计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目： 基于计算机视觉的 Transformer 研究进展
作者： 刘文婷，卢新明
网络首发日期： 2021-12-03
引用格式： 刘文婷，卢新明．基于计算机视觉的 Transformer 研究进展[J/OL]．计算机工程与应用. https://kns.cnki.net/kcms/detail/11.2127.tp.20211129.1135.004.html

网络首发时间：2021-12-03 09:10:30
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.tp.20211129.1135.004.html

基于计算机视觉的 Transformer 研究进展

刘文婷，卢新明

山东科技大学 计算机科学与工程学院，山东 黄岛 266500

摘 要：Transformer 是一种基于自注意力机制、并行化处理数据的深度神经网络。近几年基于 Transformer 的模型成为计算机视觉任务的重要研究方向。针对目前国内基于 Transformer 综述性文章的空白，对其在计算机视觉上的应用进行概述。文章回顾了 Transformer 的基本原理，重点介绍了其在图像分类、目标检测、图像分割等七个视觉任务上的应用，并对效果显著的模型进行分析。最后对 Transformer 在计算机视觉中面临的挑战以及未来的发展趋势进行了总结和展望。

关键词：Transformer；计算机视觉；自注意力机制；神经网络
文献标志码: A **中图分类号**:TP183；TP391 doi：10.3778/j.issn.1002-8331.2106-0442

Research Progress of Transformer Based on Computer Vision

LIU Wenting[1], LU Xinming[1]

College of Computer Science and Engineering, Shandong University of Science and Technology, Huangdao, Shandong 266500, China

Abstract: Transformer is a deep neural network based on the self-attention mechanism and parallel processing data. In recent years, transformer-based models have emerged as an important area of research for computer vision tasks. Aiming at the current blanks in domestic review articles based on transformer, this paper covers its application in computer vision. This paper reviews the basic principles of the transformer model, mainly focuses on the application of seven visual tasks such as image classification, object detection and segmentation, and analyzes transformer-based models with significant effects. Finally, this paper summarizes the challenges and future development trends of the transformer model in computer vision.

Key words: Transformer; computer vision; self-attention mechanism; neural network

计算机视觉通常涉及对图像或视频的评估，主要包括图像分类、目标检测、目标跟踪、语义分割等子任务。近年来，深度学习方法应用于计算机视觉的方方面面，研究人员针对不同的任务提出了各种网络模型，取得了一系列显著的研究成果。

基于深度学习的方法在计算机视觉领域中最典型的应用就是卷积神经网络（Convolutional Neural Network，CNN）[1]。CNN 中的数据表示方式是分层的，高层特征表示依赖于底层特征，由浅入深抽象地提取高级语义特征[2]。CNN 的核心是卷积核，具有平移不变性和局部敏感性等归纳偏置[3]，可以捕捉局部的时空信息。在过去的 10 年间，CNN 存在很大的优势，在计算机视觉领域被人们寄予厚望，引领了一个时代。但是卷积这种操作缺乏对图像本身的全局理解，无法建模特征之间的依赖关系，从而不能充分地利用上下文信息。此外，卷积的权重是固定的，并不能动态地适应输入的变化。因此，研究人员尝试将自然语言处理领域中的 Transformer 模型迁移到计算机视觉任务。相比 CNN，Transformer 的自注意力机制不受局部相互作用的限制，既能挖掘长距离的依赖关系又能并行计算，可以根据不同的任务目标学习最合适的归纳偏置，在诸多视觉任务中取得了良好的效果。

Transformer 是由谷歌 2017 年在文献[4]中提出的，该模型给自然语言处理领域带来极大的震动，是一个里

基金项目：国家重点研发计划资助项目（2017YFC0804406）;山东省重点研发计划资助项目（2016ZDJS02A05）。

作者简介：刘文婷(1994-)，女，博士研究生，CCF 会员，主要研究方向：图像处理，计算机视觉，E-mail：lwt_1994@163.com;卢新明（1961-），博士，教授，主要研究方向：智慧矿山信息技术、计算机辅助设计、地理信息系统、计算机图形学等。

程碑式的模型。随着研究的推进，最近一些文章[5、6、7、8、9]创新性地将 Transformer 技术跨领域的引入到计算机视觉任务中，开创了视觉领域的新时代，代表作如图 1 所示。2018 年发布的 Image Transformer[5]最早将 Transformer 架构迁移到计算机视觉领域。从 2019 年至今，基于 Transformer 的视觉模型迅速发展，出现了很多值得关注的新成果。例如，2020 年 5 月 Carion 等[6]构建了一种新的物体检测框架 DETR（Detection Transformer），第一次将 Transformer 应用于目标检测领域。2020 年 7 月，Chen 等[7]提出了 iGPT 模型，旨在探索 GPT-2[10]算法在图像上的性能及无监督准确率的表现。2020 年 10 月 Dosovitskiy 等[8]提出了 ViT（Vision Transformer）模型，一种完全基于自注意力机制的图像分类方案，这也是 Transformer 替代标准卷积的第一部作品。2021 年 1 月 Esser 等[9]构建了 VQGAN（Vector Quantised Generative Adversarial Network），将 Transformer 和 CNN 结合应用，是第一个由语义引导生成百万像素图像的 Transformer 架构。基于 Transformer 的模型如雨后春笋般涌现，给计算机视觉领域注入了新的活力，引领了新的变革。

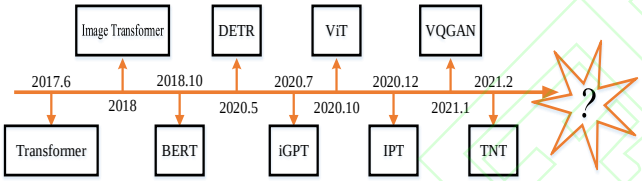

图 1 Transformer 代表作

Fig.1 Transformer masterpiece

Transformer 在计算机视觉领域能够迅速发展的原因主要有以下三个方面：

(1) 学习长距离依赖能力强。CNN 是通过不断地堆叠卷积层来实现对图像从局部信息到全局信息的提取，这种计算机制显然会导致模型臃肿，计算量大幅增加，带来梯度消失问题，甚至使整个网络无法训练收敛。而 Transformer 自带的长依赖特性，利用注意力机制来捕获全局上下文信息，抽取更强有力的特征。

(2) 多模态融合能力强。CNN 使用卷积核来获取图像信息，但不擅长融合其他模态的信息（如声音、文字、时间等）。而 Transformer 的输入不需要保持二维图像，通常可以直接对像素进行操作得到初始嵌入向量，其他模态的信息转换为向量即可直接在输入端进行融合。

(3) 模型更具可解释性。在 Transformer 的多头注意力结构中，每个头都应用独立的自注意力机制，这使得模型可以针对不同的任务在不同的表示子空间里学习相关的信息。

本文对 Transformer 在视觉领域的应用等相关工作进行整理分类，对相关模型方法进行分析，总结在该领域的研究现状，并在文末对 Transformer 的研究方向和发展趋势进行展望。

1 Transformer 基本原理

在 Transformer 提出之前，自然语言处理领域应用最广的是循环神经网络（Recurrent Neural Network, RNN）[11]，其结构如图 2 所示。RNN 中含有循环层，后一个时刻的输出来自于前面多个时刻的输入和自己当前的状态，即网络会对前面的信息进行记忆并作用于输出，因此能存储特征之间的相关性[12]。但 RNN 只能依次进行顺序计算，这种机制带来了两个问题：

(1) 当前时刻的计算依赖于前一时刻的计算结果，限制了模型的并行能力。

(2) 在计算过程中，间隔时间过长的信息会丢失，无法建立上下文的长期依赖。

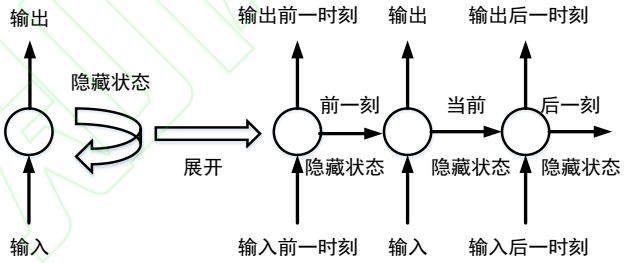

图 2 RNN 结构展开图

Fig.2 RNN structure expansion diagram

Transformer 的提出解决了上面两个问题：

(1) 模块之间并行化，提升了模型训练效率，符合现代分布式的 GPU 框架。

(2) 使用自注意力机制，将给定数据任意两个位置的距离建立联系，保留长距离信息。

1.1 编码器-解码器

Transformer 采用编码器-解码器（Encoder-Decoder）架构，由分别堆叠了 6 层的编码器和解码器组成，是一种避免循环的模型结构，如图 3 所示，输入的数据经过 6 层的编码器之后输出到每一层的解码器上计算注意力。

编码器每个层结构包含两个子层，多头注意力层（Multi-Head Attention）和前馈连接层（Feed Forward）。解码器有三个子层结构，遮掩多头注意力层（Masked Multi-Head Attention），多头注意力层(Multi-Head Attention)，前馈连接层（Feed Forward）。每个子层后面都加上残差连接（residual connection）和正则化层（layer normalization），结构如图 4 所示。

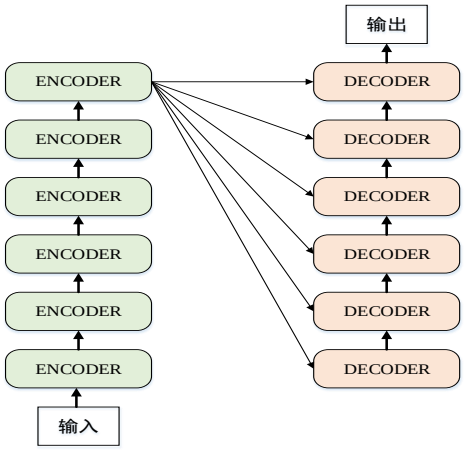

图 3 Encoder-Decoder 的 6 层结构

Fig.3 6-layer structure of Encoder-Decoder

从图 4 中可知，在解码器中多了一个遮掩多头注意力层（Masked Multi-Head Attention），由于前面编码器训练的数据长度不同，而解码器通常以数据的最大长度作为计算单元进行训练，并且只会受到之前数据对当前的影响，不需要后续数据进行参考，因此该层会遮掩掉当前位置之后的数据。

