软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(7):3408−3421 [doi: 10.13328/j.cnki.jos.006530]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于卷积神经网络的全景分割 Transformer 模型*

毛 琳, 任凤至, 杨大伟, 张汝波

(大连民族大学 机电工程学院, 辽宁 大连 116600)

通信作者: 任凤至, E-mail: renfz2019@163.com

摘 要: 提出一种基于卷积神经网络的 Transformer 模型来解决全景分割任务, 方法借鉴 CNN 在图像特征学习方面的先天优势, 避免了 Transformer 被移植到视觉任务中所导致的计算量增加. 基于卷积神经网络的 Transformer 模型由执行特征域变换的映射器和负责特征提取的提取器这两种基本结构构成, 映射器和提取器的有效结合构成了该模型的网络框架. 映射器由一种 Lattice 卷积模型实现, 通过对卷积滤波器进行设计和优化来模拟图像的空间关系. 提取器由链式网络实现, 通过链式单元堆叠提高特征提取能力. 基于全景分割的结构和功能, 构建了基于 CNN 的全景分割 Transformer 网络. 在 MS COCO 和 Cityscapes 数据集的实验结果表明, 所提方法具有优异的性能.

关键词: 全景分割; 卷积神经网络; Transformer; 语义分割; 实例分割

中图法分类号: TP391

中文引用格式: 毛琳, 任凤至, 杨大伟, 张汝波. 基于卷积神经网络的全景分割Transformer模型. 软件学报, 2023, 34(7): 3408–3421. http://www.jos.org.cn/1000-9825/6530.htm

英文引用格式: Mao L, Ren FZ, Yang DW, Zhang RB. CNN Based Transformer for Panoptic Segmentation. Ruan Jian Xue Bao/Journal of Software, 2023, 34(7): 3408–3421 (in Chinese). http://www.jos.org.cn/1000-9825/6530.htm

CNN Based Transformer for Panoptic Segmentation

MAO Lin, REN Feng-Zhi, YANG Da-Wei, ZHANG Ru-Bo

(School of Electromechanical Engineering, Dalian Minzu University, Dalian 116600, China)

Abstract: This study proposes a convolutional neural network (CNN) based Transformer to solve the panoptic segmentation task. The method draws on the inherent advantages of the CNN in image feature learning and avoids increase in the amount of calculation when the Transformer is transplanted into the vision task. The CNN-based Transformer is attributed to the two basic structures of the projector performing the feature domain transformation and the extractor responsible for the feature extraction. The effective combination of the projector and the extractor forms the framework of the CNN-based Transformer. Specifically, the projector is implemented by a lattice convolution that models the spatial relationship of the image by designing and optimizing the convolution filter configuration. The extractor is performed by a chain network that improves feature extraction capabilities by chain block stacking. Considering the framework and the substantial function of panoptic segmentation, the CNN-based Transformer is successfully applied to solve the panoptic segmentation task. The experimental results on the MS COCO and Cityscapes datasets demonstrate that the proposed method has excellent performance.

Key words: panoptic segmentation; convolutional neural network (CNN); Transformer; semantic segmentation; instance segmentation

1 介 绍

全景分割[1]是语义分割和实例分割的联合算法, 目前全景分割结构以卷积神经网络 (CNN) 为主流, 其研究围绕着语义和实例分割两种网络架构的组合而展开. 实例分割是面向前景实例的分割任务, 目前存在单阶段和双阶

* 基金项目: 国家自然科学基金 (61673084); 辽宁省自然科学基金 (20170540192, 20180550866)
收稿时间: 2021-07-23; 修改时间: 2021-09-04, 2021-10-28; 采用时间: 2021-11-05; jos 在线出版时间: 2022-09-23
CNKI 网络首发时间: 2022-11-15

段两种主流方法. 单阶段方法采用了语义分割的方法, 从分组中提取实例. 双阶段方法通过边界框的生成来进行目标检测, 继而形成实例预测结果; 凭借对候选区域的准确识别, 双阶段网络的表现甚为出色, 因此, 全景分割中的实例分割几乎被双阶段网络所覆盖, 而最常用的实例分割框架是 Mask R-CNN[2]. 不同于实例分割, 语义分割的分割对象是背景的填充目标, 其结构更加复杂多样. FCN (fully convolutional network) 的提出为语义分割开创了新的局面, 基于 FCN[3]的编解码网络得以迅速兴起, 在至关重要的编码网络上, 研究者尝试了大量的特征融合方法[4−8]和卷积变换研究[9−11]来提取上下文信息; 解码结构则由常用的上采样手段来完成. 编解码结构是语义分割的经典框架, 为其发展奠定了深厚的基础, 然而, 此类结构依然受到感受野接受范围的制约; 基于注意力机制的语义分割网络通过注意力模块来捕捉长期依赖关系, 以此来提取全局特征.

伴随着诸多体系架构的成熟, 基于卷积神经网络的全景分割网络逐渐成形. 一些算法[12−15]以 Mask R-CNN 为基础联合语义分割框架, 取得了不错的效果; 另一些工作[16,17]则是沿用语义分割的编解码架构, 也是在继承了语义或实例分割的网络结构的基础上, 以对编码器的精巧设计来提升性能. 全景分割凭借卷积神经网络得以崛起并迅速发展. 但是, 随着学界对卷积神经网络的广泛研究, 尤其是分割网络架构的日益成熟, 由 CNN 构建的全景分割算法已经形成了相对固定的结构模式, 这束缚了全景分割的前进步伐, 使其受困于网络结构的框架设计之中.

