvets, Wei Liu, Alexander C. Berg; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 9756-9764
Single-frame object detectors perform well on videos sometimes, even without tem poral context. However, challenges such as occlusion, motion blur, and rare pose s of objects are hard to resolve without temporal awareness. Thus, there is a st rong need to improve video object detection by considering long-range temporal d ependencies. In this paper, we present a light-weight modification to a single-f rame detector that accounts for arbitrary long dependencies in a video. It impro ves the accuracy of a single-frame detector significantly with negligible comput e overhead. The key component of our approach is a novel temporal relation modul

e, operating on object proposals, that learns the similarities between proposals from different frames and selects proposals from past and/or future to support current proposals. Our final "causal" model, without any offline post-processing steps, runs at a similar speed as a single-frame detector and achieves state-of-the-art video object detection on ImageNet VID dataset.
****

Transferable Contrastive Network for Generalized Zero-Shot Learning
Huajie Jiang, Ruiping Wang, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9765-9774
Zero-shot learning (ZSL) is a challenging problem that aims to recognize the target categories without seen data, where semantic information is leveraged to transfer knowledge from some source classes. Although ZSL has made great progress in recent years, most existing approaches are easy to overfit the sources classes in generalized zero-shot learning (GZSL) task, which indicates that they learn little knowledge about target classes. To tackle such problem, we propose a novel Transferable Contrastive Network (TCN) that explicitly transfers knowledge from the source classes to the target classes. It automatically contrasts one image with different classes to judge whether they are consistent or not. By exploiting the class similarities to make knowledge transfer from source images to similar target classes, our approach is more robust to recognize the target images. Experiments on five benchmark datasets show the superiority of our approach for GZSL.
****

Fast Point R-CNN
Yilun Chen, Shu Liu, Xiaoyong Shen, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9775-9784
We present a unified, efficient and effective framework for point-cloud based 3D object detection. Our two-stage approach utilizes both voxel representation and raw point cloud data to exploit respective advantages. The first stage network, with voxel representation as input, only consists of light convolutional operations, producing a small number of high-quality initial predictions. Coordinate and indexed convolutional feature of each point in initial prediction are effectively fused with the attention mechanism, preserving both accurate localization and context information. The second stage works on interior points with their fused feature for further refining the prediction. Our method is evaluated on KITTI dataset, in terms of both 3D and Bird's Eye View (BEV) detection, and achieves state-of-the-arts with a 15FPS detection rate.
****

Mesh R-CNN
Georgia Gkioxari, Jitendra Malik, Justin Johnson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9785-9795
Rapid advances in 2D perception have led to systems that accurately detect objects in real-world images. However, these systems make predictions in 2D, ignoring the 3D structure of the world. Concurrently, advances in 3D shape prediction have mostly focused on synthetic benchmarks and isolated objects. We unify advances in these two areas. We propose a system that detects objects in real-world images and produces a triangle mesh giving the full 3D shape of each detected object. Our system, called Mesh R-CNN, augments Mask R-CNN with a mesh prediction branch that outputs meshes with varying topological structure by first predicting coarse voxel representations which are converted to meshes and refined with a graph convolution network operating over the mesh's vertices and edges. We validate our mesh prediction branch on ShapeNet, where we outperform prior work on single-image shape prediction. We then deploy our full Mesh R-CNN system on Pix3D, where we jointly detect objects and predict their 3D shapes.
****

Deep Supervised Hashing With Anchor Graph
Yudong Chen, Zhihui Lai, Yujuan Ding, Kaiyi Lin, Wai Keung Wong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9796-9804
Recently, a series of deep supervised hashing methods were proposed for binary c

ode learning. However, due to the high computation cost and the limited hardware 's memory, these methods will first select a subset from the training set, and t hen form a mini-batch data to update the network in each iteration. Therefore, t he remaining labeled data cannot be fully utilized and the model cannot directly obtain the binary codes of the entire training set for retrieval. To address th ese problems, this paper proposes an interesting regularized deep model to seaml essly integrate the advantages of deep hashing and efficient binary code learnin g by using the anchor graph. As such, the deep features and label matrix can be jointly used to optimize the binary codes, and the network can obtain more discr iminative feedback from the linear combinations of the learned bits. Moreover, w e also reveal the algorithm mechanism and its computation essence. Experiments o n three large-scale datasets indicate that the proposed method achieves better r etrieval performance with less training time compared to previous deep hashing m ethods.

****

Detecting 11K Classes: Large Scale Object Detection Without Fine-Grained Boundin g Boxes

Hao Yang, Hao Wu, Hao Chen; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2019, pp. 9805-9813

Recent advances in deep learning greatly boost the performance of object detecti on. State-of-the-art methods such as Faster-RCNN, FPN and R-FCN have achieved hi gh accuracy in challenging benchmark datasets. However, these methods require fu lly annotated object bounding boxes for training, which are incredibly hard to s cale up due to the high annotation cost. Weakly-supervised methods, on the other hand, only require image-level labels for training, but the performance is far below their fully-supervised counterparts. In this paper, we propose a semi-supe rvised large scale fine-grained detection method, which only needs bounding box annotations of a smaller number of coarse-grained classes and image-level labels of large scale fine-grained classes, and can detect all classes at nearly fully -supervised accuracy. We achieve this by utilizing the correlations between coar se-grained and fine-grained classes with shared backbone, soft-attention based p roposal re-ranking, and a dual-level memory module. Experiment results show that our methods can achieve close accuracy on object detection to state-of-the-art fully-supervised methods on two large scale datasets, ImageNet and OpenImages, w ith only a small fraction of fully annotated classes.

****

Re-ID Driven Localization Refinement for Person Search

Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF International Conference on Computer Visi on (ICCV), 2019, pp. 9814-9823

Person search aims at localizing and identifying a query person from a gallery o f uncropped scene images. Different from person re-identification (re-ID), its p erformance also depends on the localization accuracy of a pedestrian detector. T he state-of-the-art methods train the detector individually, and the detected bo unding boxes may be sub-optimal for the following re-ID task. To alleviate this issue, we propose a re-ID driven localization refinement framework for providing the refined detection boxes for person search. Specifically, we develop a diffe rentiable ROI transform layer to effectively transform the bounding boxes from t he original images. Thus, the box coordinates can be supervised by the re-ID tra ining other than the original detection task. With this supervision, the detecto r can generate more reliable bounding boxes, and the downstream re-ID model can produce more discriminative embeddings based on the refined person localizations . Extensive experimental results on the widely used benchmarks demonstrate that our proposed method performs favorably against the state-of-the-art person searc h methods.

****

Hierarchical Encoding of Sequential Data With Compact and Sub-Linear Storage Cos t

Huu Le, Ming Xu, Tuan Hoang, Michael Milford; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2019, pp. 9824-9833

Snapshot-based visual localization is an important problem in several computer v
ision and robotics applications such as Simultaneous Localization And Mapping (S
LAM). To achieve real-time performance in very large-scale environments with mas
sive amounts of training and map data, techniques such as approximate nearest ne
ighbor search (ANN) algorithms are used. While several state-of-the-art variants
 of quantization and indexing techniques have demonstrated to be efficient in pr
actice, their theoretical memory cost still scales at least linearly with the tr
aining data (i.e., O(n) where n is the number of instances in the database), sin
ce each data point must be associated with at least one code vector. To address
these limitations, in this paper we present a totally new hierarchical encoding
approach that enables a sub-linear storage scale. The algorithm exploits the wid
espread sequential nature of sensor information streams in robotics and autonomo
us vehicle applications and achieves, both theoretically and experimentally, sub
-linear scalability in storage required for a given environment size. Furthermor
e, the associated query time of our algorithm is also of sub-linear complexity.
We benchmark the performance of the proposed algorithm on several real-world ben
chmark datasets and experimentally validate the theoretical sub-linearity of our
 approach, while also showing that our approach yields competitive absolute stor
age performance as well.
****

C-MIDN: Coupled Multiple Instance Detection Network With Segmentation Guidance f
or Weakly Supervised Object Detection

Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, Dongrui
Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2019, pp. 9834-9843

Weakly supervised object detection (WSOD) that only needs image-level annotation
s has obtained much attention recently. By combining convolutional neural networ
k with multiple instance learning method, Multiple Instance Detection Network (M
IDN) has become the most popular method to address the WSOD problem and been ado
pted as the initial model in many works. We argue that MIDN inclines to converge
 to the most discriminative object parts, which limits the performance of method
s based on it. In this paper, we propose a novel Coupled Multiple Instance Detec
tion Network (C-MIDN) to address this problem. Specifically, we use a pair of MI
DNs, which work in a complementary manner with proposal removal. The localizatio
n information of the MIDNs is further coupled to obtain tighter bounding boxes a
nd localize multiple objects. We also introduce a Segmentation Guided Proposal R
emoval (SGPR) algorithm to guarantee the MIL constraint after the removal and en
sure the robustness of C-MIDN. Through a simple implementation of the C-MIDN wit
h online detector refinement, we obtain 53.6% and 50.3% mAP on the challenging P
ASCAL VOC 2007 and 2012 benchmarks respectively, which significantly outperform
the previous state-of-the-arts.
****

Learning Feature-to-Feature Translator by Alternating Back-Propagation for Gener
ative Zero-Shot Learning

Yizhe Zhu, Jianwen Xie, Bingchen Liu, Ahmed Elgammal; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9844-9854

We investigate learning feature-to-feature translator networks by alternating ba
ck-propagation as a general-purpose solution to zero-shot learning (ZSL) problem
s. It is a generative model-based ZSL framework. In contrast to models based on
generative adversarial networks (GAN) or variational autoencoders (VAE) that req
uire auxiliary networks to assist the training, our model consists of a single c
onditional generator that maps class-level semantic features and Gaussian white
noise vectors accounting for instance-level latent factors to visual features, a
nd is trained by maximum likelihood estimation. The training process is a simple
 yet effective alternating back-propagation process that iterates the following
two steps: (i) the inferential back-propagation to infer the latent noise vector
 of each observed example, and (ii) the learning back-propagation to update the
model parameters. We show that, with slight modifications, our model is capable
of learning from incomplete visual features for ZSL. We conduct extensive compar
isons with existing generative ZSL methods on five benchmarks, demonstrating the

superiority of our method in not only ZSL performance but also convergence speed and computational cost. Specifically, our model outperforms the existing state-of-the-art methods by a remarkable margin up to 3.1% and 4.0% in ZSL and generalized ZSL settings, respectively.

****

Deep Constrained Dominant Sets for Person Re-Identification

Leulseged Tesfaye Alemu, Marcello Pelillo, Mubarak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9855-9864

In this work, we propose an end-to-end constrained clustering scheme to tackle the person re-identification (re-id) problem. Deep neural networks (DNN) have recently proven to be effective on person re-identification task. In particular, rather than leveraging solely a probe-gallery similarity, diffusing the similarities among the gallery images in an end-to-end manner has proven to be effective in yielding a robust probe-gallery affinity. However, existing methods do not apply probe image as a constraint, and are prone to noise propagation during the similarity diffusion process. To overcome this, we propose an intriguing scheme which treats person-image retrieval problem as a constrained clustering optimization problem, called deep constrained dominant sets (DCDS). Given a probe and gallery images, we re-formulate person re-id problem as finding a constrained cluster, where the probe image is taken as a constraint (seed) and each cluster corresponds to a set of images corresponding to the same person. By optimizing the constrained clustering in an end-to-end manner, we naturally leverage the contextual knowledge of a set of images corresponding to the given person-images. We further enhance the performance by integrating an auxiliary net alongside DCDS, which employs a multi-scale ResNet. To validate the effectiveness of our method we present experiments on several benchmark datasets and show that the proposed method can outperform state-of-the-art methods.