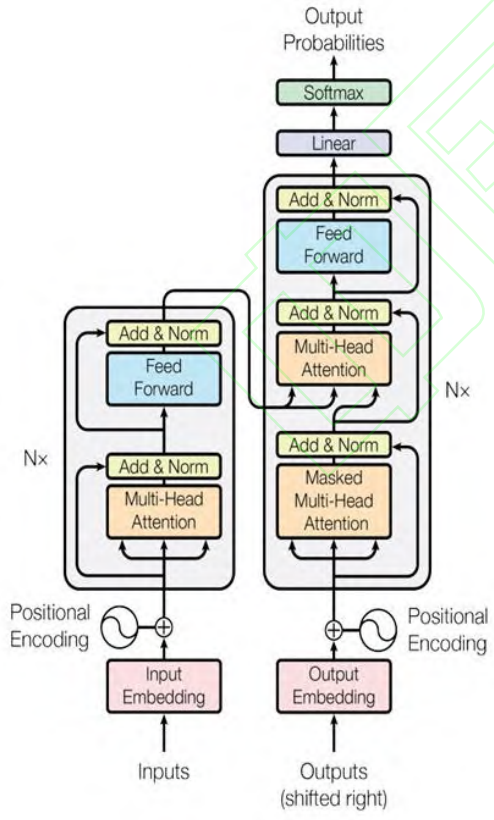

图 4 Transformer 模型结构

Fig.4 Transformer model structure

由于 Transformer 的计算抛弃了循环结构的递归和卷积，无法模拟文本中词语的位置信息，因而需要通过位置编码（Positional Encoding）进行人为添加。给句子中每个词的位置编号，每个编号对应一个向量，通过结合位置向量和词向量，给每个词引入一定的位置信息。文献[4]通过正弦/余弦函数引入位置编码，公式如（1）所示。

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d})$$
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d}) \quad (1)$$

其中，pos 表示单词在句子中的位置，d 表示位置编码的维度，2i 表示偶数的维度，2i+1 表示奇数维度 (即 2i≤d, 2i+1≤d)。

位置编码记录了序列数据之间顺序的相关性，相比较 RNN 顺序输入，Transformer 方法可以直接将数据并行输入，并存储数据之间的位置关系，大大提高了计算速度，减少了存储空间[13]。

此外，随着网络的加深，数据的分布会不断地变化。为了保证数据特征分布的稳定性，引入了层正则化（layer normalization），减少信息损失，使深度神经网络的训练更加顺利。

以机器翻译为例，Transformer 的工作流程如下：

Step1：将输入的句子通过单词嵌入算法转换为向量，使用位置编码获取单词的位置向量，两者相加获得模型的输入。

Step2：将 step1 中得到的单词向量矩阵传入编码器，经过多头注意力层，进入前馈神经网络，然后将输出向上传递到下一个编码器。

Stpe3：经过 6 个编码器后得到句子所有单词的编码信息矩阵，将矩阵分别传递到 6 个解码器，（此时解码器的输入来自编码器的输出和前一个解码器的输出）矩阵在每个编码器中依次经过遮掩多头注意力层，多头注意力层，前馈连接层。

Step4：将解码器的输出通过一个线性层之后由 Softmax 层转化为概率作为最终输出。

1.2 自注意力

注意力机制（Attention Mechanism）[14]模仿了生物观察行为的内部过程，增加部分区域观察精细度的机制。注意力机制可以快速提取稀疏数据的重要特征，因而被广泛应用于机器翻译[15]、语音识别[16]、图像处理[17]等领域。

注意力机制现在已成为神经网络领域的一个重要概念。其快速发展的原因主要有三个。首先，它是解决多任务较为先进的算法，其次被广泛用于提高神经网络的可解释性，第三有助于克服 RNN 中的一些挑战，如随着输入长度的增加导致性能下降，以及输入顺序不合理导致的计算效率低下。而自注意力机制（Self-attention Mechanism）[18]是注意力机制的改进，其减少了网络对外部信息的依赖，更擅长捕捉数据或特征内部的相关性。

Transformer 架构引入自注意力机制，避免在神经网

络中使用递归，完全依赖自注意力机制来绘制输入与输出之间的全局依赖。文献[4]中使用缩放点积注意力（Scaled Dot-Product Attention），相比一般的注意力，缩放点积注意力使用点积进行相似度计算，在实际中会更快更节省空间，基本结构如图 5 所示。在计算时，需要将输入通过线性变换得到矩阵 Q(查询)，K(键值)， V(值)，计算公式如（2）所示。

$$Attention(Q,K,V)=soft\max(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

其中，dk 是矩阵 Q,K 的列数，即向量维度。

以机器翻译为例，自注意力机制的计算过程如图 6 所示，其步骤如下：

Step1：创建 3 个向量。将输入中的每个单词线性投影到三个不同的空间中，从而产生三种新向量表示形式：查询（Query，Q），键值（Key，K），值（Value，V）。

Step2：计算得分。当我们在某个位置编码单词时，分数决定了对输入句子的其他单词的关联程度。以图 6 为例，假设计算第一个单词“thinking”的自注意力，需要根据这个单词对输入句子的每个单词进行评分。

Step3：除以缩放因子。Step2 中的评分除以缩放因子√dk（键向量维数的平方根），原始注意力值均聚集在得分最高的值，除以根号 dk，可起到缩放作用，分散注意力。

Step4：Softmax 函数标准化。Softmax 的分数决定了当前单词与句子中每个单词的相关程度。

Step5：将每个 V 向量乘以 Softmax 函数。保持对当前词关注度不变的情况下，降低不相关词的关注度。

Step6：累加权值向量。通过累加 Step5 中的向量，产生一个单词自注意力层的输出。

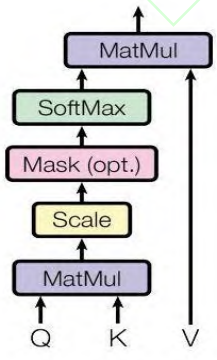

图 5　缩放点积注意力

Fig.5　Scaled Dot-Product Attention

通过自注意力机制，计算每个词和所有词之间的注意力，使得每个词都有全局的语义信息，并且可以捕获长距离依赖关系。

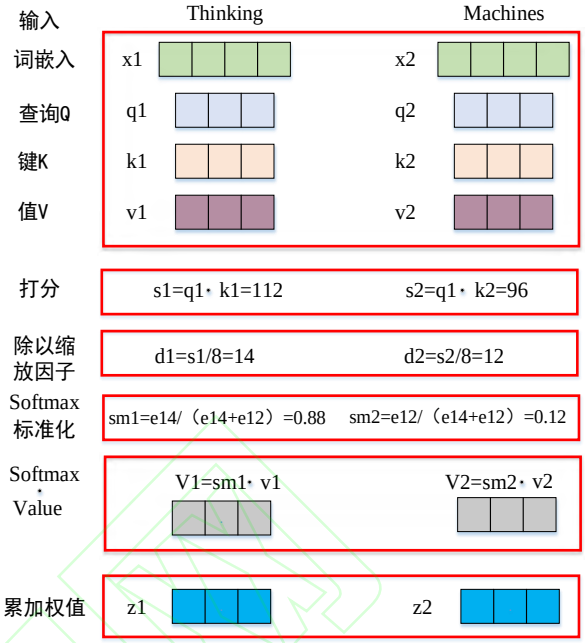

图 6　自注意力机制计算过程

Fig.6　Self-attention mechanism calculation process

在计算机视觉中，自注意力层以特征图为输入，计算每对特征之间的注意力权重，从而生成一个更新的特征图，其中每个位置都有关于同一图像中任何其他特征的信息。这些层可以直接代替卷积或与卷积层相结合，它们能够处理比常规卷积更大的感受野，因此能够获取空间上一些长距离间隔特征之间的依赖关系。

1.3　多头注意力

多头注意力机制的本质是在参数量总体不变的情况下，将查询、键、值三个参数进行多次拆分，每组拆分参数映射到高维空间的不同子空间中计算注意力权重，从而关注输入的不同部分。经过并行多次计算，最后合并所有子空间中的注意力信息，公式如（3）所示。

$$MultiHead(Q,K,V)=Concat(head_1,...,head_h)W^o \quad (3)$$
$$head_i=Attention(QW_i^Q,KW_i^K,VW_i^V)$$

其中，W^o,W_i^Q,W_i^K,W_i^V 为线性变换时的参数矩阵。

由于注意力在不同的子空间中分布不同，多头注意力实际上是寻找输入数据之间不同角度的关联，从而可以编码多个关系和细微的差别，其结构如图 7 所示。

多头注意力赋予了 Transformer 强大的结构，多个独立的头部关注不同的信息（如全局信息和局部信息），从而提取更加全面丰富的特征。

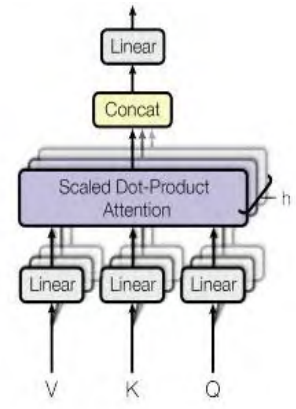

图 7　多头注意力结构

Fig.7　Multi-Head Attention

科，核心问题是研究如何对输入的图像或视频进行处理，使输出的图像或视频质量得到相当程度的改善，便于计算机对图像或视频进行分类，处理和识别。

受到文献[4]中 Transformer 架构使用自注意力机制来挖掘文本中的长距离依赖关系的启发，许多研究者提出将自注意力机制应用于计算机视觉任务，克服卷积的归纳偏置所带来的局限性，突破图像的感受野限制，计算像素与全部图像的关系，从而提取上下文的长距离依赖。

本章按照应用场景对视觉 Transformer 模型进行了分类，主要包括图像分类、目标检测、图像分割、识别任务、图像增强、图像生成和视频处理[19, 20]，分别列举了 Transformer 在视觉任务上的应用，如表 1 所示。

2　在计算机视觉领域的应用

计算机视觉是使计算机能够达到人类那样“看”的学

表 1　Transformer 在视觉任务应用一览表

Table 1　Transformer application list in visual tasks

类别	模型	重点
图像分类	iGPT[7]	像素序列、GPT-2、无监督学习
	ViT[8]	图像序列化、纯 Transformer 架构
	TNT[21]	嵌套结构、像素建模
	T2T-ViT[22]	Tokens-to-Token 机制、深窄结构
	LV-ViT[23]	Token Labeling 重新标注、CutMix 技术
	DeepViT[24]	再注意力机制、转移矩阵
目标检测	DETR[6]	集合预测、二分匹配
	Deformable DETR[25]	可形变卷积、多尺度注意力模块
	TSP[26]	纯编码器版本、新的二分匹配方案
	UP-DETR[27]	无监督预训练、随机查询补丁检测
	ACT[28]	局部敏感哈希方法、多任务知识蒸馏
图像分割	SETR[29]	序列到序列的预测任务
	Segmenter[30]	逐点线性解码器、掩码 Transformer 解码器
	Segformer[31]	MLP 解码器、编码器不同尺度信息
	max-deeplab[32]	全景率样式损失、双路径架构
	Vis-TR[33]	实例序列匹配与分割策略
识别任务	CVT[34]	注意选择性融合
	Poseformer[35]	时间和空间 Transformer 网络
	TransReID[36]	ViT-BOT 基线、SIE 模块、JPM 模块
	LSTR[37]	多项式参数模型、匈牙利损失
图像增强	IPT[38]	预训练、多任务
	TTSR[39]	高分辨率参考图像、联合学习
图像生成	Image Transformer[5]	自回归问题、像素生成
	VQGAN[9]	图像成分组成、百万像素图像
	TransGAN[40]	纯 Transformer 架构、多任务协同训练
视频处理	MEGA[41]	关联模块、长时记忆模块
	STTN[42]	联合时空变换网络、时空对抗训练、
	TimeSformer[43]	分割时间-空间注意力方案
	ConvTransformer[44]	高阶运动信息、中间插值帧和推断帧

2.1　图像分类

图像分类是根据图像的语义信息对不同类别图像进行区分并分配类别标签，是物体检测、图像分割、物体跟踪、行为分析等其他高层视觉任务的重要基础。受到 Transformer 在自然语言处理领域的成功启发，研究人员将 Transformer 迁移到图像方面，试图检验相似的模型是否可以学习更全面丰富的图像特征。典型的算法有 iGPT[7]和 ViT[8]系列。

2.1.1 iGPT

Chen 等[7]受 BERT[45]、GPT-2[10]等 Transformer 模型以及其变体在自然语言领域中无监督表征学习的影响，提出了 iGPT（image GPT），研究了 GPT-2 是否可以学习高质量的无监督图像表示。作者沿用了预先训练（pre-train）而后微调（fine-tune）的路线，并针对两者分别设计了两种不同的实验方式。首先将输入的二维图像分解为长像素序列，然后使用自回归（auto-regressive）和 BERT 目标（BERT objectives）两种方式预训练模型，最后利用线性探针（linear probe）或微调来评价预训练模型的优劣，通过合成分析，自动判定目标类别，无需人为标签的指导。

iGPT 能够理解并学习强大的图像特征，生成具有清晰可识别的物体样本，预训练得到的模型在后续任务上不弱于甚至超过监督学习的模型。实验结果表明，在 CIFAR-10 数据集上，使用线性探针实现了 96.3% 的准确度，优于有监督的 Wide ResNet[46]，并且通过完全微调实现了 99.0% 的准确度，与顶级监督预训练模型相匹配。当用像素替换 VQVAE 编码时，在 ImageNet 上与自监督的基准比较，实现了 69%的 Top-1 精度。但是该方法存在很大的局限性，由于使用了 GPT-2 模型，需要大量的计算才能达到有竞争力的效果。此外大多数的无监督方法可以处理高分辨率图像，而 iGPT 只能对低分辨率图像建模。因此，iGPT 更重要的意义是在概念上证明了 Transformer 可以无监督地学习图像特征表示。

2.1.2 ViT 及其改进算法

Dosovitskiy 等[8]首次将原始的 Transformer 模型应用于图像分类任务，提出了 ViT（Vision Transformer），一种完全基于自注意力机制的结构。作者认为在大规模数据集上，不依赖 CNN，Transformer 完全可以在分类任务中表现的很好，ViT 的框架如图 8 所示。

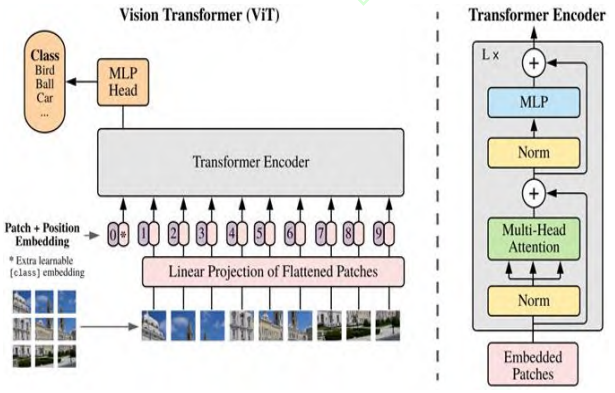

图 8 ViT 模型结构

Fig.8 ViT Structure

为了将图像转化成 Transformer 结构可以处理的序列数据，引入了图像块（patch）的概念。首先将二维图像做分块处理，每个图像块展平成一维向量，接着对每个向量进行线性投影变换，同时引入位置编码，加入序列的位置信息。此外在输入的序列数据之前添加了一个分类标志位（class），更好的表示全局信息。ViT 模型通常在大型数据集上预训练，针对较小的下游任务进行微调。在 ImageNet 数据集上，VIT-H/14 以 88.55% Top-1 的准确率超越了 EfficientNet 模型[47]，成功打破了基于卷积主导的网络在分类任务上面的垄断，比传统的 CNN 网络更具效率和可扩展性。

ViT 是 Transformer 在大规模数据集上替代标准卷积的第一部作品，为 Transformer 在计算机视觉任务的发展奠定了重要的基础。虽然它取得了突破性的进展，但缺点也十分明显，

（1）ViT 将输入图像切块并展平成向量，忽略了图像的特有性质，破坏了其内部固有的结构信息，导致学习效率不高，难以训练。

（2）ViT 所需的计算资源大，在 JFT 数据集上，ViT-L/16 的预训练样本达到 100M 时,准确率才会高于 BiT[48]。因此在有限的计算资源和数据的情况下，ViT 难以学到丰富的特征。

针对 ViT 的缺陷，Han 等[21]提出了 TNT（Transformer in Transformer），一种新型的基于结构嵌套的 Transformer 架构。通过内外两个 Transformer 联合，提取图像局部和全局的特征。具体而言，在每个 TNT 块中，外 Transformer 对图像块之间的关系进行建模，内 Transformer 对像素之间的关系进行建模，经过线性变换将像素级特征投影到图像块所在的空间中并与块信息相加。通过堆叠多个 TNT 块，形成 TNT 模型。通过这种嵌套方式，块特征可以更好的保持全局空间结构信息，像素特征可以保持局部信息，显著提高了模型的识别效果。

Yuan 等[22]提出了基于渐进式 Token 化机制的 T2T-ViT（Tokens-to-Token ViT），同时建模图像的局部结构信息与全局相关性。通过递归聚集相邻的对象逐步将图像结构化为序列组（tokens），继而连接成一个更长的序列（token）。这种渐进化机制不仅可以对局部信息建模，还能减少 token 序列的长度，降低模型的维度，减少计算量。同时为了增加特征的丰富性，借鉴 CNN 架构的设计思想提出了具有深窄结构（deep-narrow）的 ViT 骨干，减少了信息冗余，参数量和计算量显著降低。T2T-ViT 是以 ViT 为骨干网络的一次突破性探索，在标准 ImageNet 数据集上达到了 80.7%的 Top-1 精度，超越了模型大小相似的 ResNet50[49]，甚至比 MobileNet 系列[50、51]更加轻量化。

Jiang 等[23]提出了一种提高 ViT 性能的新的训练目标--Token Labeling，来探索 Transformer 在 ImageNet 分类中的潜力。作者将一张图片分成若干 patch，每个 patch 转化为 token，利用文献[52]中的 Re-labeling 技术，得到每个

token 的软标签（token-label），对图像进行重新标注，从而将图像分类问题转化为多个 token-label 识别问题。同时在训练模型时使用了 CutMix 技术，它能提高模型的性能和鲁棒性。Token Labeling 技术可以改善不同规模的 ViT 模型的性能，以具有 26M 可学习参数的视觉 Transformer 为例，可以在 ImageNet 上达到 84.4%的 Top-1 精度。

ViT 浅层在视觉任务上有良好的表现，一个很自然的问题被提出："Transformer 能否可以像 CNN 一样做的更深？"Zhou 等[24]加深了 VIT 模型层次，性能迅速饱和。通过研究发现，在 ViT 的深层，自注意力机制无法学习到有效的特征，特征图逐渐趋于相似，阻碍了模型获得预期的性能提升。因此作者提出了再注意力机制（Re-attention），解决了深层 ViT 架构的注意力坍塌（attention collapse）问题。在 ViT 的深层中，同一个序列在不同层之间的注意力图差别较小，但同一层不同的头之间差距明显。通过在每一层中加入一个转移矩阵，以一种可学习的方式交换来自不同注意力头的信息，从而再生注意力图。DeepViT 能够以可忽略的计算和存储成本重新生成注意图以增加其在不同层的多样性，并使得 ViT 模型的性能可以随着层数的加深而增加。