****

Invariant Information Clustering for Unsupervised Image Classification and Segmentation

Xu Ji, Joao F. Henriques, Andrea Vedaldi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9865-9874

We present a novel clustering objective that learns a neural network classifier from scratch, given only unlabelled data samples. The model discovers clusters that accurately match semantic classes, achieving state-of-the-art results in eight unsupervised clustering benchmarks spanning image classification and segmentation. These include STL10, an unsupervised variant of ImageNet, and CIFAR10, where we significantly beat the accuracy of our closest competitors by 6.6 and 9.5 absolute percentage points respectively. The method is not specialised to computer vision and operates on any paired dataset samples; in our experiments we use random transforms to obtain a pair from each image. The trained network directly outputs semantic labels, rather than high dimensional representations that need external processing to be usable for semantic clustering. The objective is simply to maximise mutual information between the class assignments of each pair. It is easy to implement and rigorously grounded in information theory, meaning we effortlessly avoid degenerate solutions that other clustering methods are susceptible to. In addition to the fully unsupervised mode, we also test two semi-supervised settings. The first achieves 88.8% accuracy on STL10 classification, setting a new global state-of-the-art over all existing methods (whether supervised, semi-supervised or unsupervised). The second shows robustness to 90% reductions in label coverage, of relevance to applications that wish to make use of small amounts of labels. github.com/xu-ji/IIC

****

Subspace Structure-Aware Spectral Clustering for Robust Subspace Clustering

Masataka Yamaguchi, Go Irie, Takahito Kawanishi, Kunio Kashino; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9875-9884

Subspace clustering is the problem of partitioning data drawn from a union of multiple subspaces. The most popular subspace clustering framework in recent years is the graph clustering-based approach, which performs subspace clustering in t

wo steps: graph construction and graph clustering. Although both steps are equal ly important for accurate clustering, the vast majority of work has focused on i mproving the graph construction step rather than the graph clustering step. In t his paper, we propose a novel graph clustering framework for robust subspace clu stering. By incorporating a geometry-aware term with the spectral clustering obj ective, we encourage our framework to be robust to noise and outliers in given a ffinity matrices. We also develop an efficient expectation-maximization-based al gorithm for optimization. Through extensive experiments on four real-world datas ets, we demonstrate that the proposed method outperforms existing methods.
****

Order-Preserving Wasserstein Discriminant Analysis
Bing Su, Jiahuan Zhou, Ying Wu; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2019, pp. 9885-9894
Supervised dimensionality reduction for sequence data projects the observations in sequences onto a low-dimensional subspace to better separate different sequen ce classes. It is typically more challenging than conventional dimensionality re duction for static data, because measuring the separability of sequences involve s non-linear procedures to manipulate the temporal structures. This paper presen ts a linear method, namely Order-preserving Wasserstein Discriminant Analysis (O WDA), which learns the projection by maximizing the inter-class distance and min imizing the intra-class scatter. For each class, OWDA extracts the order-preserv ing Wasserstein barycenter and constructs the intra-class scatter as the dispers ion of the training sequences around the barycenter. The inter-class distance is measured as the order-preserving Wasserstein distance between the corresponding barycenters. OWDA is able to concentrate on the distinctive differences among c lasses by lifting the geometric relations with temporal constraints. Experiments show that OWDA achieves competitive results on three 3D action recognition data sets.
****

LayoutVAE: Stochastic Scene Layout Generation From a Label Set
Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, Greg Mori; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9895-9904
Recently there is an increasing interest in scene generation within the research community. However, models used for generating scene layouts from textual descr iption largely ignore plausible visual variations within the structure dictated by the text. We propose LayoutVAE, a variational autoencoder based framework for generating stochastic scene layouts. LayoutVAE is a versatile modeling framewor k that allows for generating full image layouts given a label set, or per label layouts for an existing image given a new label. In addition, it is also capable of detecting unusual layouts, potentially providing a way to evaluate layout ge neration problem. Extensive experiments on MNIST-Layouts and challenging COCO 20 17 Panoptic dataset verifies the effectiveness of our proposed framework.
****

Robust Variational Bayesian Point Set Registration
Jie Zhou, Xinke Ma, Li Liang, Yang Yang, Shijin Xu, Yuhe Liu, Sim-Heng Ong ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2019, pp. 9905-9914
In this work, we propose a hierarchical Bayesian network based point set registr ation method to solve missing correspondences and various massive outliers. We c onstruct this network first using the finite Student s t latent mixture model (T LMM), in which distributions of latent variables are estimated by a tree-structu red variational inference (VI) so that to obtain a tighter lower bound under the Bayesian framework. We then divide the TLMM into two different mixtures with is otropic and anisotropic covariances for correspondences recovering and outliers identification, respectively. Finally, the parameters of mixing proportion and c ovariances are both taken as latent variables, which benefits explaining of miss ing correspondences and heteroscedastic outliers. In addition, a cooling schedul e is adopted to anneal prior on covariances and scale variables within designed two phases of transformation, it anneal priors on global and local variables to

perform a coarse-to- fine registration. In experiments, our method outperforms f ive state-of-the-art methods in synthetic point set and realistic imaging regist rations.
****

Is an Affine Constraint Needed for Affine Subspace Clustering?
Chong You, Chun-Guang Li, Daniel P. Robinson, Rene Vidal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9915-9924
Subspace clustering methods based on expressing each data point as a linear comb ination of other data points have achieved great success in computer vision appl ications such as motion segmentation, face and digit clustering. In face cluster ing, the subspaces are linear and subspace clustering methods can be applied dir ectly. In motion segmentation, the subspaces are affine and an additional affine constraint on the coefficients is often enforced. However, since affine subspac es can always be embedded into linear subspaces of one extra dimension, it is un clear if the affine constraint is really necessary. This paper shows, both theor etically and empirically, that when the dimension of the ambient space is high r elative to the sum of the dimensions of the affine subspaces, the affine constra int has a negligible effect on clustering performance. Specifically, our analysi s provides conditions that guarantee the correctness of affine subspace clusteri ng methods both with and without the affine constraint, and shows that these con ditions are satisfied for high-dimensional data. Underlying our analysis is the notion of affinely independent subspaces, which not only provides geometrically interpretable correctness conditions, but also clarifies the relationships betwe en existing results for affine subspace clustering.
****

Meta-Learning to Detect Rare Objects
Yu-Xiong Wang, Deva Ramanan, Martial Hebert; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 9925-9934
Few-shot learning, i.e., learning novel concepts from few examples, is fundament al to practical visual recognition systems. While most of existing work has focu sed on few-shot classification, we make a step towards few-shot object detection , a more challenging yet under-explored task. We develop a conceptually simple b ut powerful meta-learning based framework that simultaneously tackles few-shot c lassification and few-shot localization in a unified, coherent way. This framewo rk leverages meta-level knowledge about "model parameter generation" from base c lasses with abundant data to facilitate the generation of a detector for novel c lasses. Our key insight is to disentangle the learning of category-agnostic and category-specific components in a CNN based detection model. In particular, we i ntroduce a weight prediction meta-model that enables predicting the parameters o f category-specific components from few examples. We systematically benchmark th e performance of modern detectors in the small-sample size regime. Experiments i n a variety of realistic scenarios, including within-domain, cross-domain, and l ong-tailed settings, demonstrate the effectiveness and generality of our approac h under different notions of novel classes.
****

New Convex Relaxations for MRF Inference With Unknown Graphs
Zhenhua Wang, Tong Liu, Qinfeng Shi, M. Pawan Kumar, Jianhua Zhang; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp . 9935-9943
Treating graph structures of Markov random fields as unknown and estimating them jointly with labels have been shown to be useful for modeling human activity re cognition and other related tasks. We propose two novel relaxations for solving this problem. The first is a linear programming (LP) relaxation, which is provab ly tighter than the existing LP relaxation. The second is a non-convex quadratic programming (QP) relaxation, which admits an efficient concave-convex procedure (CCCP). The CCCP algorithm is initialized by solving a convex QP relaxation of the problem, which is obtained by modifying the diagonal of the matrix that spec ifies the non-convex QP relaxation. We show that our convex QP relaxation is opt imal in the sense that it minimizes the L1 norm of the diagonal modification vec tor. While the convex QP relaxation is not as tight as the existing and the new

LP relaxations, when used in conjunction with the CCCP algorithm for the non-con
vex QP relaxation, it provides accurate solutions. We demonstrate the efficacy o
f our new relaxations for both synthetic data and human activity recognition.
****

Cluster Alignment With a Teacher for Unsupervised Domain Adaptation
Zhijie Deng, Yucen Luo, Jun Zhu; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2019, pp. 9944-9953
Deep learning methods have shown promise in unsupervised domain adaptation, whic
h aims to leverage a labeled source domain to learn a classifier for the unlabel
ed target domain with a different distribution. However, such methods typically
learn a domain-invariant representation space to match the marginal distribution
s of the source and target domains, while ignoring their fine-level structures.
In this paper, we propose Cluster Alignment with a Teacher (CAT) for unsupervise
d domain adaptation, which can effectively incorporate the discriminative cluste
ring structures in both domains for better adaptation. Technically, CAT leverage
s an implicit ensembling teacher model to reliably discover the class-conditiona
l structure in the feature space for the unlabeled target domain. Then CAT force
s the features of both the source and the target domains to form discriminative
class-conditional clusters and aligns the corresponding clusters across domains.
Empirical results demonstrate that CAT achieves state-of-the-art results in sev
eral unsupervised domain adaptation scenarios.
****

Analyzing the Variety Loss in the Context of Probabilistic Trajectory Prediction
Luca Anthony Thiede, Pratik Prabhanjan Brahma; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2019, pp. 9954-9963
Trajectory or behavior prediction of traffic agents is an important component of
autonomous driving and robot planning in general. It can be framed as a probabi
listic future sequence generation problem and recent literature has studied the
applicability of generative models in this context. The variety or Minimum over
N (MoN) loss, which tries to minimize the error between the ground truth and the
closest of N output predictions, has been used in these recent learning models
to improve the diversity of predictions. In this work, we present a proof to sho
w that the MoN loss does not lead to the ground truth probability density functi
on, but approximately to its square root instead. We validate this finding with
extensive experiments on both simulated toy as well as real world datasets. We a
lso propose multiple solutions to compensate for the dilation to show improvemen
t of log likelihood of the ground truth samples in the corrected probability den
sity function.
****

Deep Mesh Reconstruction From Single RGB Images via Topology Modification Networ
ks
Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, Kui Jia; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9964
-9973
Reconstructing the 3D mesh of a general object from a single image is now possib
le thanks to the latest advances of deep learning technologies. However, due to
the nontrivial difficulty of generating a feasible mesh structure, the state-of-
the-art approaches often simplify the problem by learning the displacements of a
template mesh that deforms it to the target surface. Though reconstructing a 3D
shape with complex topology can be achieved by deforming multiple mesh patches,
it remains difficult to stitch the results to ensure a high meshing quality. In
this paper, we present an end-to-end single-view mesh reconstruction framework
that is able to generate high-quality meshes with complex topologies from a sing
le genus-0 template mesh. The key to our approach is a novel progressive shaping
framework that alternates between mesh deformation and topology modification. W
hile a deformation network predicts the per-vertex translations that reduce the
gap between the reconstructed mesh and the ground truth, a novel topology modifi
cation network is employed to prune the error-prone faces, enabling the evolutio
n of topology. By iterating over the two procedures, one can progressively modif
y the mesh topology while achieving higher reconstruction accuracy. Moreover, a

boundary refinement network is designed to refine the boundary conditions to further improve the visual quality of the reconstructed mesh. Extensive experiments demonstrate that our approach outperforms the current state-of-the-art methods both qualitatively and quantitatively, especially for the shapes with complex topologies.
****

UprightNet: Geometry-Aware Camera Orientation Estimation From Single Images
Wenqi Xian, Zhengqi Li, Matthew Fisher, Jonathan Eisenmann, Eli Shechtman, Noah Snavely; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9974-9983
We introduce UprightNet, a learning-based approach for estimating 2DoF camera orientation from a single RGB image of an indoor scene. Unlike recent methods that leverage deep learning to perform black-box regression from image to orientation parameters, we propose an end-to-end framework that incorporates explicit geometric reasoning. In particular, we design a network that predicts two representations of scene geometry, in both the local camera and global reference coordinate systems, and solves for the camera orientation as the rotation that best aligns these two predictions via a differentiable least squares module. This network can be trained end-to-end, and can be supervised with both ground truth camera poses and intermediate representations of surface geometry. We evaluate UprightNet on the single-image camera orientation task on synthetic and real datasets, and show significant improvements over prior state-of-the-art approaches.
****

Escaping Plato's Cave: 3D Shape From Adversarial Rendering
Philipp Henzler, Niloy J. Mitra, Tobias Ritschel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9984-9993
We introduce PlatonicGAN to discover the 3D structure of an object class from an unstructured collection of 2D images, i.e., where no relation between photos is known, except that they are showing instances of the same category. The key idea is to train a deep neural network to generate 3D shapes which, when rendered to images, are indistinguishable from ground truth images (for a discriminator) under various camera poses. Discriminating 2D images instead of 3D shapes allows tapping into unstructured 2D photo collections instead of relying on curated (e.g., aligned, annotated, etc.) 3D data sets. To establish constraints between 2D image observation and their 3D interpretation, we suggest a family of rendering layers that are effectively differentiable. This family includes visual hull, absorption-only (akin to x-ray), and emission-absorption. We can successfully reconstruct 3D shapes from unstructured 2D images and extensively evaluate PlatonicGAN on a range of synthetic and real data sets achieving consistent improvements over baseline methods. We further show that PlatonicGAN can be combined with 3D supervision to improve on and in some cases even surpass the quality of 3D-supervised methods.
****