2.1.3 图像分类算法总结

目前，基于 Transformer 的图像分类研究大致可以分为 iGPT 和 ViT 系列，本小节对部分图像分类方法从参数量和 Top1 上的准确率进行了对比，如表 2 所示，"-"表示没有相关数据。此外，分析了数据集的大小对模型性能的影响，类比了 BiT 和 ViT 的多个变体，如图 9 所示。

iGPT 是 Transformer 首次直接应用于图像分类的模型，它本身不含卷积网络，并且在 CIFAR-10 和 CIFAR-100 数据集上优于如 ResNet 等 CNN 模型。但是它需要对大量数据进行训练，并且需要庞大的参数量来实现最佳结果，如表 2 所示，iGPT 在 CIFAR 数据集上的参数量达到 1362M，约是 ResNet-50 的 54 倍，而准确率却比 ResNet-50 低了 7.2%。ViT 可以获得与当前最优卷积网络相媲美的结果，其训练所需的计算资源大大减少。但是 ViT 也具有很大的局限性，首先，ViT 处理图片的方式不够好，无法建模图片的空间信息；其次，模型深度不够，无法像 CNN 一样扩大层数；最后，ViT 模型需要基于 CNN 模型的预训练，其效果很大程度上取决于预训练模型的结果，这些缺陷为后续的改进工作提供了诸多的思路。

此外，ViT 模型通常需要大量的数据进行训练，如图 9 所示，分别在 9M，30M，90M 和 300M 上训练 BiT 模型和 ViT 模型，从图上可以看出，在较小的数据集上 BiT 模型表现较好，随着训练数据量的增多，ViT 的准确率超过 BiT。因此，怎样提高 ViT 在小样本上的性能也是未来一个非常值得研究的方向。

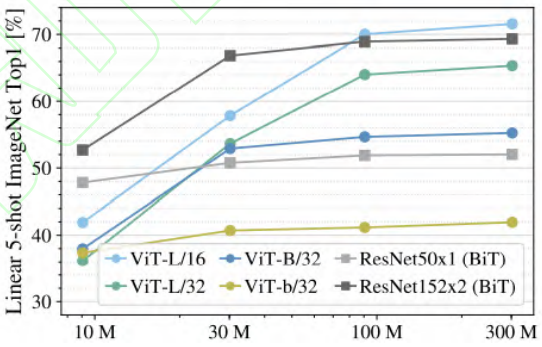

图 9 数据量对模型的影响

Fig.9 The impact of the amount of data on the model

表 2 部分图像分类方法对比

Table 2 Comparison of some Image Classification methods

算法		数据集	Params (M)	ACC（Top-1）
CNN based				
ResNet[49]	ResNet-50	ImageNet	25	76.2
	ResNet-110	CIFAR-100	—	72.8
	ResNet-152	ImageNet	60	78.3
EfficientNet[47]	EfficientNet-B5	ImageNet	30	83.6
	EfficientNet-B7		66	84.3
BiT[48]	BiT-L	CIFAR-100	60	93.51
	BiT-L	ImageNet	60	87.54
Transformer based				
iGPT[7]	GPT-2	ImageNet	1362	69.0
	GPT-2	CIFAR-10	1362	96.3
	GPT-2	CIFAR-100	1362	82.8
ViT 变体[8]	ViT-B/16	CIFAR-10	86	98.13
		CIFAR-100	86	87.1
		ImageNet	86	77.9
	VIT-H/14	CIFAR-10	632	99.5
		CIFAR-100	632	94.6
		ImageNet	632	88.6
ViT 改进	TNT-B[21]	ImageNe	66	82.8
	T2T-ViT[22]	ImageNet	22	80.7
	LV-ViT[23]	ImageNet	56	84.4
	DeepViT-L[24]	ImageNet	55	82.2

2.2 目标检测

目标检测任务是给定一张图像，找出其中所有目标的位置，并给出每个目标的具体类别。由于目标会出现在图像中的任何位置，形态大小各异，图像背景千差万别，诸多的干扰因素都使得目标检测在计算机视觉领域是一个具有挑战性的研究。目前，主流的深度学习目标检测算法主要分为双阶段检测算法和单阶段检测算法，分别以 R-CNN[53]系列和 YOLO[54]系列为代表。它们通常需要手动定义一些锚点或窗口[55]，将检测构建成分类或回归问题间接完成任务。

Transformer 在图像分类上有着良好的表现，研究人员将其扩展到目标检测领域，其中以 DETR 为典型代表，后续的一些目标检测算法几乎都是根据 DETR 进行的改进。

2.2.1 DETR 及其改进算法

Carion 等[6]重新设计了目标检测框架，构建了 Detection Transformer（DETR），一种基于 Transformer 的物体检测框架，总体结构如图 10 所示。DETR 首先使用 CNN 提取图像特征，将提取的特征与其位置编码相加传递到编码器中，然后将一组对象查询（object queries）和编码器的输出一起作为解码器的输入进行处理，最后解码器的每个输出传递到前馈网络（Feed Forward Network，FFN），独立地解码成框坐标和类标签，得到最终的预测。同时将检测结果与真实值（ground truth）进行基于匈牙利算法(Hungarian Algorithm)的二分图匹配计算损失。DETR 将检测视为集合预测问题，简化了目标检测的整体流程，根据对象和全局上下文的关系，直接并行输出最终的预测集，将需要手动设计的技巧如非极大值抑制（Non-Maximum Suppression，NMS）和锚点删除，实现了端到端的自动训练和学习。与许多其他检测算法不同，DETR 在概念上很简单，不需要专门的库。经过 COCO 数据集测试，DETR 的平均精确度（Average Precision，AP）为 42%，在速度和精度上都比 Faster-RCNN 高，这是第一次将 Transformer 用于目标检测领域。

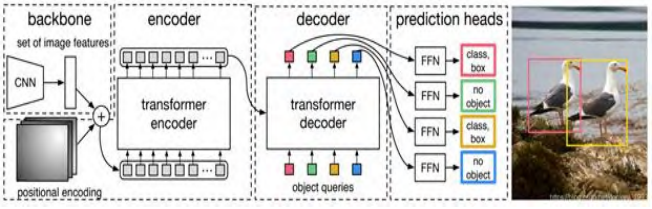

图 10 DETR 模型结构

Fig.10 DETR Structure

虽然 DETR 有良好的表现，但其仍存在两点局限性。一方面，在 COCO 数据集上，小目标的平均准确度（AP_S，AP for small objects）仅为 20.5%，检测效果较差。现有的检测器通常利用多尺度特征从高分辨率图像上检测小目标，但对于 DETR，高分辨率图像通常带来了极高的计算复杂度；另一方面，与现有的目标检测器相比，DETR 需要更长的训练时间来收敛。这两点缺陷实则是由 Transformer 的固有结构引起的。Transformer 初始化时，特征图上像素的注意力权重几乎分布均匀，所以需要更多的训练轮次来让注意力权重集中在稀疏有意义的位置。同时，注意力权重是基于像素个数的平方计算，这使得处理高分辨率特征图的计算复杂度和内存复杂度都非常高。

Zhu 等[25]针对 DETR 的缺陷，提出了可形变 DETR（Deformable DETR），融合了可变形卷积[56]（Deformable Convolution）良好的稀疏空间采样优势和 Transformer 强大的关系建模能力。可变形卷积将固定形状的卷积改造成了能适应物体形状的可变卷积，从而使结构适应物体形变的能力更强。在 Deformable DETR 中，作者使用了可变形注意力模块替换 Transformer 的注意力模块来处理特征图，对所有的特征图像素点进行预筛，只关注参考点周围少量的关键采样点，而不考虑特征图的空间大小，大大降低了计算复杂度，缓解了收敛性和特征空间分辨率问题。该模块无需 FPN[57]的帮助，对多个分辨率特征图统一计算注意力，实现了不同尺度信息的自动融合，能天然地扩展到聚合多尺度特征上。在 COCO 数据集上，相比 DETR，Deformable DETR 的训练周期少了 10 倍，特别是在小目标检测上提升了 5.9%AP_S。

Sun 等[26]研究了 DETR 模型收敛慢的问题，并分析了 Transformer 解码器中交叉注意模块的瓶颈，提出了 DETR 的纯编码器版本（encoder-only）。交叉注意力模块是解码器中的对象查询从编码器中获取目标信息的关键模块，不精确的交叉注意力可能导致解码器从图像中提取不够准确的上下文信息。作者通过计算交叉注意力图负熵（用来表示注意力图的稀疏性）的变化，发现随着交叉注意力的持续增加，任何一层的稀疏性都在一直变化并没有趋于稳定，因此，认为交叉注意力部分是影响 DETR 模型优化的主要原因，在直接去掉了 Transformer 的解码器部分后，检测精度和训练收敛性上都有了很大的提高。此外，设计了一种新的二分匹配方案，在 FCOS 网络[58]的基础上改良了两个模型，即 TSP-FCOS 和 TSP-RCNN，改进了纯编码器 DETR，比原始的 DETR 具备更好的性能。

Dai 等[27]受 NLP 领域无监督预训练的成功启发，提出了 UP-DETR（Unsupervised Pre-training DETR）。设计了一种“随机查询块检测（random query patch detection）”的预训练任务，在没有任何人工注释的情况下，对 DETR 中的 Transformer 进行无监督预训练。具体而言，将原图输入编码器，从原图中随机裁剪若干个图像块输入解码器，从原始图像中检测出这些图像块。经过预训练之后的 DETR，当输入一个图像块时，网络就可以定位到这个

块在图像中的位置。同时，引入了冻结预训练的 CNN 网络和块特征重构分支，以保留 Transformer 的特征判别能力。此外，为了同时支持多查询块定位，提出了注意力掩码和对象查询洗牌机制，解决了查询块和对象查询之间的分配问题。UP-DETR 在 ImageNet 数据集上预训练，在 PASCAL VOC 和 COCO 数据集上微调，训练 150 个轮次后，UP-DETR 的 AP 分别为 56.1%和 40.5%，较 DETR 提升了 6.2%AP 和 0.8%AP，且收敛速度更快。