Deep End-to-End Alignment and Refinement for Time-of-Flight RGB-D Module
Di Qiu, Jiahao Pang, Wenxiu Sun, Chengxi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9994-10003
Recently, it is increasingly popular to equip mobile RGB cameras with Time-of-Flight (ToF) sensors for active depth sensing. However, for off-the-shelf ToF sensors, one must tackle two problems in order to obtain high-quality depth with respect to the RGB camera, namely 1) online calibration and alignment; and 2) complicated error correction for ToF depth sensing. In this work, we propose a framework for jointly alignment and refinement via deep learning. First, a cross-modal optical flow between the RGB image and the ToF amplitude image is estimated for alignment. The aligned depth is then refined via an improved kernel predicting network that performs kernel normalization and applies the bias prior to the dynamic convolution. To enrich our data for end-to-end training, we have also synthesized a dataset using tools from computer graphics. Experimental results demonstrate the effectiveness of our approach, achieving state-of-the-art for ToF refinement.
****

GEOBIT: A Geodesic-Based Binary Descriptor Invariant to Non-Rigid Deformations for RGB-D Images
Erickson R. Nascimento, Guilherme Potje, Renato Martins, Felipe Cadar, Mario F. M. Campos, Ruzena Bajcsy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10004-10012
At the core of most three-dimensional alignment and tracking tasks resides the critical problem of point correspondence. In this context, the design of descriptors that efficiently and uniquely identifies keypoints, to be matched, is of central importance. Numerous descriptors have been developed for dealing with affine/perspective warps, but few can also handle non-rigid deformations. In this paper, we introduce a novel binary RGB-D descriptor invariant to isometric deformations. Our method uses geodesic isocurves on smooth textured manifolds. It combines appearance and geometric information from RGB-D images to tackle non-rigid transformations. We used our descriptor to track multiple textured depth maps and demonstrate that it produces reliable feature descriptors even in the presence of strong non-rigid deformations and depth noise. The experiments show that our descriptor outperforms different state-of-the-art descriptors in both precision-recall and recognition rate metrics. We also provide to the community a new dataset composed of annotated RGB-D images of different objects (shirts, cloths, paintings, bags), subjected to strong non-rigid deformations, to evaluate point correspondence algorithms.
****
CDTB: A Color and Depth Visual Object Tracking Dataset and Benchmark
Alan Lukezic, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, Matej Kristan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10013-10022
We propose a new color-and-depth general visual object tracking benchmark (CDTB). CDTB is recorded by several passive and active RGB-D setups and contains indoor as well as outdoor sequences acquired in direct sunlight. The CDTB dataset is the largest and most diverse dataset in RGB-D tracking, with an order of magnitude larger number of frames than related datasets. The sequences have been carefully recorded to contain significant object pose change, clutter, occlusion, and periods of long-term target absence to enable tracker evaluation under realistic conditions. Sequences are per-frame annotated with 13 visual attributes for detailed analysis. Experiments with RGB and RGB-D trackers show that CDTB is more challenging than previous datasets. State-of-the-art RGB trackers outperform the recent RGB-D trackers, indicating a large gap between the two fields, which has not been previously detected by the prior benchmarks. Based on the results of the analysis we point out opportunities for future research in RGB-D tracker design.
****
Learning Joint 2D-3D Representations for Depth Completion
Yun Chen, Bin Yang, Ming Liang, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10023-10032
In this paper, we tackle the problem of depth completion from RGBD data. Towards this goal, we design a simple yet effective neural network block that learns to extract joint 2D and 3D features. Specifically, the block consists of two domain-specific sub-networks that apply 2D convolution on image pixels and continuous convolution on 3D points, with their output features fused in image space. We build the depth completion network simply by stacking the proposed block, which has the advantage of learning hierarchical representations that are fully fused between 2D and 3D spaces at multiple levels. We demonstrate the effectiveness of our approach on the challenging KITTI depth completion benchmark and show that our approach outperforms the state-of-the-art.
****
Make a Face: Towards Arbitrary High Fidelity Face Manipulation
Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, Ran He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10033-10042
Recent studies have shown remarkable success in face manipulation task with the

advance of GANs and VAEs paradigms, but the outputs are sometimes limited to low-resolution and lack of diversity. In this work, we propose Additive Focal Variational Auto-encoder (AF-VAE), a novel approach that can arbitrarily manipulate high-resolution face images using a simple yet effective model and only weak supervision of reconstruction and KL divergence losses. First, a novel additive Gaussian Mixture assumption is introduced with an unsupervised clustering mechanism in the structural latent space, which endows better disentanglement and boosts multi-modal representation with external memory. Second, to improve the perceptual quality of synthesized results, two simple strategies in architecture design are further tailored and discussed on the behavior of Human Visual System (HVS) for the first time, allowing for fine control over the model complexity and sample quality. Human opinion studies and new state-of-the-art Inception Score (IS) / Frechet Inception Distance (FID) demonstrate the superiority of our approach over existing algorithms, advancing both the fidelity and extremity of face manipulation task.

****

M2FPA: A Multi-Yaw Multi-Pitch High-Quality Dataset and Benchmark for Facial Pose Analysis

Peipei Li, Xiang Wu, Yibo Hu, Ran He, Zhenan Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10043-10051

Facial images in surveillance or mobile scenarios often have large view-point variations in terms of pitch and yaw angles. These jointly occurred angle variations make face recognition challenging. Current public face databases mainly consider the case of yaw variations. In this paper, a new large-scale Multi-yaw Multi-pitch high-quality database is proposed for Facial Pose Analysis (M2FPA), including face frontalization, face rotation, facial pose estimation and pose-invariant face recognition. It contains 397,544 images of 229 subjects with yaw, pitch, attribute, illumination and accessory. M2FPA is the most comprehensive multi-view face database for facial pose analysis. Further, we provide an effective benchmark for face frontalization and pose-invariant face recognition on M2FPA with several state-of-the-art methods, including DR-GAN, TP-GAN and CAPG-GAN. We believe that the new database and benchmark can significantly push forward the advance of facial pose analysis in real-world applications. Moreover, a simple yet effective parsing guided discriminator is introduced to capture the local consistency during GAN optimization. Extensive quantitative and qualitative results on M2FPA and Multi-PIE demonstrate the superiority of our face frontalization method. Baseline results for both face synthesis and face recognition from state-of-the-art methods demonstrate the challenge offered by this new database.

****

Fair Loss: Margin-Aware Reinforcement Learning for Deep Face Recognition

Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, Yaohai Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10052-10061

Recently, large-margin softmax loss methods, such as angular softmax loss (SphereFace), large margin cosine loss (CosFace), and additive angular margin loss (ArcFace), have demonstrated impressive performance on deep face recognition. These methods incorporate a fixed additive margin to all the classes, ignoring the class imbalance problem. However, imbalanced problem widely exists in various real-world face datasets, in which samples from some classes are in a higher number than others. We argue that the number of a class would influence its demand for the additive margin. In this paper, we introduce a new margin-aware reinforcement learning based loss function, namely fair loss, in which each class will learn an appropriate adaptive margin by Deep Q-learning. Specifically, we train an agent to learn a margin adaptive strategy for each class, and make the additive margins for different classes more reasonable. Our method has better performance than present large-margin loss functions on three benchmarks, Labeled Face in the Wild (LFW), Youtube Faces (YTF) and MegaFace, which demonstrates that our method could learn better face representation on imbalanced face datasets.

****

Face De-Occlusion Using 3D Morphable Model and Generative Adversarial Network

Xiaowei Yuan, In Kyu Park; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10062-10071
In recent decades, 3D morphable model (3DMM) has been commonly used in image-based photorealistic 3D face reconstruction. However, face images are often corrupted by serious occlusion by non-face objects including eyeglasses, masks, and hands. Such objects block the correct capture of landmarks and shading information. Therefore, the reconstructed 3D face model is hardly reusable. In this paper, a novel method is proposed to restore de-occluded face images based on inverse use of 3DMM and generative adversarial network. We utilize the 3DMM prior to the proposed adversarial network and combine a global and local adversarial convolutional neural network to learn face de-occlusion model. The 3DMM serves not only as geometric prior but also proposes the face region for the local discriminator. Experiment results confirm the effectiveness and robustness of the proposed algorithm in removing challenging types of occlusions with various head poses and illumination. Furthermore, the proposed method reconstructs the correct 3D face model with de-occluded textures.
****

Detecting Photoshopped Faces by Scripting Photoshop
Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, Alexei A. Efros; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10072-10081
Most malicious photo manipulations are created using standard image editing tools, such as Adobe Photoshop. We present a method for detecting one very popular Photoshop manipulation -- image warping applied to human faces -- using a model trained entirely using fake images that were automatically generated by scripting Photoshop itself. We show that our model outperforms humans at the task of recognizing manipulated images, can predict the specific location of edits, and in some cases can be used to "undo" a manipulation to reconstruct the original, unedited image. We demonstrate that the system can be successfully applied to artist-created image manipulations.
****

Ego-Pose Estimation and Forecasting As Real-Time PD Control
Ye Yuan, Kris Kitani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10082-10092
We propose the use of a proportional-derivative (PD) control based policy learned via reinforcement learning (RL) to estimate and forecast 3D human pose from egocentric videos. The method learns directly from unsegmented egocentric videos and motion capture data consisting of various complex human motions (e.g., crouching, hopping, bending, and motion transitions). We propose a video-conditioned recurrent control technique to forecast physically-valid and stable future motions of arbitrary length. We also introduce a value function based fail-safe mechanism which enables our method to run as a single pass algorithm over the video data. Experiments with both controlled and in-the-wild data show that our approach outperforms previous art in both quantitative metrics and visual quality of the motions, and is also robust enough to transfer directly to real-world scenarios. Additionally, our time analysis shows that the combined use of our pose estimation and forecasting can run at 30 FPS, making it suitable for real-time applications.
****

End-to-End Learning for Graph Decomposition
Jie Song, Bjoern Andres, Michael J. Black, Otmar Hilliges, Siyu Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10093-10102
Deep neural networks provide powerful tools for pattern recognition, while classical graph algorithms are widely used to solve combinatorial problems. In computer vision, many tasks combine elements of both pattern recognition and graph reasoning. In this paper, we study how to connect deep networks with graph decomposition into an end-to-end trainable framework. More specifically, the minimum cost multicut problem is first converted to an unconstrained binary cubic formulation where cycle consistency constraints are incorporated into the objective funct

ion. The new optimization problem can be viewed as a Conditional Random Field (CRF) in which the random variables are associated with the binary edge labels. Cycle constraints are introduced into the CRF as high-order potentials. A standard Convolutional Neural Network (CNN) provides the front-end features for the fully differentiable CRF. The parameters of both parts are optimized in an end-to-end manner. The efficacy of the proposed learning algorithm is demonstrated via experiments on clustering MNIST images and on the challenging task of real-world multi-people pose estimation.
****

Laplace Landmark Localization
Joseph P. Robinson, Yuncheng Li, Ning Zhang, Yun Fu, Sergey Tulyakov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10103-10112
Landmark localization in images and videos is a classic problem solved in various ways. Nowadays, with deep networks prevailing throughout machine learning, there are revamped interests in pushing facial landmark detectors to handle more challenging data. Most efforts use network objectives based on L1 or L2 norms, which have several disadvantages. First of all, the generated heatmaps translate to the locations of landmarks (i.e. confidence maps) from which predicted landmark locations (i.e. the means) get penalized without accounting for the spread: a high- scatter corresponds to low confidence and vice-versa. For this, we introduce a LaplaceKL objective that penalizes for low confidence. Another issue is a dependency on labeled data, which are expensive to obtain and susceptible to error. To address both issues, we propose an adversarial training framework that leverages unlabeled data to improve model performance. Our method claims state-of-the-art on all of the 300W benchmarks and ranks second-to-best on the Annotated Facial Landmarks in the Wild (AFLW) dataset. Furthermore, our model is robust with a reduced size: 1/8 the number of channels (i.e. 0.0398 MB) is comparable to the state-of-the-art in real-time on CPU. Thus, this work is of high practical value to real-life application.
****

Through-Wall Human Mesh Recovery Using Radio Signals
Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Tianhong Li, Hang Zhao, Antonio Torralba, Dina Katabi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10113-10122
This paper presents RF-Avatar, a neural network model that can estimate 3D meshes of the human body in the presence of occlusions, baggy clothes, and bad lighting conditions. We leverage that radio frequency (RF) signals in the WiFi range traverse clothes and occlusions and bounce off the human body. Our model parses such radio signals and recovers 3D body meshes. Our meshes are dynamic and smoothly track the movements of the corresponding people. Further, our model works both in single and multi-person scenarios. Inferring body meshes from radio signals is a highly under-constrained problem. Our model deals with this challenge using: 1) a combination of strong and weak supervision, 2) a multi-headed self-attention mechanism that attends differently to temporal information in the radio signal, and 3) an adversarially trained temporal discriminator that imposes a prior on the dynamics of human motion. Our results show that RF-Avatar accurately recovers dynamic 3D meshes in the presence of occlusions, baggy clothes, bad lighting conditions, and even through walls.
****

Discriminatively Learned Convex Models for Set Based Face Recognition
Hakan Cevikalp, Golara Ghorban Dordinejad; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10123-10132
Majority of the image set based face recognition methods use a generatively learned model for each person that is learned independently by ignoring the other persons in the gallery set. In contrast to these methods, this paper introduces a novel method that searches for discriminative convex models that best fit to an individual's face images but at the same time are as far as possible from the images of other persons in the gallery. We learn discriminative convex models for both affine and convex hulls of image sets. During testing, distances from the q

uery set images to these models are computed efficiently by using simple matrix multiplications, and the query set is assigned to the person in the gallery whose image set is closest to the query images. The proposed method significantly outperforms other methods using generative convex models in terms of both accuracy and testing time, and achieves the state-of-the-art results on four of the five tested datasets. Especially, the accuracy improvement is significant on the challenging PaSC, COX and ESOGU video datasets.
****

Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image
Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10133-10142
Although significant improvement has been achieved recently in 3D human pose estimation, most of the previous methods only treat a single-person case. In this work, we firstly propose a fully learning-based, camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. The pipeline of the proposed system consists of human detection, absolute 3D human root localization, and root-relative 3D single-person pose estimation modules. Our system achieves comparable results with the state-of-the-art 3D single-person pose estimation models without any groundtruth information and significantly outperforms previous 3D multi-person pose estimation methods on publicly available datasets. The code is available in (https://github.com/mks0601/3DMPPE_ROOTNET_RELEASE), (https://github.com/mks0601/3DMPPE_POSENET_RELEASE).
****

Context-Aware Emotion Recognition Networks
Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, Kwanghoon Sohn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10143-10152
Traditional techniques for emotion recognition have focused on the facial expression analysis only, thus providing limited ability to encode context that comprehensively represents the emotional responses. We present deep networks for context-aware emotion recognition, called CAER-Net, that exploit not only human facial expression but also context information in a joint and boosting manner. The key idea is to hide human faces in a visual scene and seek other contexts based on an attention mechanism. Our networks consist of two sub-networks, including two-stream encoding networks to separately extract the features of face and context regions, and adaptive fusion networks to fuse such features in an adaptive fashion. We also introduce a novel benchmark for context-aware emotion recognition, called CAER, that is appropriate than existing benchmarks both qualitatively and quantitatively. On several benchmarks, CAER-Net proves the effect of context for emotion recognition. Our dataset is available at http://caer-dataset.github.io.
****

Deep Head Pose Estimation Using Synthetic Images and Partial Adversarial Domain Adaption for Continuous Label Spaces
Felix Kuhnke, Jorn Ostermann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10164-10173
Head pose estimation aims at predicting an accurate pose from an image. Current approaches rely on supervised deep learning, which typically requires large amounts of labeled data. Manual or sensor-based annotations of head poses are prone to errors. A solution is to generate synthetic training data by rendering 3D face models. However, the differences (domain gap) between rendered (source-domain) and real-world (target-domain) images can cause low performance. Advances in visual domain adaptation allow reducing the influence of domain differences using adversarial neural networks, which match the feature spaces between domains by enforcing domain-invariant features. While previous work on visual domain adaptation generally assumes discrete and shared label spaces, these assumptions are both invalid for pose estimation tasks. We are the first to present domain adaptation for head pose estimation with a focus on partially shared and continuous label spaces. More precisely, we adapt the predominant weighting approaches to cont

inuous label spaces by applying a weighted resampling of the source domain durin g training. To evaluate our approach, we revise and extend existing datasets res ulting in a new benchmark for visual domain adaption. Our experiments show that our method improves the accuracy of head pose estimation for real-world images d espite using only labels from synthetic images.
****

Flare in Interference-Based Hyperspectral Cameras
Eden Sassoon, Yoav Y. Schechner, Tali Treibitz; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2019, pp. 10174-10182
Stray light (flare) is formed inside cameras by internal reflections between opt ical elements. We point out a flare effect of significant magnitude and implicat ion to snapshot hyperspectral imagers. Recent technologies enable placing interf erence-based filters on individual pixels in imaging sensors. These filters have narrow transmission bands around custom wavelengths and high transmission effic iency. Cameras using arrays of such filters are compact, robust and fast. Howeve r, as opposed to traditional broad-band filters, which often absorb unwanted lig ht, narrow band-pass interference filters reflect non-transmitted light. This is a source of very significant flare which biases hyperspectral measurements. The bias in any pixel depends on spectral content in other pixels. We present a the oretical image formation model for this effect, and quantify it through simulati ons and experiments. In addition, we test deflaring of signals affected by such flare.
****

Computational Hyperspectral Imaging Based on Dimension-Discriminative Low-Rank T ensor Recovery
Shipeng Zhang, Lizhi Wang, Ying Fu, Xiaoming Zhong, Hua Huang; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 101 83-10192
Exploiting the prior information is fundamental for the image reconstruction in computational hyperspectral imaging. Existing methods usually unfold the 3D sign al as a 1D vector and treat the prior information within different dimensions in an indiscriminative manner, which ignores the high-dimensionality nature of hyp erspectral image (HSI) and thus results in poor quality reconstruction. In this paper, we propose to make full use of the high-dimensionality structure of the d esired HSI to boost the reconstruction quality. We first build a high-order tens or by exploiting the nonlocal similarity in HSI. Then, we propose a dimension-di scriminative low-rank tensor recovery (DLTR) model to characterize the structure prior adaptively in each dimension. By integrating the structure prior in DLTR with the system imaging process, we develop an optimization framework for HSI re construction, which is finally solved via the alternating minimization algorithm . Extensive experiments implemented with both synthetic and real data demonstrat e that our method outperforms state-of-the-art methods.
****

Deep Optics for Monocular Depth Estimation and 3D Object Detection
Julie Chang, Gordon Wetzstein; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2019, pp. 10193-10202
Depth estimation and 3D object detection are critical for scene understanding bu t remain challenging to perform with a single image due to the loss of 3D inform ation during image capture. Recent models using deep neural networks have improv ed monocular depth estimation performance, but there is still difficulty in pred icting absolute depth and generalizing outside a standard dataset. Here we intro duce the paradigm of deep optics, i.e. end-to-end design of optics and image pro cessing, to the monocular depth estimation problem, using coded defocus blur as an additional depth cue to be decoded by a neural network. We evaluate several o ptical coding strategies along with an end-to-end optimization scheme for depth estimation on three datasets, including NYU Depth v2 and KITTI. We find an optim ized freeform lens design yields the best results, but chromatic aberration from a singlet lens offers significantly improved performance as well. We build a ph ysical prototype and validate that chromatic aberrations improve depth estimatio n on real-world results. In addition, we train object detection networks on the

KITTI dataset and show that the lens optimized for depth estimation also results in improved 3D object detection performance.
****

Physics-Based Rendering for Improving Robustness to Rain
Shirsendu Sukanta Halder, Jean-Francois Lalonde, Raoul de Charette; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10203-10212
To improve the robustness to rain, we present a physically-based rain rendering pipeline for realistically inserting rain into clear weather images. Our rendering relies on a physical particle simulator, an estimation of the scene lighting and an accurate rain photometric modeling to augment images with arbitrary amount of realistic rain or fog. We validate our rendering with a user study, proving our rain is judged 40% more realistic that state-of-the-art. Using our generated weather augmented Kitti and Cityscapes dataset, we conduct a thorough evaluation of deep object detection and semantic segmentation algorithms and show that their performance decreases in degraded weather, on the order of 15% for object detection and 60% for semantic segmentation. Furthermore, we show refining existing networks with our augmented images improves the robustness of both object detection and semantic segmentation algorithms. We experiment on nuScenes and measure an improvement of 15% for object detection and 35% for semantic segmentation compared to original rainy performance. Augmented databases and code are available on the project page.
****

ARGAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal
Bin Ding, Chengjiang Long, Ling Zhang, Chunxia Xiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10213-10222
In this paper we propose an attentive recurrent generative adversarial network (ARGAN) to detect and remove shadows in an image. The generator consists of multiple progressive steps. At each step a shadow attention detector is firstly exploited to generate an attention map which specifies shadow regions in the input image. Given the attention map, a negative residual by a shadow remover encoder will recover a shadow-lighter or even a shadow-free image. The discriminator is designed to classify whether the output image in the last progressive step is real or fake. Moreover, ARGAN is suitable to be trained with a semi-supervised strategy to make full use of sufficient unsupervised data. The experiments on four public datasets have demonstrated that our ARGAN is robust to detect both simple and complex shadows and to produce more realistic shadow removal results. It outperforms the state-of-the-art methods, especially in detail of recovering shadow areas.
****

Deep Tensor ADMM-Net for Snapshot Compressive Imaging
Jiawei Ma, Xiao-Yang Liu, Zheng Shou, Xin Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10223-10232
Snapshot compressive imaging (SCI) systems have been developed to capture high-dimensional (> 3) signals using low-dimensional off-the-shelf sensors, i.e., mapping multiple video frames into a single measurement frame. One key module of a SCI system is an accurate decoder that recovers the original video frames. However, existing model-based decoding algorithms require exhaustive parameter tuning with prior knowledge and cannot support practical applications due to the extremely long running time. In this paper, we propose a deep tensor ADMM-Net for video SCI systems that provides high-quality decoding in seconds. Firstly, we start with a standard tensor ADMM algorithm, unfold its inference iterations into a layer-wise structure, and design a deep neural network based on tensor operations. Secondly, instead of relying on a pre-specified sparse representation domain, the network learns the domain of low-rank tensor through stochastic gradient descent. It is worth noting that the proposed deep tensor ADMM-Net has potentially mathematical interpretations. On public video data, the simulation results show the proposed method achieves average 0.8 ~ 2.5 dB improvement in PSNR and 0.07 ~ 0.1 in SSIM, and 1500x~ 3600 xspeedups over the state-of-the-art methods. On r

eal data captured by SCI cameras, the experimental results show comparable visua l results with the state-of-the-art methods but in much shorter running time.
****

Convex Relaxations for Consensus and Non-Minimal Problems in 3D Vision
Thomas Probst, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 102 33-10242
In this paper, we formulate a generic non-minimal solver using the existing tool s of Polynomials Optimization Problems (POP) from computational algebraic geomet ry. The proposed method exploits the well known Shor's or Lasserre's relaxations , whose theoretical aspects are also discussed. Notably, we further exploit the POP formulation of non-minimal solver also for the generic consensus maximizatio n problems in 3D vision. Our framework is simple and straightforward to implemen t, which is also supported by three diverse applications in 3D vision, namely ri gid body transformation estimation, Non-Rigid Structure-from-Motion (NRSfM), and camera autocalibration. In all three cases, both non-minimal and consensus maxi mization are tested, which are also compared against the state-of-the-art method s. Our results are competitive to the compared methods, and are also coherent wi th our theoretical analysis. The main contribution of this paper is the claim th at a good approximate solution for many polynomial problems involved in 3D visio n can be obtained using the existing theory of numerical computational algebra. This claim leads us to reason about why many relaxed methods in 3D vision behave so well? And also allows us to offer a generic relaxed solver in a rather strai ghtforward way. We further show that the convex relaxation of these polynomials can easily be used for maximizing consensus in a deterministic manner. We suppor t our claim using several experiments for aforementioned three diverse problems in 3D vision.
****

Pareto Meets Huber: Efficiently Avoiding Poor Minima in Robust Estimation
Christopher Zach, Guillaume Bourmaud; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10243-10251
Robust cost optimization is the task of fitting parameters to data points contai ning outliers. In particular, we focus on large-scale computer vision problems, such as bundle adjustment, where Non-Linear Least Square (NLLS) solvers are the current workhorse. In this context, NLLS-based state of the art algorithms have been designed either to quickly improve the target objective and find a local mi nimum close to the initial value of the parameters, or to have a strong ability to escape poor local minima. In this paper, we propose a novel algorithm relying on multi-objective optimization which allows to match those two properties. We experimentally demonstrate that our algorithm has an ability to escape poor loca l minima that is on par with the best performing algorithms with a faster decrea se of the target objective.
****

K-Best Transformation Synchronization
Yifan Sun, Jiacheng Zhuo, Arnav Mohan, Qixing Huang; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2019, pp. 10252-10261
In this paper, we introduce the problem of K-best transformation synchronization for the purpose of multiple scan matching. Given noisy pair-wise transformation s computed between a subset of depth scan pairs, K-best transformation synchroni zation seeks to output multiple consistent relative transformations. This proble m naturally arises in many geometry reconstruction applications, where the under lying object possesses self-symmetry. For approximately symmetric or even non-sy mmetric objects, K-best solutions offer an intermediate presentation for recover ing the underlying single-best solution. We introduce a simple yet robust iterat ive algorithm for K-best transformation synchronization, which alternates betwee n transformation propagation and transformation clustering. We present theoretic al guarantees on the robust and exact recoveries of our algorithm. Experimental results demonstrate the advantage of our approach against state-of-the-art trans formation synchronization techniques on both synthetic and real datasets.
****