Zheng 等[28]针对 DETR 的高计算复杂性问题，提出了 ACT（Adaptive Clustering Transformer），一种自适应聚类的 Transformer 变体，无需任何训练过程，降低了 DETR 的推理成本，实现了准确率和计算成本之间的良好平衡。ACT 使用局部敏感哈希（Locality Sensitivity Hashing，LSH）自适应的对查询特征进行聚类，并将原型键交互近似为查询键交互。ACT 是可嵌入式模块，无需任何重新训练即可代替 DETR 中的自注意力模块，与原始的 Transformer 完全兼容。通过实验，DETR 的计算量（floating point operations，FLOPS）从 73.4 降低到 58.2，而 AP 的损失仅为 0.7%。此外，作者同时提出了多任务知识蒸馏（Multi-Task Knowledge Distillation，MTKD），该方法利用原始的 Transformer 通过少许的微调来蒸馏 ACT 模块，进一步将 AP 的损失降低到 0.2%，实现 ACT 和原始 Transformer 之间的无缝转换。

2.2.2 目标检测算法总结

与基于 CNN 的目标检测算法相比，基于 Transformer 骨干网络的方法无论是在精度还是运行速度上都表现出了强大的性能，最典型的模型是 DETR。

DETR 架构从根本上进行了改变，这是第一个将 Transformer 成功整合为端到端训练的目标检测框架。在性能上，DETR 可以媲美当前的最先进的方法（state of the art，SOTA）方法，但架构得到了极大简化。在 COCO 数据集上将 DETR 与 Faster R-CNN 基线方法进行了对比，如表 3 所示。结果发现 DETR 在大型目标上的检测性能要优于 Faster R-CNN，但在小目标的检测上性能不如后者，另外需要更长的训练时间来收敛，这也为今后 DETR 的改进提供了新的方向。

表 3 目标识别算法性能比较

Table 3 Performance comparison of target recognition algorithms

算法	GFLOPS/FPS	参数(M)	AP	AP50	AP75	APS	APM	APL
CNN based								
Faster RCNN-FPN[53]	180/26	42	42.0	62.1	45.5	26.6	45.4	53.4
Transformer based								
DETR[6]	86/ 28	41	42.0	62.4	44.2	20.5	45.8	61.1
Deformable DETR[25]	173/19	40	46.2	65.2	50.0	28.8	49.2	61.7
TSP-FCOS[26]	189/20	-	43.1	62.3	47.0	26.6	46.8	55.9
TSP-RCNN[26]	188/15	-	45.0	64.5	49.6	29.7	47.7	58.0
UP-DETR[27]	-/-	41	42.8	63.0	45.3	20.8	47.1	61.7
ACT+MKKD (L=32)[28]	169/14	-	43.1	-	-	61.4	47.1	22.2

2.3 图像分割

图像分割是根据某些规则把图片中的像素分成不同的部分（加不同的标签），可以看作是图像分类从图像层次到像素级的延伸。图像分割包含语义分割、全景分割、实例分割等子任务，为理解完整的现实场景铺平了道路，是非常重要且基础的研究方向。

此前，图像分割大多数是基于全卷积网络（Fully Convolutional Network，FCN）[59]体系结构进行的改进。FCN 通过多次下采样以牺牲空间分辨率为代价来抽取局部或全局特征，网络层固定后，每一层的感受野是受限的，感受野的大小决定了特征是否能捕获更大范围的周边信息甚至是全局信息。因此，如何构造更大的感受野进行上下文建模，达到既能抽取特征信息又尽量不损失空间分辨率，一直是图像分割的难点。由于图像分类和分割之间有着密切的联系，因此许多分割算法将 ViT 作为骨干网络，以 SETR[29]、Segmenter[30]和 Segformer[31]为典型代表。

2.3.1 SETR

Zheng 等[29]为语义分割方法设计了一个新的视角，提出了 SETR（Segmentation Transformer），将语义分割转变为序列到序列的预测任务，摒弃了模型需要通过降低分辨率来学习局部到全局的特征。SETR 部署了一个纯 Transformer（即不进行卷积和降低分辨率），借鉴 ViT[8]模型，首先将图像分解为若干固定大小的块，进行线性变换，将每个块的像素向量和位置编码相加作为编码器的输入。经过 24 层的 Transformer 学习得到图像的全局特征，最后使用解码器恢复原始图像的分辨率。此外，作者设计了三种复杂度不同的解码器结构，对自注意力进行了更深入的研究。SETR 在空间分辨率上没有进行下采样，而是在编码器 Transformer 的每一层进行全局上下文建模，完全用注意力机制实现了编码器的功能。实验表明，与现有的基于 FCN 的方法相比，SETR 可以学习到更好的特征表示。在 ADE20K 数据集上，SETR 单尺度的推理在均交并比（mean Intersection over Union，mIoU）指标上达到了 48.64%，比 ACNET[60]方法提升了 2.74%。

2.3.2 Segmenter

Strudel 等[30]基于 ViT 的研究成果，提出了 Segmenter，

一种用于语义分割的转换器模型。图像分割在单个图像块级别通常是不明确的，并且需要上下文信息来达成标签共识。Segmenter 在编码阶段采用了 ViT 模型结构，将图像分割成块，并进行线性映射，经过编码器处理后输出嵌入序列。在解码阶段引入可学习类别嵌入，将编码器的输出与类别嵌入一起送进解码器，这里使用逐点线性解码器（point-wise linear decoder）或掩码 Transformer 解码器（mask Transformer decoder），从而获得类标签，经过 softmax 及上采样等一系列的操作后输出最终的像素分割图。作者在图像分类上预训练模型，在语义分割上进行微调，通过实验发现逐点线性解码器可以获得不错的效果，使用类掩码 Transformer 解码器可以进一步提高 0.2%-1.22%mIoU。

2.3.3 SegFormer

Xie 等[31]提出了 SegFormer，一种简单、高效但功能强大的语义分割框架，它将 Transformer 与轻量级多层感知器 (MLP) 解码器相结合。SegFormer 使用一种分层特征表示的方法，编码阶段每个 transformer 层的输出特征尺寸逐层递减，通过这种方式捕获不同尺度的特征信息，同时舍弃了 ViT 中的位置嵌入，避免了测试图像与训练图像尺寸不同而导致模型性能下降的问题。所提出的 MLP 解码器（Lightweight All-MLP Decoder）采用简单的 MLP 结构，聚合编码器层不同尺度的特征，从而融合了局部注意力和全局注意力，并证明这些简单和轻量级的设计是在 Transformers 上进行高效分割的关键。在 ADE20K 数据集上以 64M 的参数实现了 51.8% mIoU，比 SETR 参数量减少了 4 倍，mIoU 提高了 1.6%。此外，SegFormer 比现有方法对常见的腐蚀和扰动更为鲁棒。

2.3.4 MaX-DeepLab

Wang 等[32]受 DETR 的启发，提出了 MaX-DeepLab，简化了依赖子任务和手动设计的组件，是第一个用于全景分割的端到端模型。该模型直接预测一组不重叠的掩码及其对应的语义标签，并通过使用全景质量（Panoptic Quality，PQ）样式进行目标优化，输出掩码和类别。MaX-DeepLab 采用双路径架构，除了 CNN 路径外，还引入全局内存路径，使 CNN 可以在任何层上读写全局内存，从而提供了一种将 Transformer 与 CNN 结合的新方法。MaX-DeepLab 在不增加测试时间的情况下，在 COCO 测试集上实现了最新的 51.3%PQ。

2.3.5 VisTR

Wang 等[33]提出了一种新的视频实例分割框架 VisTR（Video Instance Segmentation Transformer），它将视频实例分割任务建模为一个端到端的并行序列的解码、预测问题，其核心是一种高效的实例序列匹配与分割策略。给定一个由多个图像组成的视频片段作为输入，VisTR 从相似性学习的新角度，在序列级别上对实例进行整体监控和分段，最后直接按顺序输出视频中每个实例的掩码序列，在相同的实例分割框架下，可以无缝、自然地实现实例跟踪，大大简化了视频实例分割的流程，与现有的方法大不相同。实验表明，在 YouTube-VIS 数据集上，使用 ResNet-50 相同的主干，VisTR 比 MaskTrack R-CNN[61]的精确度提升了 3.8%AP，在不考虑数据加载过程时，速度可以到达 57.7fps。在使用单一模型的方法中速度和精确度都是最优的。

2.3.6 语义分割算法总结

由于图像分类、目标检测和分割有着密切的联系，因此，目前分割任务中的算法也是对 ViT 和 DETR 的延伸与改进。如 SETR 是第一个尝试将 ViT 引入语义分割领域，并取得了不错的效果，迈出了重要的一步，但是 SETR 将 ViT 作为骨干网络仍存在一些问题，

（1）ViT 是柱状结构，输出分辨率低且单一，但语义分割对像素的分类及边缘等轮廓细节要求比较精细；

（2）ViT 使用固定的位置编码，但在语义分割测试时往往图片的分辨率不固定，要么对位置编码进行线性插值，这会损害性能，要么做固定分辨率的滑动窗口，这样效率很低且不灵活。

因此后续工作可以针对以上两点进行进一步的改进。表 4 总结了语义分割算法在多个数据集上的检测精度（mAP）。基于 Transformers 的模型总体比基于 CNN 的模型表现要好，在不同的数据集上都要优于基准线甚至优于最新的 CNN 网络模型。

表 4 语义分割算法性能比较

Table 4 Performance comparison of target recognition algorithms

算法	骨干网络	mIoU			
		ADE20K	Pascal Context	Cityscapes	COCO
FCN[59]	ResNet-101	41.4	45.74	76.61	-
ACNet[60]	ResNet-101	45.9	54.1	-	-
SETR[29]	ViT	50.28	55.83	82.15	-
Segmenter[30]	ViT	50.77	80.7	55.6	-
SegFormer[31]	MiT	51.8	-	83.1	46.7