Parametric Majorization for Data-Driven Energy Minimization Methods
Jonas Geiping, Michael Moeller; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10262-10273
Energy minimization methods are a classical tool in a multitude of computer vision applications. While they are interpretable and well-studied, their regularity assumptions are difficult to design by hand. Deep learning techniques on the other hand are purely data-driven, often provide excellent results, but are very difficult to constrain to predefined physical or safety-critical models. A possible combination between the two approaches is to design a parametric energy and train the free parameters in such a way that minimizers of the energy correspond to desired solution on a set of training examples. Unfortunately, such formulations typically lead to bi-level optimization problems, on which common optimization algorithms are difficult to scale to modern requirements in data processing and efficiency. In this work, we present a new strategy to optimize these bi-level problems. We investigate surrogate single-level problems that majorize the target problems and can be implemented with existing tools, leading to efficient algorithms without collapse of the energy function. This framework of strategies enables new avenues to the training of parameterized energy minimization models from large data.
****
A Bayesian Optimization Framework for Neural Network Compression
Xingchen Ma, Amal Rannen Triki, Maxim Berman, Christos Sagonas, Jacques Cali, Matthew B. Blaschko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10274-10283
Neural network compression is an important step for deploying neural networks where speed is of high importance, or on devices with limited memory. It is necessary to tune compression parameters in order to achieve the desired trade-off between size and performance. This is often done by optimizing the loss on a validation set of data, which should be large enough to approximate the true risk and therefore yield sufficient generalization ability. However, using a full validation set can be computationally expensive. In this work, we develop a general Bayesian optimization framework for optimizing functions that are computed based on U-statistics. We propagate Gaussian uncertainties from the statistics through the Bayesian optimization framework yielding a method that gives a probabilistic approximation certificate of the result. We then apply this to parameter selection in neural network compression. Compression objectives that can be written as U-statistics are typically based on empirical risk and knowledge distillation for deep discriminative models. We demonstrate our method on VGG and ResNet models, and the resulting system can find optimal compression parameters for relatively high-dimensional parametrizations in a matter of minutes on a standard desktop machine, orders of magnitude faster than competing methods.
****
HiPPI: Higher-Order Projected Power Iterations for Scalable Multi-Matching
Florian Bernard, Johan Thunberg, Paul Swoboda, Christian Theobalt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10284-10293
The matching of multiple objects (e.g. shapes or images) is a fundamental problem in vision and graphics. In order to robustly handle ambiguities, noise and repetitive patterns in challenging real-world settings, it is essential to take geometric consistency between points into account. Computationally, the multi-matching problem is difficult. It can be phrased as simultaneously solving multiple (NP-hard) quadratic assignment problems (QAPs) that are coupled via cycle-consistency constraints. The main limitations of existing multi-matching methods are that they either ignore geometric consistency and thus have limited robustness, or they are restricted to small-scale problems due to their (relatively) high computational cost. We address these shortcomings by introducing a Higher-order Projected Power Iteration method, which is (i) efficient and scales to tens of thousands of points, (ii) straightforward to implement, (iii) able to incorporate geometric consistency, (iv) guarantees cycle-consistent multi-matchings, and (iv) comes with theoretical convergence guarantees. Experimentally we show that our ap

proach is superior to existing methods.
****
Language-Conditioned Graph Networks for Relational Reasoning
Ronghang Hu, Anna Rohrbach, Trevor Darrell, Kate Saenko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10294-10303
Solving grounded language tasks often requires reasoning about relationships between objects in the context of a given task. For example, to answer the question "What color is the mug on the plate?" we must check the color of the specific mug that satisfies the "on" relationship with respect to the plate. Recent work has proposed various methods capable of complex relational reasoning. However, most of their power is in the inference structure, while the scene is represented with simple local appearance features. In this paper, we take an alternate approach and build contextualized representations for objects in a visual scene to support relational reasoning. We propose a general framework of Language-Conditioned Graph Networks (LCGN), where each node represents an object, and is described by a context-aware representation from related objects through iterative message passing conditioned on the textual input. E.g., conditioning on the "on" relationship to the plate, the object "mug" gathers messages from the object "plate" to update its representation to "mug on the plate", which can be easily consumed by a simple classifier for answer prediction. We experimentally show that our LCGN approach effectively supports relational reasoning and improves performance across several tasks and datasets. Our code is available at http://ronghanghu.com/lcgn.
****
Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction
Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, Graham W. Taylor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10304-10312
Conditional text-to-image generation is an active area of research, with many possible applications. Existing research has primarily focused on generating a single image from available conditioning information in one step. One practical extension beyond one-step generation is a system that generates an image iteratively, conditioned on ongoing linguistic input or feedback. This is significantly more challenging than one-step generation tasks, as such a system must understand the contents of its generated images with respect to the feedback history, the current feedback, as well as the interactions among concepts present in the feedback history. In this work, we present a recurrent image generation model which takes into account both the generated output up to the current step as well as all past instructions for generation. We show that our model is able to generate the background, add new objects, and apply simple transformations to existing objects. We believe our approach is an important step toward interactive generation. Code and data is available at: https://www.microsoft.com/en-us/research/project/generative-neural-visual-artist-geneva/.
****
Relation-Aware Graph Attention Network for Visual Question Answering
Linjie Li, Zhe Gan, Yu Cheng, Jingjing Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10313-10322
In order to answer semantically-complicated questions about an image, a Visual Question Answering (VQA) model needs to fully understand the visual scene in the image, especially the interactive dynamics between different objects. We propose a Relation-aware Graph Attention Network (ReGAT), which encodes each image into a graph and models multi-type inter-object relations via a graph attention mechanism, to learn question-adaptive relation representations. Two types of visual object relations are explored: (i) Explicit Relations that represent geometric positions and semantic interactions between objects; and (ii) Implicit Relations that capture the hidden dynamics between image regions. Experiments demonstrate that ReGAT outperforms prior state-of-the-art approaches on both VQA 2.0 and VQA

-CP v2 datasets. We further show that ReGAT is compatible to existing VQA architectures, and can be used as a generic relation encoder to boost the model performance for VQA.
****

Unpaired Image Captioning via Scene Graph Alignments
Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, Gang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10323-10332
Most of current image captioning models heavily rely on paired image-caption datasets. However, getting large scale image-caption paired data is labor-intensive and time-consuming. In this paper, we present a scene graph-based approach for unpaired image captioning. Our framework comprises an image scene graph generator, a sentence scene graph generator, a scene graph encoder, and a sentence decoder. Specifically, we first train the scene graph encoder and the sentence decoder on the text modality. To align the scene graphs between images and sentences, we propose an unsupervised feature alignment method that maps the scene graph features from the image to the sentence modality. Experimental results show that our proposed model can generate quite promising results without using any image-caption training pairs, outperforming existing methods by a wide margin.
****

Modeling Inter and Intra-Class Relations in the Triplet Loss for Zero-Shot Learning
Yannick Le Cacheux, Herve Le Borgne, Michel Crucianu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10333-10342
Recognizing visual unseen classes, i.e. for which no training data is available, is known as Zero Shot Learning (ZSL). Some of the best performing methods apply the triplet loss to seen classes to learn a mapping between visual representations of images and attribute vectors that constitute class prototypes. They nevertheless make several implicit assumptions that limit their performance on real use cases, particularly with fine-grained datasets comprising a large number of classes. We identify three of these assumptions and put forward corresponding novel contributions to address them. Our approach consists in taking into account both inter-class and intra-class relations, respectively by being more permissive with confusions between similar classes, and by penalizing visual samples which are atypical to their class. The approach is tested on four datasets, including the large-scale ImageNet, and exhibits performances significantly above recent methods, even generative methods based on more restrictive hypotheses.
****

Occlusion-Shared and Feature-Separated Network for Occlusion Relationship Reasoning
Rui Lu, Feng Xue, Menghan Zhou, Anlong Ming, Yu Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10343-10352
Occlusion relationship reasoning demands closed contour to express the object, and orientation of each contour pixel to describe the order relationship between objects. Current CNN-based methods neglect two critical issues of the task: (1) simultaneous existence of the relevance and distinction for the two elements, i.e, occlusion edge and occlusion orientation; and (2) inadequate exploration to the orientation features. For the reasons above, we propose the Occlusion-shared and Feature-separated Network (OFNet). On one hand, considering the relevance between edge and orientation, two sub-networks are designed to share the occlusion cue. On the other hand, the whole network is split into two paths to learn the high semantic features separately. Moreover, a contextual feature for orientation prediction is extracted, which represents the bilateral cue of the foreground and background areas. The bilateral cue is then fused with the occlusion cue to precisely locate the object regions. Finally, a stripe convolution is designed to further aggregate features from surrounding scenes of the occlusion edge. The proposed OFNet remarkably advances the state-of-the-art approaches on PIOD and BSDS ownership dataset.
****

Mixture-Kernel Graph Attention Network for Situation Recognition

Mohammed Suhail, Leonid Sigal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10363-10372
Understanding images beyond salient actions involves reasoning about scene context, objects, and the roles they play in the captured event. Situation recognition has recently been introduced as the task of jointly reasoning about the verbs (actions) and a set of semantic-role and entity (noun) pairs in the form of action frames. Labeling an image with an action frame requires an assignment of values (nouns) to the roles based on the observed image content. Among the inherent challenges are the rich conditional structured dependencies between the output role assignments and the overall semantic sparsity. In this paper, we propose a novel mixture-kernel attention graph neural network (GNN) architecture designed to address these challenges. Our GNN enables dynamic graph structure during training and inference, through the use of a graph attention mechanism, and context-aware interactions between role pairs. We illustrate the efficacy of our model and design choices by conducting experiments on imSitu benchmark dataset, with accuracy improvements of up to 10% over the state-of-the-art.
****

Learning Similarity Conditions Without Explicit Supervision
Reuben Tan, Mariya I. Vasileva, Kate Saenko, Bryan A. Plummer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10373-10382
Many real-world tasks require models to compare images along multiple similarity conditions (e.g. similarity in color, category or shape). Existing methods often reason about these complex similarity relationships by learning condition-aware embeddings. While such embeddings aid models in learning different notions of similarity, they also limit their capability to generalize to unseen categories since they require explicit labels at test time. To address this deficiency, we propose an approach that jointly learns representations for the different similarity conditions and their contributions as a latent variable without explicit supervision. Comprehensive experiments across three datasets, Polyvore-Outfits, Maryland-Polyvore and UT-Zappos50k, demonstrate the effectiveness of our approach: our model outperforms the state-of-the-art methods, even those that are strongly supervised with pre-defined similarity conditions, on fill-in-the-blank, outfit compatibility prediction and triplet prediction tasks. Finally, we show that our model learns different visually-relevant semantic sub-spaces that allow it to generalize well to unseen categories.
****

Joint Prediction for Kinematic Trajectories in Vehicle-Pedestrian-Mixed Scenes
Huikun Bi, Zhong Fang, Tianlu Mao, Zhaoqi Wang, Zhigang Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10383-10392
Trajectory prediction for objects is challenging and critical for various applications (e.g., autonomous driving, and anomaly detection). Most of the existing methods focus on homogeneous pedestrian trajectories prediction, where pedestrians are treated as particles without size. However, they fall short of handling crowded vehicle-pedestrian-mixed scenes directly since vehicles, limited with kinematics in reality, should be treated as rigid, non-particle objects ideally. In this paper, we tackle this problem using separate LSTMs for heterogeneous vehicles and pedestrians. Specifically, we use an oriented bounding box to represent each vehicle, calculated based on its position and orientation, to denote its kinematic trajectories. We then propose a framework called VP-LSTM to predict the kinematic trajectories of both vehicles and pedestrians simultaneously. In order to evaluate our model, a large dataset containing the trajectories of both vehicles and pedestrians in vehicle-pedestrian-mixed scenes is specially built. Through comparisons between our method with state-of-the-art approaches, we show the effectiveness and advantages of our method on kinematic trajectories prediction in vehicle-pedestrian-mixed scenes.
****

Learning to Caption Images Through a Lifetime by Asking Questions
Tingke Shen, Amlan Kar, Sanja Fidler; Proceedings of the IEEE/CVF Internationa

l Conference on Computer Vision (ICCV), 2019, pp. 10393-10402
In order to bring artificial agents into our lives, we will need to go beyond su pervised learning on closed datasets to having the ability to continuously expan d knowledge. Inspired by a student learning in a classroom, we present an agent that can continuously learn by posing natural language questions to humans. Our agent is composed of three interacting modules, one that performs captioning, an other that generates questions and a decision maker that learns when to ask ques tions by implicitly reasoning about the uncertainty of the agent and expertise o f the teacher. As compared to current active learning methods which query images for full captions, our agent is able to ask pointed questions to improve the ge nerated captions. The agent trains on the improved captions, expanding its knowl edge. We show that our approach achieves better performance using less human sup ervision than the baselines on the challenging MSCOCO dataset.
****

VrR-VG: Refocusing Visually-Relevant Relationships
Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, Tao Mei; Procee dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10403-10412
Relationships encode the interactions among individual instances and play a crit ical role in deep visual scene understanding. Suffering from the high predictabi lity with non-visual information, relationship models tend to fit the statistica l bias rather than "learning" to infer the relationships from images. To encoura ge further development in visual relationships, we propose a novel method to min e more valuable relationships by automatically pruning visually-irrelevant relat ionships. We construct a new scene graph dataset named Visually-Relevant Relatio nships Dataset (VrR-VG) based on Visual Genome. Compared with existing datasets, the performance gap between learnable and statistical method is more significan t in VrR-VG, and frequency-based analysis does not work anymore. Moreover, we pr opose to learn a relationship-aware representation by jointly considering instan ces, attributes and relationships. By applying the representation-aware feature learned on VrR-VG, the performances of image captioning and visual question answ ering are systematically improved, which demonstrates the effectiveness of both our dataset and features embedding schema. Both our VrR-VG dataset and represent ation-aware features will be made publicly available soon.
****

TAPA-MVS: Textureless-Aware PAtchMatch Multi-View Stereo
Andrea Romanoni, Matteo Matteucci; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2019, pp. 10413-10422
One of the most successful approaches in Multi-View Stereo estimates a depth map and a normal map for each view via PatchMatch-based optimization and fuses them into a consistent 3D points cloud. This approach relies on photo-consistency to evaluate the goodness of a depth estimate. It generally produces very accurate results; however, the reconstructed model often lacks completeness, especially i n correspondence of broad untextured areas where the photo-consistency metrics a re unreliable. Assuming the untextured areas piecewise planar, in this paper we generate novel PatchMatch hypotheses so to expand reliable depth estimates in ne ighboring untextured regions. At the same time, we modify the photo-consistency measure such to favor standard or novel PatchMatch depth hypotheses depending on the textureness of the considered area. We also propose a depth refinement step to filter wrong estimates and to fill the gaps on both the depth maps and norma l maps while preserving the discontinuities. The effectiveness of our new method s has been tested against several state of the art algorithms in the publicly av ailable ETH3D dataset containing a wide variety of high and low-resolution image s.
****

U4D: Unsupervised 4D Dynamic Scene Understanding
Armin Mustafa, Chris Russell, Adrian Hilton; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2019, pp. 10423-10432
We introduce the first approach to solve the challenging problem of unsupervised 4D visual scene understanding for complex dynamic scenes with multiple interact

ing people from multi-view video. Our approach simultaneously estimates a detail ed model that includes a per-pixel semantically and temporally coherent reconstr uction, together with instance-level segmentation exploiting photo-consistency, semantic and motion information. We further leverage recent advances in 3D pose estimation to constrain the joint semantic instance segmentation and 4D temporal ly coherent reconstruction. This enables per person semantic instance segmentati on of multiple interacting people in complex dynamic scenes. Extensive evaluatio n of the joint visual scene understanding framework against state-of-the-art met hods on challenging indoor and outdoor sequences demonstrates a significant (app rox 40%) improvement in semantic segmentation, reconstruction and scene flow acc uracy.

****

Hierarchical Point-Edge Interaction Network for Point Cloud Semantic Segmentatio n

Li Jiang, Hengshuang Zhao, Shu Liu, Xiaoyong Shen, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10433-10441

We achieve 3D semantic scene labeling by exploring semantic relation between eac h point and its contextual neighbors through edges. Besides an encoder-decoder b ranch for predicting point labels, we construct an edge branch to hierarchically integrate point features and generate edge features. To incorporate point featu res in the edge branch, we establish a hierarchical graph framework, where the g raph is initialized from a coarse layer and gradually enriched along the point d ecoding process. For each edge in the final graph, we predict a label to indicat e the semantic consistency of the two connected points to enhance point predicti on. At different layers, edge features are also fed into the corresponding point module to integrate contextual information for message passing enhancement in l ocal regions. The two branches interact with each other and cooperate in segment ation. Decent experimental results on several 3D semantic labeling datasets demo nstrate the effectiveness of our work.

****

Multi-Angle Point Cloud-VAE: Unsupervised Feature Learning for 3D Point Clouds F rom Multiple Angles by Joint Self-Reconstruction and Half-to-Half Prediction

Zhizhong Han, Xiyang Wang, Yu-Shen Liu, Matthias Zwicker; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10442-104 51

Unsupervised feature learning for point clouds has been vital for large-scale po int cloud understanding. Recent deep learning based methods depend on learning g lobal geometry from self-reconstruction. However, these methods are still suffer ing from ineffective learning of local geometry, which significantly limits the discriminability of learned features. To resolve this issue, we propose MAP-VAE to enable the learning of global and local geometry by jointly leveraging global and local self-supervision. To enable effective local self-supervision, we intr oduce multi-angle analysis for point clouds. In a multi-angle scenario, we first split a point cloud into a front half and a back half from each angle, and then , train MAP-VAE to learn to predict a back half sequence from the corresponding front half sequence. MAP-VAE performs this half-to-half prediction using RNN to simultaneously learn each local geometry and the spatial relationship among them . In addition, MAP-VAE also learns global geometry via self-reconstruction, wher e we employ a variational constraint to facilitate novel shape generation. The o utperforming results in four shape analysis tasks show that MAP-VAE can learn mo re discriminative global or local features than the state-of-the-art methods.

****

P-MVSNet: Learning Patch-Wise Matching Confidence Aggregation for Multi-View Ste reo

Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, Yawei Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10452-104 61

Learning-based methods are demonstrating their strong competitiveness in estimat ing depth for multi-view stereo reconstruction in recent years. Among them the a

pproaches that generate cost volumes based on the plane-sweeping algorithm and t
hen use them for feature matching have shown to be very prominent recently. The
plane-sweep volumes are essentially anisotropic in depth and spatial directions,
 but they are often approximated by isotropic cost volumes in those methods, whi
ch could be detrimental. In this paper, we propose a new end-to-end deep learnin
g network of P-MVSNet for multi-view stereo based on isotropic and anisotropic 3
D convolutions. Our P-MVSNet consists of two core modules: a patch-wise aggregat
ion module learns to aggregate the pixel-wise correspondence information of extr
acted features to generate a matching confidence volume, from which a hybrid 3D
U-Net then infers a depth probability distribution and predicts the depth maps.
We perform extensive experiments on the DTU and Tanks & Temples benchmark datase
ts, and the results show that the proposed P-MVSNet achieves the state-of-the-ar
t performance over many existing methods on multi-view stereo.
****

SME-Net: Sparse Motion Estimation for Parametric Video Prediction Through Reinfo
rcement Learning

Yung-Han Ho, Chuan-Yuan Cho, Wen-Hsiao Peng, Guo-Lun Jin; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10462-104
70

This paper leverages a classic prediction technique, known as parametric overlap
ped block motion compensation (POBMC), in a reinforcement learning framework for
 video prediction. Learning-based prediction methods with explicit motion models
 often suffer from having to estimate large numbers of motion parameters with ar
tificial regularization. Inspired by the success of sparse motion-based predicti
on for video compression, we propose a parametric video prediction on a sparse m
otion field composed of few critical pixels and their motion vectors. The predic
tion is achieved by gradually refining the estimate of a future frame in iterati
ve, discrete steps. Along the way, the identification of critical pixels and the
ir motion estimation are addressed by two neural networks trained under a reinfo
rcement learning setting. Our model achieves the state-of-the-art performance on
 CaltchPed, UCF101 and CIF datasets in one-step and multi-step prediction tests.
 It shows good generalization results and is able to learn well on small trainin
g data.
****

ClothFlow: A Flow-Based Model for Clothed Person Generation

Xintong Han, Xiaojun Hu, Weilin Huang, Matthew R. Scott; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10471-1048
0

We present ClothFlow, an appearance-flow-based generative model to synthesize cl
othed person for posed-guided person image generation and virtual try-on. By est
imating a dense flow between source and target clothing regions, ClothFlow effec
tively models the geometric changes and naturally transfers the appearance to sy
nthesize novel images as shown in Figure 1. We achieve this with a three-stage f
ramework: 1) Conditioned on a target pose, we first estimate a person semantic l
ayout to provide richer guidance to the generation process. 2) Built on two feat
ure pyramid networks, a cascaded flow estimation network then accurately estimat
es the appearance matching between corresponding clothing regions. The resulting
 dense flow warps the source image to flexibly account for deformations. 3) Fina
lly, a generative network takes the warped clothing regions as inputs and render
s the target view. We conduct extensive experiments on the DeepFashion dataset f
or pose-guided person image generation and on the VITON dataset for the virtual
try-on task. Strong qualitative and quantitative results validate the effectiven
ess of our method.
****

LADN: Local Adversarial Disentangling Network for Facial Makeup and De-Makeup

Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, Chi-Keung Tang; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp.
 10481-10490

We propose a local adversarial disentangling network (LADN) for facial makeup an
d de-makeup. Central to our method are multiple and overlapping local adversaria

l discriminators in a content-style disentangling network for achieving local de
tail transfer between facial images, with the use of asymmetric loss functions f
or dramatic makeup styles with high-frequency details. Existing techniques do no
t demonstrate or fail to transfer high-frequency details in a global adversarial
 setting, or train a single local discriminator only to ensure image structure c
onsistency and thus work only for relatively simple styles. Unlike others, our p
roposed local adversarial discriminators can distinguish whether the generated l
ocal image details are consistent with the corresponding regions in the given re
ference image in cross-image style transfer in an unsupervised setting. Incorpor
ating these technical contributions, we achieve not only state-of-the-art result
s on conventional styles but also novel results involving complex and dramatic s
tyles with high-frequency details covering large areas across multiple facial fe
atures. A carefully designed dataset of unpaired before and after makeup images
is released at https://georgegu1997.github.io/LADN-project-page.
****

Point-to-Point Video Generation
Tsun-Hsuan Wang, Yen-Chi Cheng, Chieh Hubert Lin, Hwann-Tzong Chen, Min Sun;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2019, pp. 10491-10500
While image synthesis achieves tremendous breakthroughs (e.g., generating realis
tic faces), video generation is less explored and harder to control, which limit
s its applications in the real world. For instance, video editing requires tempo
ral coherence across multiple clips and thus poses both start and end constraint
s within a video sequence. We introduce point-to-point video generation that con
trols the generation process with two control points: the targeted start- and en
d-frames. The task is challenging since the model not only generates a smooth tr
ansition of frames but also plans ahead to ensure that the generated end-frame c
onforms to the targeted end-frame for videos of various lengths. We propose to m
aximize the modified variational lower bound of conditional data likelihood unde
r a skip-frame training strategy. Our model can generate end-frame-consistent se
quences without loss of quality and diversity. We evaluate our method through ex
tensive experiments on Stochastic Moving MNIST, Weizmann Action, Human3.6M, and
BAIR Robot Pushing under a series of scenarios. The qualitative results showcase
 the effectiveness and merits of point-to-point generation.
****

Semantics-Enhanced Adversarial Nets for Text-to-Image Synthesis
Hongchen Tan, Xiuping Liu, Xin Li, Yi Zhang, Baocai Yin; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10501-105
10
This paper presents a new model, Semantics-enhanced Generative Adversarial Netwo
rk (SEGAN), for fine-grained text-to-image generation. We introduce two modules,
 a Semantic Consistency Module (SCM) and an Attention Competition Module (ACM),
to our SEGAN. The SCM incorporates image-level semantic consistency into the tra
ining of the Generative Adversarial Network (GAN), and can diversify the generat
ed images and improve their structural coherence. A Siamese network and two type
s of semantic similarities are designed to map the synthesized image and the gro
undtruth image to nearby points in the latent semantic feature space. The ACM co
nstructs adaptive attention weights to differentiate keywords from unimportant w
ords, and improves the stability and accuracy of SEGAN. Extensive experiments de
monstrate that our SEGAN significantly outperforms existing state-of-the-art met
hods in generating photo-realistic images. All source codes and models will be r
eleased for comparative study.
****

VTNFP: An Image-Based Virtual Try-On Network With Body and Clothing Feature Pres
ervation
Ruiyun Yu, Xiaoqi Wang, Xiaohui Xie; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2019, pp. 10511-10520
Image-based virtual try-on systems with the goal of transferring a desired cloth
ing item onto the corresponding region of a person have made great strides recen
tly, but challenges remain in generating realistic looking images that preserve

both body and clothing details. Here we present a new virtual try-on network, ca lled VTNFP, to synthesize photo-realistic images given the images of a clothed p erson and a target clothing item. In order to better preserve clothing and body features, VTNFP follows a three-stage design strategy. First, it transforms the target clothing into a warped form compatible with the pose of the given person. Next, it predicts a body segmentation map of the person wearing the target clot hing, delineating body parts as well as clothing regions. Finally, the warped cl othing, body segmentation map and given person image are fused together for fine -scale image synthesis. A key innovation of VTNFP is the body segmentation map p rediction module, which provides critical information to guide image synthesis i n regions where body parts and clothing intersects, and is very beneficial for p reventing blurry pictures and preserving clothing and body part details. Experim ents on a fashion dataset demonstrate that VTNFP generates substantially better results than state-of-the-art methods.