2.4 识别任务

识别任务是一个综合性的任务，它囊括了视觉领域的多种技术，如图像分类、目标检测、语义分割、多实例匹配、部件与整体关系的学习、行为推理和时空关系

等。本节从面部表情识别、姿态估计、行人重识别和车道线检测四个方面分别列举了目前基于 Transformer 的典型算法。

2.4.1 CVT

面部表情识别（Facial Expression Recognition，FER）随着人脸识别的研究而发展，在过去几十年取得了实质性的进展，但以前的研究主要是在实验室收集的数据集上实现的，现实世界中的遮挡、头部姿势变化和复杂的背景无疑增加了表情识别的难度。

Ma 等[34]提出了 CVT（Convolutional Visual Transformers），认为将人脸图像转换为视觉单词序列并从全局角度执行表情识别是可行的，设计了一种注意选择性融合（Attentional Selective Fusion，ASF）方法，来汇总全局和局部面部信息，引导主干提取所需要的信息，以端到端的方式压缩无用信息。此外，对具有全局自注意力的这些视觉单词之间的关系进行建模，使整个网络能够从全局角度学习特征序列之间的关系，从而忽略信息不足的区域。这是首次将 Transformer 应用于面部表情识别，CVT 在 RAF-DB 数据集上的正确率达到了 88.14%，比之前 SOTA 的 SCN[62]方法提升了 1.11%。

2.4.2 PoseFormer

人体姿态估计（PoseEstimation）是从输入的图像或视频中定位人体部位并建立人体表征（如人体骨骼）。近年来受到了广泛的关注，并已被应用与人机交互、运动分析、增强现实和虚拟现实等任务中。目前在人体姿态估计领域，卷积结构仍占主导。

Zheng 等[35]提出了 PoseFormer，设计了一种不含卷积的时空 Transformer 结构，用于视频中的 3D 人体姿态估计。PoseFormer 使用两个维度不同的 Transformer 模块直接对时间和空间进行建模。具体来说，首先构建了一个空间模块，提取每帧中二维骨架关键点之间的关节联系，空间自注意力层会考虑二维关节的位置信息并返回该帧的潜在特征表示。然后构建一个时间模块，分析每个空间特征之间的全局依赖关系，并捕捉多帧输入的时间相关性，最后输出一个精确的三维人体姿态中心帧。在 Human3.6M 数据集上，PoseFormer 模型在评估关键点位置时，产生 44.3mm 的最低 MPJPE（Mean Per Joint Postion Error），与 METRO[63]相比，MPJPE 降低了大约 18%，此外 METRO 中忽略了时间一致性，这限制了其姿态估计的稳健性。PoseFormer 可以形象化地估算出 3D 姿态，并产生更平滑可靠的结果，甚至在户外、快速移动和高遮挡的情况下均能达到不错的效果。

2.4.3 TransReID

重识别（Re-identification，ReID）是从给定图像或视频中判断是否存在特定对象的技术。例如在监控视频中，由于相机分辨率和拍摄角度的缘故，人脸识别有时会失效，行人重识别就成了非常重要的辅助技术。

He 等[36]提出了 TransReID（Transformer-based Object Re-Identification），第一个使用纯 Transformer 进行对象重识别的研究。受 Bag of Tricks（BoT）[64]的启发，以 ViT 模型作为特征提取主干，构建了一个强大的基准模型 ViT-BOT，它在几个 ReID 基准中取得了与基于 CNN 框架相当的结果。作者考虑到 ReID 数据的特殊性，设计了 SIE（Side Information Embedding）模块，通过向量投影来编码不同类型的边界信息，以消除由各种相机参数或视角等非可视信息导致的特征偏差。将 ViT-BOT 的最后一层调整为双并行分支结构，设计了 Jigsaw 分支，与全局分支平衡。在 Jigsaw 分支中构建了一个 JPM (Jigsaw Patch Module)模块，通过对图像块打乱重组，使模型适应扰动，新构建的块中依然包含全局信息，从而学习更鲁棒的特征表达。在行人重识别方面，TransReID 在数据集 MSMT17 上比之前最先进的 ABDNet[65]方法提升了 8.6%mAP（mean average precision）；在遮挡重识别方面，在 Occluded-Duke 数据集上在实现了 55.7% 的 mAP，与 PGFA[66]方法相比，提升了约 11.9% mAP；车辆重新识别方面，在 Veri-776 数据集上，TransReID* 达到 81.7% mAP，超过 SAVER [67]2.1% mAP。

2.4.4 LSTR

车道线检测是将车道识别为近似曲线的过程，被广泛应用于自动驾驶汽车的车道线偏离警告和自适应巡航控制。目前，传统的车道检测算法通常首先生成分割结果然后采用后处理，这使得在学习全局上下文和车道的细长结构方面效率很低且存在缺陷。

Liu 等[37]提出了 LSTR（Lane Shape Transformers），一种可以直接输出车道形状模型参数的端到端方法。车道模型借助道路结构和摄像头内参的设定，采用多项式参数模型来描述车道线，为网络输出的参数提供了物理解释。开发了基于 Transformer 的网络，利用自注意力机制对非局部交互进行建模，从任何成对的视觉特征中总结信息，使其能学习丰富的结构和上下文信息。整个结构快速预测输出，并采用匈牙利拟合损失在预测参数和车道真值之间进行匹配，保证一对一的无序分配，利用匹配结果优化路径相关的回归损失，使模型消除了显性的非极大抑制过程。在 TuSimple 基准中，相比 PolyLane Net[68]方法，LSTR 的准确度提升 2.82%，参数量减少 5 倍，运行速度提高了 3.6 倍；与最先进的 Line-CNN[69]相比，LSTR 的准确率仅低 0.69%，但是运行速度比它快 14 倍。此外，在具有挑战性的自收集车道线检测数据集中显示出了出色的适应性。

2.5 图像增强

图像增强是图像处理中一种常用的技术，它的目的是增强图像中全局或局部有用的信息。合理利用图像增

强技术能够针对性地增强图像中感兴趣的特征，抑制不感兴趣的特征，有效的改善图像质量。

2.5.1 IPT

Chen 等[38]提出了 IPT（Image Processing Transformer）预训练模型，完成超分辨率、降噪、去雨等低级视觉任务。由图 11 可知，IPT 框架由多头结构、编码器、解码器和多尾结构组成。首先图像经过头结构变换为特征图，进行分块与展平，将每个特征向量等同于一个单词送入 Transformer 进行处理。经过整型与拼接，还原为与输入相同维度的特征图，并通过尾结构解码为目标图像。IPT 的多个头结构与尾结构负责维度变换，不同的任务共享同一个 Transformer 模块，只需要增加新的头结构与尾结构即可，这使得多任务的扩展变得简单。为了更好地适应不同的图像处理任务，研究者根据特征块之间的相关性引入了对比学习方法作为自监督损失函数，使来自于同一图像的特征块相互接近，不同图像的特征块远离。经过预训练的 IPT 模型，只需要在特征任务的数据集上进行微调即可达到很好的效果。在微调阶段，只有特定任务相关的头尾结构被激活训练，无关的模块暂时被冻结。对于不同倍率的超分辨率任务，IPT 在 Urban100 数据集上，相较于其他方法普遍能够提升 0.4dB，而对于去噪和去雨任务则提升了 1.6-2.0dB。

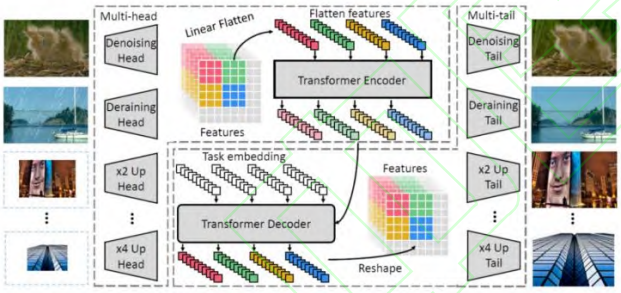

图 11 IPT 模型结构

Fig.11 IPT Structure

2.5.2 TTSR

超分辨率技术（Super-Resolution，SR），即从低分辨率图像（Low Resolution，LR)中恢复出自然、清晰的高分辨率图像(High Resolution，HR)。Yang 等[39]提出了 TTSR（Texture Transformer Network for Image Super），一种新颖的超分辨率纹理 Transformer 网络。TTSR 包含四个紧密相关的模块，可学习纹理提取器（Learnable Texture Extractor，LTE）是一个浅层的神经网络，在训练过程中不断更新自己的参数，以便于提取到最合适的纹理信息；相关性嵌入模块（Relevance Embedding module，RE）用来建立低分辨率输入图像与参考图像之间的关系，输出一个硬注意力图和软注意力图；硬注意力模块（Hard-Attention module for feature transfer，HA）利用硬注意力图中所记录的位置，从原参考图像的纹理信息中迁移对应位置的特征块，组合成迁移纹理特征图，随后与骨干网络中的特征进行通道级联，并通过一个卷积层得到融合的特征；软注意力模块（Soft-Attention module for feature synthesis，SA）将融合特征与软注意力图进行对应位置的点乘，使迁移过来的高频纹理特征得到更准确的利用。TTSR 通过引入一张高分辨率参考图像来指引整个超分辨率的计算过程，将复杂的图像纹理恢复过程转化为简单的纹理迁移。这种设计鼓励低分辨率图像和参考图像之间进行联合特征学习，通过注意力发现深度特征的对应关系，从而实现纹理特征的准确迁移。该模型可以最大程度地利用参考图像，有效地搜索与高清的迁移纹理信息，解决纹理模糊和失真的问题。从定量和定性的角度，TTSR 在 Sun80 和 Manga109 数据集上明显优于 SISR 和 RefSR 方法。

2.6 图像生成

图像生成是指根据输入向量（随机噪声或指定的条件向量）生成目标图像，这就需要一个能够理解图像全局组件的模型，使其具有局部真实和全局一致的特性。当前的图像生成任务主要是借助生成对抗网络（Generative Adversarial Network，GAN）[70]来实现。GAN 通常依赖于生成器（generator）和鉴别器（discriminator）。但自然图像的生成门槛较高，GAN 的训练具有较高的不稳定性，且优化困难，还可能导致特征分辨率和精细细节的损失（例如图像模糊）。