****

Boundless: Generative Adversarial Networks for Image Extension

Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger , Ce Liu, William T. Freeman; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2019, pp. 10521-10530

Image extension models have broad applications in image editing, computational p hotography and computer graphics. While image inpainting has been extensively st udied in the literature, it is challenging to directly apply the state-of-the-ar t inpainting methods to image extension as they tend to generate blurry or repet itive pixels with inconsistent semantics. We introduce semantic conditioning to the discriminator of a generative adversarial network (GAN), and achieve strong results on image extension with coherent semantics and visually pleasing colors and textures. We also show promising results in extreme extensions, such as pano rama generation.

****

Image Synthesis From Reconfigurable Layout and Style

Wei Sun, Tianfu Wu; Proceedings of the IEEE/CVF International Conference on Com puter Vision (ICCV), 2019, pp. 10531-10540

Despite remarkable recent progress on both unconditional and conditional image s ynthesis, it remains a long- standing problem to learn generative models that ar e capable of synthesizing realistic and sharp images from re- configurable spati al layout (i.e., bounding boxes + class labels in an image lattice) and style (i .e., structural and appearance variations encoded by latent vectors), especially at high resolution. By reconfigurable, it means that a model can preserve the i ntrinsic one-to-many mapping from a given layout to multiple plausible images wi th different styles, and is adaptive with respect to perturbations of a layout a nd style latent code. In this paper, we present a layout- and style-based archit ecture for generative adversarial networks (termed LostGANs) that can be trained end-to-end to generate images from reconfigurable layout and style. Inspired by the vanilla StyleGAN, the proposed LostGAN consists of two new components: (i) learning fine-grained mask maps in a weakly-supervised manner to bridge the gap between layouts and images, and (ii) learning object instance-specific layout-aw are feature normalization (ISLA-Norm) in the generator to realize multi-object s tyle generation. In experiments, the proposed method is tested on the COCO-Stuff dataset and the Visual Genome dataset with state-of-the-art performance obtaine d. The code and pretrained models are available at https://github.com/iVMCL/Lost GANs.

****

Attribute Manipulation Generative Adversarial Networks for Fashion Images

Kenan E. Ak, Joo Hwee Lim, Jo Yew Tham, Ashraf A. Kassim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10541-105 50

Recent advances in Generative Adversarial Networks (GANs) have made it possible to conduct multi-domain image-to-image translation using a single generative net work. While recent methods such as Ganimation and SaGAN are able to conduct tran slations on attribute-relevant regions using attention, they do not perform well

when the number of attributes increases as the training of attention masks most ly rely on classification losses. To address this and other limitations, we intr oduce Attribute Manipulation Generative Adversarial Networks (AMGAN) for fashion images. While AMGAN's generator network uses class activation maps (CAMs) to em power its attention mechanism, it also exploits perceptual losses by assigning r eference (target) images based on attribute similarities. AMGAN incorporates an additional discriminator network that focuses on attribute-relevant regions to d etect unrealistic translations. Additionally, AMGAN can be controlled to perform attribute manipulations on specific regions such as the sleeve or torso regions . Experiments show that AMGAN outperforms state-of-the-art methods using traditi onal evaluation metrics as well as an alternative one that is based on image ret rieval.

****

Few-Shot Unsupervised Image-to-Image Translation

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtine n, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10551-10560

Unsupervised image-to-image translation methods learn to map images in a given c lass to an analogous image in a different class, drawing on unstructured (non-re gistered) datasets of images. While remarkably successful, current methods requi re access to many images in both source and destination classes at training time . We argue this greatly limits their use. Drawing inspiration from the human cap ability of picking up the essence of a novel object from a small number of examp les and generalizing from there, we seek a few-shot, unsupervised image-to-image translation algorithm that works on previously unseen target classes that are s pecified, at test time, only by a few example images. Our model achieves this fe w-shot generation capability by coupling an adversarial training scheme with a n ovel network design. Through extensive experimental validation and comparisons t o several baseline methods on benchmark datasets, we verify the effectiveness of the proposed framework. Our implementation and datasets are available at https: //github.com/NVlabs/FUNIT

****

Very Long Natural Scenery Image Prediction by Outpainting

Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, Shuicheng Yan; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10561-1 0570

Comparing to image inpainting, image outpainting receives less attention due to two challenges in it. The first challenge is how to keep the spatial and content consistency between generated images and original input. The second challenge i s how to maintain high quality in generated results, especially for multi-step g enerations in which generated regions are spatially far away from the initial in put. To solve the two problems, we devise some innovative modules, named Skip Ho rizontal Connection and Recurrent Content Transfer, and integrate them into our designed encoder-decoder structure. By this design, our network can generate hig hly realistic outpainting prediction effectively and efficiently. Other than tha t, our method can generate new images with very long sizes while keeping the sam e style and semantic content as the given input. To test the effectiveness of th e proposed architecture, we collect a new scenery dataset with diverse, complica ted natural scenes. The experimental results on this dataset have demonstrated t he efficacy of our proposed network.

****

Scaling Recurrent Models via Orthogonal Approximations in Tensor Trains

Ronak Mehta, Rudrasis Chakraborty, Yunyang Xiong, Vikas Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1057 1-10579

Modern deep networks have proven to be very effective for analyzing real world i mages. However, their application in medical imaging is still in its early stage s, primarily due to the large size of three-dimensional images, requiring enormo us convolutional or fully connected layers - if we treat an image (and not image patches) as a sample. These issues only compound when the focus moves towards l

ongitudinal analysis of 3D image volumes through recurrent structures, and when a point estimate of model parameters is insufficient in scientific applications where a reliability measure is necessary. Using insights from differential geome try, we adapt the tensor train decomposition to construct networks with signific antly fewer parameters, allowing us to train powerful recurrent networks on whol e brain image volume sequences. We describe the "orthogonal" tensor train, and d emonstrate its ability to express a standard network layer both theoretically an d empirically. We show its ability to effectively reconstruct whole brain volume s with faster convergence and stronger confidence intervals compared to the stan dard tensor train decomposition. We provide code and show experiments on the ADN I dataset using image sequences to regress on a cognition related outcome.
****

A Deep Cybersickness Predictor Based on Brain Signal Analysis for Virtual Realit y Contents
Jinwoo Kim, Woojae Kim, Heeseok Oh, Seongmin Lee, Sanghoon Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10 580-10589
What if we could interpret the cognitive state of a user while experiencing a vi rtual reality (VR) and estimate the cognitive state from a visual stimulus? In t his paper, we address the above question by developing an electroencephalography (EEG) driven VR cybersickness prediction model. The EEG data has been widely ut ilized to learn the cognitive representation of brain activity. In the first sta ge, to fully exploit the advantages of the EEG data, it is transformed into the multi-channel spectrogram which enables to account for the correlation of spectr al and temporal coefficient. Then, a convolutional neural network (CNN) is appli ed to encode the cognitive representation of the EEG spectrogram. In the second stage, we train a cybersickness prediction model on the VR video sequence by des igning a Recurrent Neural Network (RNN). Here, the encoded cognitive representat ion is transferred to the model to train the visual and cognitive features for c ybersickness prediction. Through the proposed framework, it is possible to predi ct the cybersickness level that reflects brain activity automatically. We use 8- channels EEG data to record brain activity while more than 200 subjects experien ce 44 different VR contents. After rigorous training, we demonstrate that the pr oposed framework reliably estimates cognitive states without the EEG data. Furth ermore, it achieves state-of-the-art performance comparing to existing VR cybers ickness prediction models.
****

Learning With Unsure Data for Medical Image Diagnosis
Botong Wu, Xinwei Sun, Lingjing Hu, Yizhou Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10590-10599
In image-based disease prediction, it can be hard to give certain cases a determ inistic "disease/normal" label due to lack of enough information, e.g., at its e arly stage. We call such cases "unsure" data. Labeling such data as unsure sugge sts follow-up examinations so as to avoid irreversible medical accident/loss in contrast to incautious prediction. This is a common practice in clinical diagnos is, however, mostly neglected by existing methods. Learning with unsure data als o interweaves with two other practical issues: (i) data imbalance issue that may incur model-bias towards the majority class, and (ii) conservative/aggressive s trategy consideration, i.e., the negative (normal) samples and positive (disease) samples should NOT be treated equally \-- the former should be detected with a high precision (conservativeness) and the latter should be detected with a high recall (aggression) to avoid missing opportunity for treatment. Mixed with thes e issues, learning with unsure data becomes particularly challenging. In this pa per, we raise "learning with unsure data" problem and formulate it as an ordinal regression and propose a unified end-to-end learning framework, which also cons iders the aforementioned two issues: (i) incorporate cost-sensitive parameters t o alleviate the data imbalance problem, and (ii) execute the conservative and ag gressive strategies by introducing two parameters in the training procedure. The benefits of learning with unsure data and validity of our models are demonstrat ed on the prediction of Alzheimer's Disease and lung nodules.

****
Recursive Cascaded Networks for Unsupervised Medical Image Registration
Shengyu Zhao, Yue Dong, Eric I-Chao Chang, Yan Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10600-10610
We present recursive cascaded networks, a general architecture that enables learning deep cascades, for deformable image registration. The proposed architecture is simple in design and can be built on any base network. The moving image is warped successively by each cascade and finally aligned to the fixed image; this procedure is recursive in a way that every cascade learns to perform a progressive deformation for the current warped image. The entire system is end-to-end and jointly trained in an unsupervised manner. In addition, enabled by the recursive architecture, one cascade can be iteratively applied for multiple times during testing, which approaches a better fit between each of the image pairs. We evaluate our method on 3D medical images, where deformable registration is most commonly applied. We demonstrate that recursive cascaded networks achieve consistent, significant gains and outperform state-of-the-art methods. The performance reveals an increasing trend as long as more cascades are trained, while the limit is not observed. Code is available at https://github.com/zsyzzsoft/Recursive-Cascaded-Networks.
****
DUAL-GLOW: Conditional Flow-Based Generative Model for Modality Transfer
Haoliang Sun, Ronak Mehta, Hao H. Zhou, Zhichun Huang, Sterling C. Johnson, Vivek Prabhakaran, Vikas Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10611-10620
Positron emission tomography (PET) imaging is an imaging modality for diagnosing a number of neurological diseases. In contrast to Magnetic Resonance Imaging (MRI), PET is costly and involves injecting a radioactive substance into the patient. Motivated by developments in modality transfer in vision, we study the generation of certain types of PET images from MRI data. We derive new flow-based generative models which we show perform well in this small sample size regime (much smaller than dataset sizes available in standard vision tasks). Our formulation, DUAL-GLOW, is based on two invertible networks and a relation network that maps the latent spaces to each other. We discuss how given the prior distribution, learning the conditional distribution of PET given the MRI image reduces to obtaining the conditional distribution between the two latent codes w.r.t. the two image types. We also extend our framework to leverage "side" information (or attributes) when available. By controlling the PET generation through "conditioning" on age, our model is also able to capture brain FDG-PET (hypometabolism) changes, as a function of age. We present experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with 826 subjects, and obtain good performance in PET image synthesis, qualitatively and quantitatively better than recent works.
****
Dilated Convolutional Neural Networks for Sequential Manifold-Valued Data
Xingjian Zhen, Rudrasis Chakraborty, Nicholas Vogt, Barbara B. Bendlin, Vikas Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10621-10631
Efforts are underway to study ways via which the power of deep neural networks can be extended to non-standard data types such as structured data (e.g., graphs) or manifold-valued data (e.g., unit vectors or special matrices). Often, sizable empirical improvements are possible when the geometry of such data spaces are incorporated into the design of the model, architecture, and algorithms. Motivated by neuroimaging applications, we study formulations where the data are sequential manifold-valued measurements. This case is common in brain imaging, where the samples correspond to symmetric positive definite matrices or orientation distribution functions. Instead of a recurrent model which poses computational/technical issues, and inspired by recent results showing the viability of dilated convolutional models for sequence prediction, we develop a dilated convolutional neural network architecture for this task. On the technical side, we show how the modules needed in our network can be derived while explicitly taking the Rie

mannian manifold structure into account. We show how the operations needed can l
everage known results for calculating the weighted Frechet Mean (wFM). Finally,
we present scientific results for group difference analysis in Alzheimer's disea
se (AD) where the groups are derived using AD pathology load: here the model fin
ds several brain fiber bundles that are related to AD even when the subjects are
 all still cognitively healthy.
****