2.6.1 Image Transformer

Parmar 等[5]受卷积神经网络的启发，迈出了 Transformer 模型到图像转换和生成任务的第一步，提出了 Image Transformer。其采用了一种图像生成公式，类似于 Transformer 框架内的序列建模，由两部分组成：用于提取图像表示的编码器和用于生成像素的解码器。将图像生成视为一个自回归问题，即每个新像素的生成只考虑图像中已知的像素值，在每次特征生成中，自注意力都将各个特征块作为上下文，生成未知像素值表示。但是该模型有个明显的缺点，只重点关注局部注意范围，图像的生成依赖于每个像素其周围的取值，一次只能执行一步，并且要以失去全局接受域为代价，增加了存储和计算成本。

2.6.2 VQGAN

Esser 等[9]提出了 VQGAN（Vector Quantised Generative Adversarial Network），结合了 CNN 的有效归纳偏置和 Transformer 的表达能力，这是第一个由语义引导生成百万像素图像的 Transformer 架构。作者使用 CNN 架构对图像成分进行建模，Transformer 架构对成分进行合成，将图像表示为由感知丰富的图像成分组成，克服了直接在像素空间中对图像进行建模时难以实现的平方级复杂度，充分挖掘了两者互补优势的潜力。此外，作者使用

对抗性的方法来确保局部字典捕获感知上重要的结构，减少 Transformer 架构建模低层统计的需要，专注于建模远程关系的独特优势，从而生成高分辨率图像。VQGAN 无需重新学习图像局部结构中已知的、规律性的所有知识，在有效编码归纳偏置的同时，保持了 Transformer 的灵活性。在 CIFAR10 数据集上，VQGAN 比直接在像素空间中建模的 Transformer 方法 FID（Fréchet Inception Distance）分数提高了 18.63%，图像采样速度提高了 14.08 倍。

2.6.3 TransGAN

“对于更加困难的视觉任务，相比于生成对抗网络，Transformer 的表现又如何？”Jiang 等[40]怀着这样的疑问进行了一次试验性的研究，构建了一个只使用纯 Transformer 架构，完全没卷积的 TransGAN（Transformer Generative Adversarial Network）。从结构上来看，TransGAN 包括两个部分：一个是内存友好的生成器，该生成器可以逐步提高特征分辨率，同时减少每个阶段的嵌入维数；另一个是 patch 级判别器，将图像块而不是像素作为输入，并在真实图像和生成图像之间进行分类。同时，使用多任务协同训练策略以及本地初始化自注意力机制来增强自然图像的邻域平滑度（提高图像平滑，减少模糊）。实验结果表明，TransGAN 在 CIFAR10 数据集上得到具有竞争力的 IS （Inception Score）8.63 和 FID 11.89 ，但略逊于 StyleGAN v2[71]；在规模更大、分辨率更高的 STL-10 基准上，IS 为 10.10，FID 为 25.32，优于当前所有基于卷积网络的 GAN 模型。因此得出结论，纯 Transformer 有足够的能力应对困难的图像生成任务。

2.7 视频处理

Transformer 是一个非常有潜力的结构，可以统一不同模态的数据和不同的视觉任务。Transformer 所具有长依赖的优势，可以对时序信息进行建模，其核心自注意力机制，能够基于帧与帧之间的相似性，不断地学习，更新。所以，Transformer 非常适用于视频相关的任务。

2.7.1 MEGA

针对视频检测物体任务中相机失焦、物体遮挡等问题，Chen 等[41]提出了基于记忆增强的全局-局部整合网络（Memory Enhanced Global-Local Aggregation, MEGA）。作者认为在视频中检测物体可以利用时序信息来辅助质量较差的帧，设计了一个简洁的基础模块，使用区域候选网络从关键帧的相邻帧和全局帧中生成一些候选区域，使用关联模块（relation module）将全局帧中候选区域对应的特征整合到局部帧的候选区域的特征中，局部帧内部经过若干层关联模块得到同时包含全局和局部信息的关键帧特征。此外，设计了一个长时记忆模块(Long Range Memory，LRM)，将某一帧检测后的特征保存下来，并在下一帧的检测中使用该特征来辅助检测。在只增加非常小的计算开销的前提下，整合大量的全局信息和局部信息来辅助关键帧的检测，从而显著地提升了视频物体检测器的性能。实验结果表明，在 ImageNet VID 数据集上，MEGA 以 ResNet-101 和 ResNeXt-101 作为骨干网络的 mAP 分别达到了 82.9%和 84.1%，取得了该数据集上至今最佳效果。

2.7.2 STTN

视频修复是一项旨在视频帧缺失区域中填补合理内容的任务。Zeng 等[42]提出了一种用于视频修复的联合时空变换网络（Spatial-Temporal Transformer Network，STTN）。作者将视频修复描述为一个“多到多”的映射问题，以相邻帧和远帧作为输入，通过一个基于多尺度的注意力模块沿着空间和时间维度进行搜索，从所有帧中提取不同尺度的块，以覆盖由复杂运动引起的不同外观变化。Transformer 的不同头部计算不同尺度上空间块的相似性，从而检测出缺少区域并为其转换最相似的块。此外，在联合域的优化中引入时空对抗训练，以学习感知良好的、连贯的视频内容。使用固定掩码和移动掩码进行定量和定性评估以模拟现实世界的应用（例如水印去除和对象去除）。STTN 可以填充移动物体后面的缺失像素，并以更高的准确性和更少的模糊度重建整个视频，在 YouTube-VOS 和 DAVIS 数据集上，峰值信噪比 (Peak Signal to Noise Ratio ，PSNR)、流扭曲误差（flowwarping error）和 VFID（video-based Fréchet Inception Distance）三个指标上相对基线分别提升了 2.4%、1.3%和 19.7%。

2.7.3 TimeSformer

为了训练和理解模型，目前最好的 3D CNN 只能使用几秒长的视频片段。Bertasius 等[43]提出了一个新型视频理解架构 TimeSformer（Time-Space Transformer），它是首个无卷积，完全基于 Transformer 的视频架构。作者提出了分割时间-空间注意力方案，将输入视频分解为一组不重叠的图像块，通过应用时间注意力，使得每个块只与其他帧中相同空间位置的块进行比较。应用空间注意力，使得每个块仅与同一帧内的块比较，避免了所有成对块之间详尽的比较，降低了计算成本，提高了训练速度。该模型通过将每个块与视频中其他块进行显示比较来捕获每个块的语义，进而捕获相邻块之间的短程依赖性以及长距离块之间的远程关联。作者测量了 Kinetics-400 中 20K 的验证视频的实际推理运行时间，在成本相当的情况下，SlowFast[72]需要 14.88 个小时完成推理，TimeSformer 需要 36 分钟，运行时间要低的多。此外，TimeSformer 的可扩展性使其能在更长的视频片段上训练，来执行超远程时域建模，对于相同的单个剪辑覆盖范围，TimeSformer 在 HowTo100M 数据集上 Top-1 的正确率比 SlowFast 高 8%-11%，距离越远，效果越佳，

极大地促进了机器理解视频中复杂长动作的研究。

2.7.4 ConvTransformer

Liu 等[44]提出了一种新型的端到端架构，称为卷积 Transformer（ConvTransformer），用于视频帧序列学习和视频帧合成。研究者表示这是在视频合成方面卷积神经网络与 Transformer 的首度结合。ConvTransformer 将视频帧合成简化为一个编码器和解码器的问题，通过提出的多头卷积自注意机制，提取视频序列中存在的高阶运动信息，并将其用于合成目标插值帧。首先使用基于多头卷积自注意力层的编码器将输入的视频帧映射成特征图序列，然后使用解码器从特征图序列中对目标合成帧进行解码，解码后的特征图最终通过综合前馈网络生成中间插值帧或推断帧。实验证明，在下一帧推断任务中，ConvTransformer 在 Vimeo90K 数据集上，PSNR 值比 DVF[73]和 MCNet[74]模型高 2.7140dB 和 1.8983dB。ConvTransformer 可以有效的对视频帧中长序列的依赖性进行建模，然后推断出高质量的未来帧。

3 应用展望

Transformer 突破了 RNN 模型不能并行计算的限制，克服了 CNN 模型无法建模长距离依赖的缺点，通过自注意力机制，使模型更具可解释性。

在计算机视觉领域，现有的 Transformer 模型通常是从自然语言处理领域迁移过来，根据不同的视觉任务做了一些初步的探索，大致可以分为两类：一类是将自注意力机制与常见的 CNN 架构结合，另一类是用自注意力机制完全代替卷积。随着 Transformer 结构在越来越多的视觉任务中应用，有人不禁要问“在视觉领域，Transformer 会不会像在自然语言处理中代替 RNN 那样完全取代 CNN 吗？”。就目前的研究来看，Transformer 结构有其巨大的优势，但其缺点也十分明显。首先，Transformer 模型缺乏归纳偏置能力，并不具备卷积的平移不变性和局部敏感性，因此在数据不足时，不能很好地泛化任务。其次，无法处理高分辨率特征图，会使图像中的小目标丢失。最后，Transformer 结构是顺序无关的，会丢失输入数据的位置信息。尽管许多研究中将位置编码嵌入输入的特征向量中，但并没有改变其结构上的固有缺陷。因此，本文认为应该对 CNN 与 Transformer 取长补短，相互融合，而不存在取代关系。

Transformer 激起了计算机视觉领域各个方向的热潮，基于目前的研究现状，对未来的研究方向进行展望。

（1）冗余性问题。在机器翻译中，输入是对应的单词，但在视觉任务中，通常输入的是被分块之后的图像，由于图像具备局部相关性，相邻的块之间相关度较高，这就造成了输入的冗余度非常高。因此，如何优化算法性能，从而解决输入的冗余性会成为未来一个非常值得研究的方向。

（2）通用问题。以往的视觉 Transformer 模型一般用于单任务，近年来一些模型可以做多任务，如 IPT 模型可以完成超分辨率、降噪、去雨等多任务。未来是否可以有一个通用的模型来处理所有任务。

（3）效率问题。Transformer 的计算量通常很大，在 ImageNet 数据集上，ViT 需要 180 亿 FLOPs 才能达到 78%的准确率，而普通的 CNN 模型如 GhostNet 只需要 6 亿 FLOPs，准确率即可达到 79%以上，所以需要开发高效的 Transformer 模型，提高运算效率。

（4）数据规模问题。Transformer 需要依赖大量的数据集来进行训练，而部分视觉任务的数据不能完全满足 Transformer 的训练需求。如何构建丰富、有效且全面的数据集以及如何减少 Transformer 对大量数据的依赖是未来研究的一个热点。

（5）可解释性问题。Transformer 结构不具备卷积的归纳偏置，却在视觉领域中表现优异。这对神经网络的可解释性问题提供一个研究方向。

每一种新结构的发展都是不断地发现问题，提出问题，再到解决问题，逐步不停迭代的过程。因此，Transformer 作为视觉领域新引入的模型，其本身还存在许多不足，需要不断地改进。未来的 Transformer 将会应用于更多的领域，以探索其本身巨大的潜力，实现更优、更合理的结果。

4 结语

Transformer 已成为计算机视觉领域的研究热点，由于其巨大的潜力，该模型一直受到研究者们的关注。本文对近几年来 Transformer 模型在图像分类、目标检测、图像分割等七个视觉任务中的应用进行分类和分析，并对其在计算机视觉中面临的挑战以及未来的发展趋势进行了总结和探讨。

参考文献

[1] Lecun Y, Bottou L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[2] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(06):1229-1251.

Zhou F Y, Jin L P, Dong J. Summary of Research on Convolutional Neural Networks[J]. Chinese Journal of Computers, 2017, 40(06):1229-1251.

[3] Goodfellow I J , Pouget-Abadie J , Mirza M , et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[5] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[C]//International Conference on Machine Learning. PMLR, 2018: 4055-4064.

[6] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Springer, Cham, 2020: 213-229.

[7] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C]//International Conference on Machine Learning. PMLR, 2020: 1691-1703.

[8] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[9] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12873-12883.

[10] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[11] Tsai Y H, Bai S, Liang P. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, 2019: 6558.

[12] 杨丽,吴雨茜,王俊丽.循环神经网络研究综述[J].计算机应用, 2018,38(S2):1-6.

Yang L, Wu Y Q, Wang J L. Research review of cyclic neural networks[J]. Journal of Computer Applications, 2018, 38(S2): 1-6.

[13] 任欢,王旭光.注意力机制综述[J].计算机应用, 2021, 41(S1): 1-6.

Ren H, Wang X G. Summary of Attention Mechanism[J]. Computer Applications, 2021, 41(S1):1-6.

[14] 刘金花.基于主动半监督极限学习机多类图像分类方法研究[D].东南大学,2016.

Liu J H. Research on multi-class image classification method based on active semi-supervised extreme learning machine [D]. Southeast University, 2016.

[15] 王红,史金钏,张志伟.基于注意力机制的 LSTM 的语义关系抽取[J].计算机应用研究,2018,35(05):1417-1420+1440.

Wang H, Shi J C, Zhang Z W. Semantic relation extraction of LSTM based on attention mechanism[J]. Application Research of Computers, 2018, 35(05): 1417-1420+1440.

[16] 唐海桃,薛嘉宾,韩纪庆.一种多尺度前向注意力模型的语音识别方法[J].电子学报,2020,48(07):1255-1260.

Tang H T, Xue J B, Han J Q. A multi-scale forward attention model speech recognition method[J]. Chinese Journal of Electronics, 2020, 48(07): 1255-1260.

[17] Wang W, Shen J, Yu Y, et al. Stereoscopic thumbnail creation via efficient stereo saliency detection[J]. IEEE transactions on visualization and computer graphics, 2016, 23(8): 2014-2027.

[18] Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding[C]//Proceedings of the International Conference on Learning Representations, Toulon, France. 2017.

[19] Han K, Wang Y, Chen H, et al. A Survey on Visual Transformer[J]. arXiv preprint arXiv:2012.12556, 2020.

[20] Khan S, Naseer M, Hayat M, et al. Transformers in Vision: A Survey[J]. arXiv preprint arXiv:2101.01169, 2021.

[21] Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. arXiv preprint arXiv:2103.00112, 2021.

[22] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[J]. arXiv preprint arXiv:2101.11986, 2021.

[23] Jiang Z, Hou Q, Yuan L, et al. Token labeling: Training a 85.5% top-1 accuracy vision transformer with 56m parameters on imagenet[J]. arXiv preprint arXiv:2104.10858, 2021.

[24] Zhou D, Kang B, Jin X, et al. Deepvit: Towards deeper vision transformer[J]. arXiv preprint arXiv:2103.11886, 2021.

[25] Zhu X, Su W, Lu L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[26] Sun Z, Cao S, Yang Y. Rethinking Transformer-based Set Prediction for Object Detection[J]. arXiv preprint arXiv:2011.10881, 2020.

[27] Dai Z, Cai B, Lin Y, et al. UP-DETR: Unsupervised Pre-training for Object Detection with Transformers[J]. arXiv preprint arXiv:2011.09094, 2020.

[28] Zheng M, Gao P, Wang X, et al. End-to-End Object Detection with Adaptive Clustering Transformer[J]. arXiv preprint arXiv:2011.09315, 2020.

[29] Zheng S, Lu J, Zhao H, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers[J]. arXiv preprint arXiv:2012.15840, 2020.

[30] Strudel R, Garcia R, Laptev I, et al. Segmenter: Transf-ormer for Semantic Segmentation[J]. arXivpreprint arXiv:2105.05633, 2021.

[31] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers[J]. arXiv preprint arXiv:2105.15203, 2021.

[32] Wang H, Zhu Y, Adam H, et al. MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers[J]. arXiv preprint arXiv:2012.00759, 2020.

[33] Wang Y, Xu Z, Wang X, et al. End-to-End Video Insta-nce Segmentation with Transformers[J]. arXiv preprint arXiv:2011.14503, 2020.

[34] Ma F, Sun B, Li S. Robust Facial Expression Recognition with Convolutional Visual Transformers[J]. arXiv preprint arXiv:2103.16854, 2021.

[35] Zheng C, Zhu S, Mendieta M, et al. 3d human pose estimation with spatial and temporal transformers[J]. arXiv preprint arXiv:2103.10455, 2021.

[36] He S, Luo H, Wang P, et al. TransReID: Transformer-based Object Re-Identification [J]. arXiv preprint arXiv:2102.04378, 2021.

[37] Liu R, Yuan Z, Liu T, et al. End-to-end lane shape prediction

with transformers[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3694-3702.

[38] Chen H, Wang Y, Guo T, et al. Pre-trained image processing transformer[J]. arXiv preprint arXiv:2012.00364, 2020.

[39] Yang F, Yang H, Fu J, et al. Learning texture transformer network for image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5791-5800.

[40] Jiang Y, Chang S, Wang Z. Transgan: Two transformers can make one strong gan[J]. arXiv preprint arXiv:2102.07074, 2021.

[41] Chen Y, Cao Y, Hu H, et al. Memory enhanced global-local aggregation for video object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10337-10346.

[42] Zeng Y, Fu J, Chao H. Learning joint spatial-temporal transformations for video inpainting[C]//European Conference on Computer Vision. Springer, Cham, 2020: 528-543.

[43] Bertasius G, Wang H, Torresani L. Is Space-Time Attention All You Need for Video Understanding?[J]. arXiv preprint arXiv:2102.05095, 2021.

[44] Liu Z, Luo S, Li W, et al. ConvTransformer: A Convolutional Transformer Network for Video Frame Synthesis[J]. arXiv preprint arXiv:2011.10185, 2020.

[45] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[46] Zagoruyko S, Komodakis N. Wide Residual Networks[J]. British Machine Vision Conference 2016, 2016.

[47] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[48] Kolesnikov A, Beyer L, Zhai X, et al. Big transfer (bit): General visual representation learning[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020: 491-507.

[49] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[50] Howard A G, Zhu M, Chen B. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[51] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[52] Yun S, Oh S J, Heo B. Re-labeling imagenet: from single to multi-labels, from global to localized labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2340-2350.

[53] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[54] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[55] 李彦冬. 基于卷积神经网络的计算机视觉关键技术研究[D].电子科技大学,2017.

Li Y D. Research on key technologies of computer vision based on convolutional neural networks [D]. University of Electronic Science and Technology of China, 2017

[56] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conf-erence on computer vision. 2017: 764-773

[57] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[58] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional onestage object detection[C]//Proceedings of the IEEE/CVF Int-ernational Conference on Computer Vision. 2019: 9627-9636

[59] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7103-7112.

[60] Ding X, Guo Y, Ding G, et al. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1911-1920.

[61] Yang L, Fan Y, Xu N. Video instance segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5188-5197.

[62] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897-6906.

[63] Lin K, Wang L, Liu Z. End-to-end human pose and mesh reconstruction with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1954-1963.

[64] Hao L .Bags of Tricks and A Strong Baseline for Deep Person Re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2019.

[65] Chen T, Ding S, Xie J, et al. Abd-net: Attentive but diverse person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8351-8361.

[66] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 542-551.

[67] Khorramshahi P, Peri N, Chen J, et al. The devil is in the details: Self-supervised attention for vehicle re-identification[C]//European Conference on Computer Vision. Springer, Cham, 2020: 369-386.
[68] Tabelini L, Berriel R, Paixao T M, et al. Polylanenet: Lane estimation via deep polynomial regression[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 6150-6156.
[69] Li X, Li J, Hu X, et al. Line-cnn: End-to-end traffic line detection with line proposal unit[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(1): 248-258.
[70] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3:2672-2680.
[71] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8110-8119.
[72] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211
[73] Liu Z, Yeh R A, Tang X, et al. Video frame synthesis using deep voxel flow[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4463-4471.
[74] Villegas R, Yang J, Hong S, et al. Decomposing motion and content for natural video sequence prediction[J]. arXiv preprint arXiv:1706.08033, 2017.