Align, Attend and Locate: Chest X-Ray Diagnosis via Contrast Induced Attention N
etwork With Limited Supervision
Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, Yizhou Yu; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019
, pp. 10632-10641
Obstacles facing accurate identification and localization of diseases in chest X
-ray images lie in the lack of high-quality images and annotations. In this pape
r, we propose a Contrast Induced Attention Network (CIA-Net), which exploits the
 highly structured property of chest X-ray images and localizes diseases via con
trastive learning on the aligned positive and negative samples. To force the att
ention module to focus only on sites of abnormalities, we also introduce a learn
able alignment module to adjust all the input images, which eliminates variation
s of scales, angles, and displacements of X-ray images generated under bad scan
conditions. We show that the use of contrastive attention and alignment module a
llows the model to learn rich identification and localization information using
only a small amount of location annotations, resulting in state-of-the-art perfo
rmance in NIH chest X-ray dataset.
****

Joint Acne Image Grading and Counting via Label Distribution Learning
Xiaoping Wu, Ni Wen, Jie Liang, Yu-Kun Lai, Dongyu She, Ming-Ming Cheng, J
ufeng Yang; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2019, pp. 10642-10651
Accurate grading of skin disease severity plays a crucial role in precise treatm
ent for patients. Acne vulgaris, the most common skin disease in adolescence, ca
n be graded by evidence-based lesion counting as well as experience-based global
 estimation in the medical field. However, due to the appearance similarity of a
cne with close severity, it is challenging to count and grade acne accurately. I
n this paper, we address the problem of acne image analysis via Label Distributi
on Learning (LDL) considering the ambiguous information among acne severity. Bas
ed on the professional grading criterion, we generate two acne label distributio
ns considering the relationship between the similar number of lesions and severi
ty of acne, respectively. We also propose a unified framework for joint acne ima
ge grading and counting, which is optimized by the multi-task learning loss. In
addition, we further build the ACNE04 dataset with annotations of acne severity
and lesion number of each image for evaluation. Experiments demonstrate that our
 proposed framework performs favorably against state-of-the-art methods. We make
 the code and dataset publicly available at https://github.com/xpwu95/ldl.
****

An Alarm System for Segmentation Algorithm Based on Shape Model
Fengze Liu, Yingda Xia, Dong Yang, Alan L. Yuille, Daguang Xu; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 106
52-10661
It is usually hard for a learning system to predict correctly on rare events tha
t never occur in the training data, and there is no exception for segmentation a
lgorithms. Meanwhile, manual inspection of each case to locate the failures beco
mes infeasible due to the trend of large data scale and limited human resource.
Therefore, we build an alarm system that will set off alerts when the segmentati
on result is possibly unsatisfactory, assuming no corresponding ground truth mas
k is provided. One plausible solution is to project the segmentation results int
o a low dimensional feature space; then learn classifiers/regressors to predict
their qualities. Motivated by this, in this paper, we learn a feature space usin
g the shape information which is a strong prior shared among different datasets
and robust to the appearance variation of input data. The shape feature is captu

red using a Variational Auto-Encoder (VAE) network that trained with only the gr ound truth masks. During testing, the segmentation results with bad shapes shall not fit the shape prior well, resulting in large loss values. Thus, the VAE is able to evaluate the quality of segmentation result on unseen data, without usin g ground truth. Finally, we learn a regressor in the one-dimensional feature spa ce to predict the qualities of segmentation results. Our alarm system is evaluat ed on several recent state-of-art segmentation algorithms for 3D medical segment ation tasks. Compared with other standard quality assessment methods, our system consistently provides more reliable prediction on the qualities of segmentation results.
****

HistoSegNet: Semantic Segmentation of Histological Tissue Type in Whole Slide Im ages

Lyndon Chan, Mahdi S. Hosseini, Corwyn Rowsell, Konstantinos N. Plataniotis, Savvas Damaskinos; Proceedings of the IEEE/CVF International Conference on Comp uter Vision (ICCV), 2019, pp. 10662-10671

In digital pathology, tissue slides are scanned into Whole Slide Images (WSI) an d pathologists first screen for diagnostically-relevant Regions of Interest (ROI s) before reviewing them. Screening for ROIs is a tedious and time-consuming vis ual recognition task which can be exhausting. The cognitive workload could be re duced by developing a visual aid to narrow down the visual search area by highli ghting (or segmenting) regions of diagnostic relevance, enabling pathologists to spend more time diagnosing relevant ROIs. In this paper, we propose HistoSegNet , a method for semantic segmentation of histological tissue type (HTT). Using th e HTT-annotated Atlas of Digital Pathology (ADP) database, we train a Convolutio nal Neural Network on the patch annotations, infer Gradient-Weighted Class Activ ation Maps, average overlapping predictions, and post-process the segmentation w ith a fully-connected Conditional Random Field. Our method out-performs more com plicated weakly-supervised semantic segmentation methods and can generalize to o ther datasets without retraining.
****

Prior-Aware Neural Network for Partially-Supervised Multi-Organ Segmentation

Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fis hman, Alan L. Yuille; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2019, pp. 10672-10681

Accurate multi-organ abdominal CT segmentation is essential to many clinical app lications such as computer-aided intervention. As data annotation requires massi ve human labor from experienced radiologists, it is common that training data is usually partially-labeled. However, these background labels can be misleading i n multi-organ segmentation since the "background" usually contains some other or gans of interest. To address the background ambiguity in these partially-labeled datasets, we propose Prior-aware Neural Network (PaNN) via explicitly incorpora ting anatomical priors on abdominal organ sizes, guiding the training process wi th domain-specific knowledge. More specifically, PaNN assumes that the average o rgan size distributions in the abdomen should approximate their empirical distri butions, a prior statistics obtained from the fully-labeled dataset. As our obje ctive is difficult to be directly optimized using stochastic gradient descent, i t is reformulated as a min-max form and optimized via the stochastic primal-dual gradient algorithm. PaNN achieves state-of-the-art performance on the MICCAI201 5 challenge "Multi-Atlas Labeling Beyond the Cranial Vault", a competition on or gan segmentation in the abdomen. We report an average Dice score of 84.97%, surp assing the prior art by a large margin of 3.27%. Code and models will be made pu blicly available.
****

CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentat ion

Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, Wei Xu; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2019, pp. 10682-10691

Histopathology image analysis plays a critical role in cancer diagnosis and trea

tment. To automatically segment the cancerous regions, fully supervised segmentation algorithms require labor-intensive and time-consuming labeling at the pixel level. In this research, we propose CAMEL, a weakly supervised learning framework for histopathology image segmentation using only image-level labels. Using multiple instance learning (MIL)-based label enrichment, CAMEL splits the image into latticed instances and automatically generates instance-level labels. After label enrichment, the instance-level labels are further assigned to the corresponding pixels, producing the approximate pixel-level labels and making fully supervised training of segmentation models possible. CAMEL achieves comparable performance with the fully supervised approaches in both instance-level classification and pixel-level segmentation on CAMELYON16 and a colorectal adenoma dataset. Moreover, the generality of the automatic labeling methodology may benefit future weakly supervised learning studies for histopathology image analysis.

****

Conditional Recurrent Flow: Conditional Generation of Longitudinal Samples With Applications to Neuroimaging

Seong Jae Hwang, Zirui Tao, Won Hwa Kim, Vikas Singh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10692-10701

We develop a conditional generative model for longitudinal image datasets based on sequential invertible neural networks. Longitudinal image acquisitions are common in various scientific and biomedical studies where often each image sequence sample may also come together with various secondary (fixed or temporally dependent) measurements. The key goal is not only to estimate the parameters of a deep generative model for the given longitudinal data, but also to enable evaluation of how the temporal course of the generated longitudinal samples are influenced as a function of induced changes in the (secondary) temporal measurements (or events). Our proposed formulation incorporates recurrent subnetworks and temporal context gating, which provides a smooth transition in a temporal sequence of generated data that can be easily informed or modulated by secondary temporal conditioning variables. We show that the formulation works well despite the smaller sample sizes common in these applications. Our model is validated on two video datasets and a longitudinal Alzheimer's disease (AD) dataset for both quantitative and qualitative evaluations of the generated samples. Further, using our generated longitudinal image samples, we show that we can capture the pathological progressions in the brain that turn out to be consistent with the existing literature, and could facilitate various types of downstream statistical analysis.

****

Multi-Stage Pathological Image Classification Using Semantic Segmentation

Shusuke Takahama, Yusuke Kurose, Yusuke Mukuta, Hiroyuki Abe, Masashi Fukayama, Akihiko Yoshizawa, Masanobu Kitagawa, Tatsuya Harada; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10702-10711

Histopathological image analysis is an essential process for the discovery of diseases such as cancer. However, it is challenging to train CNN on whole slide images (WSIs) of gigapixel resolution considering the available memory capacity. Most of the previous works divide high resolution WSIs into small image patches and separately input them into the model to classify it as a tumor or a normal tissue. However, patch-based classification uses only patch-scale local information but ignores the relationship between neighboring patches. If we consider the relationship of neighboring patches and global features, we can improve the classification performance. In this paper, we propose a new model structure combining the patch-based classification model and whole slide-scale segmentation model in order to improve the prediction performance of automatic pathological diagnosis. We extract patch features from the classification model and input them into the segmentation model to obtain a whole slide tumor probability heatmap. The classification model considers patch-scale local features, and the segmentation model can take global information into account. We also propose a new optimization method that retains gradient information and trains the model partially for end-to-end learning with limited GPU memory capacity. We apply our method to the tumor/normal prediction on WSIs and the classification performance is improved comp

ared with the conventional patch-based method.
****
Semantic-Transferable Weakly-Supervised Endoscopic Lesions Segmentation
Jiahua Dong, Yang Cong, Gan Sun, Dongdong Hou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10712-10721
Weakly-supervised learning under image-level labels supervision has been widely applied to semantic segmentation of medical lesions regions. However, 1) most existing models rely on effective constraints to explore the internal representation of lesions, which only produces inaccurate and coarse lesions regions; 2) they ignore the strong probabilistic dependencies between target lesions dataset (e.g., enteroscopy images) and well-to-annotated source diseases dataset (e.g., gastroscope images). To better utilize these dependencies, we present a new semantic lesions representation transfer model for weakly-supervised endoscopic lesions segmentation, which can exploit useful knowledge from relevant fully-labeled diseases segmentation task to enhance the performance of target weakly-labeled lesions segmentation task. More specifically, a pseudo label generator is proposed to leverage seed information to generate highly-confident pseudo pixel labels by incorporating class balance and super-pixel spatial prior. It can iteratively include more hard-to-transfer samples from weakly-labeled target dataset into training set. Afterwards, dynamically-searched feature centroids for same class among different datasets are aligned by accumulating previously-learned features. Meanwhile, adversarial learning is also employed in this paper, to narrow the gap between the lesions among different datasets in output space. Finally, we build a new medical endoscopic dataset with 3659 images collected from more than 1100 volunteers. Extensive experiments on our collected dataset and several benchmark datasets validate the effectiveness of our model.
****
Unsupervised Microvascular Image Segmentation Using an Active Contours Mimicking Neural Network
Shir Gur, Lior Wolf, Lior Golgher, Pablo Blinder; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10722-10731
The task of blood vessel segmentation in microscopy images is crucial for many diagnostic and research applications. However, vessels can look vastly different, depending on the transient imaging conditions, and collecting data for supervised training is laborious. We present a novel deep learning method for unsupervised segmentation of blood vessels. The method is inspired by the field of active contours and we introduce a new loss term, which is based on the morphological Active Contours Without Edges (ACWE) optimization method. The role of the morphological operators is played by novel pooling layers that are incorporated to the network's architecture. We demonstrate the challenges that are faced by previous supervised learning solutions, when the imaging conditions shift. Our unsupervised method is able to outperform such previous methods in both the labeled dataset, and when applied to similar but different datasets. Our code, as well as efficient pytorch reimplementations of the baseline methods VesselNN and DeepVess are attached as supplementary.
****
GLAMpoints: Greedily Learned Accurate Match Points
Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, Sandro De Zanet; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10732-10741
We introduce a novel CNN-based feature point detector - Greedily Learned Accurate Match Points (GLAMpoints) - learned in a semi-supervised manner. Our detector extracts repeatable, stable interest points with a dense coverage, specifically designed to maximize the correct matching in a specific domain, which is in contrast to conventional techniques that optimize indirect metrics. In this paper, we apply our method on challenging retinal slitlamp images, for which classical detectors yield unsatisfactory results due to low image quality and insufficient amount of low-level features. We show that GLAMpoints significantly outperforms classical detectors as well as state-of-the-art CNN-based methods in matching and registration quality for retinal images.

****