为推进图像分割的发展, 学者尝试将 Transformer 的架构套用于分割网络以打造新型网络架构. Transformer 本质是通过特征域变换来寻求新的视角解决图像任务. 图像 Transformer 依靠映射 (projection)、注意力机制和 MLP 来实现变换的目的. Projection 将图像抽象为具有位置信息嵌入的序列; 注意力机制通过全局感受野获取上下文语义信息及全局依赖关系; MLP 辅助注意力结构进行信息提取. SETR[18] 将 Transformer 作为编码器, 通过线性映射和注意力结构把输入转换为全局语义特征, 以适用语义分割的要求. DETR[19]将 CNN 和 Transformer 进行联合来实现全景分割, Transformer 在图像分割上表现出了优秀性能. 但是, 当 Transformer 被应用于视觉任务时, 由 projection 和注意力网络在特征转换和处理中带来的计算复杂性是不可忽视的[20]. Transformer 网络的复杂程度使其在图像任务中的应用仍然面临很多困难; 更重要的, Transformer 的优势集中在全局语义理解上, 在图像空间和局部的感知能力明显弱于 CNN[21], 这更降低了 Transformer 在解决图像任务时的竞争力.

基于这一问题, 我们致力于构建基于卷积神经网络的 Transformer 模型 (CNN based Transformer, CBT) 来解决全景分割问题. 通过 CNN 来构造 Transformer, 能够在一种新的特征域内解释图像分割任务, 利用 CNN 对图像特征的敏感性设计映射器, 进行特征域变换; 提出提取器配合映射器使用, 对域变换前后的特征进行提取操作, 保障分割质量. 3 种典型的全景分割网络架构如图 1 所示.

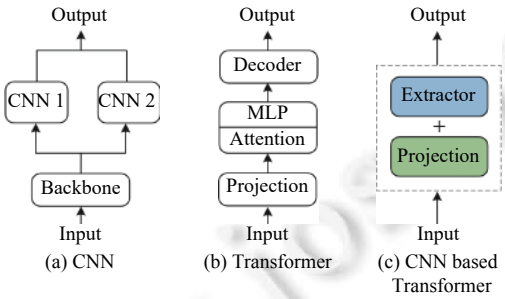


图 1 网络架构 (CNN 1 代表语义分割, CNN 2 代表实例分割)

2 相关工作

基于 CNN 的全景分割网络的工作重点为语义分割和实例分割的有效联合, 因此, 语义和实例分割网络的研究及其联合是该类工作的重要部分.

实例分割: 实例分割的目标是区分同一对象的不同实例, 因此, 其任务关键在于提取目标的显著信息从而区分不同实例. 当前主要存在两类实例分割网络: 以 Mask R-CNN 为代表的双阶段网络和单阶段网络. 双阶段网

络[22−24]先将输入图像映射为目标候选区域以检测目标的存在, 而后区分不同实例进行密集预测; 单阶段网络[25−27]则依靠语义分割的方法来执行像素级分割, 然后再从分组中提取不同的实例. 这一方法采用和语义分割相同的特征映射方式, 依赖于像素级分割的质量. 基于区域候选方法的准确性和独立性, 双阶段实例分割方法在全景分割网络中更为常见.

语义分割: 语义分割的任务是为图像中的所有像素分配类别标签, 从而实现像素级分割, 因此, 全局特征和上下文语义信息的获取是语义分割的重点. 目前存在两类语义分割网络结构: 第 1 类网络是编解码网络[3,4,7,28–30], 编解码网络的本质将输入映射为上下文信息丰富的特征, 为达到这一目标, 有两类研究方法: (1) 通过多层次特征融合来丰富特征、聚合语义信息, 根植于这一原理, 形成了 FCN[3]、U-Net[7]等众多语义分割的经典网络. (2) 以扩大视觉感受野为目标, 对卷积方法做出的变换和改进工作, 具有代表性的有 ASPP[9,31]、大卷积核[11]和可变形卷积[32]等方法. 其中, 基于空间采样能力的提升, 由可变形卷积所构成的编解码结构具有不俗的表现, 因此被全景分割算法广泛使用, 曾经一度占据了算法精度的排行榜. 第 2 类网络是基于注意力机制的语义分割网络, 该类网络通过在空间和通道维度中捕获远程上下文信息来攻克像素级分割任务, 借该方法将输入图像变换为一种全局逻辑关系. PAN[33]利用全局注意力模块收获全局语义特征来指导空间金字塔进行特征提取; SANet[34]和 DANet[35]在空间和通道上建立语义依赖模型, 整合全局和局部逻辑关系; CCNet[36] 创新性地设计出交叉注意力模块, 挖掘交叉路径上像素的上下文信息; PSANet[37]提出自适应预测注意力机制来加强特征图中各个位置点的联系, 以降低常规卷积滤波器对特征信息流动的束缚, 并设计双向信息传播路径, 以理解复杂场景.

全景分割: Panoptic FPN[38]主干网络的提出为分割网络的结合提供了有利的条件, 目前绝大多数全景分割网络的基本结构是 FPN 作为网络主干与 Mask R-CNN 实例分割网络及标准语义分割网络的组合, 在这种框架之下, UPSNet[15]将可变形卷积应用于语义分割网络来提升语义分割效果; EfficientPS[39]使用多路不同参数的可分离卷积并行处理网络来捕获语义分割中不同尺度的特征; TASCNet[12]和 AUNet[13]通过建立两种分割网络的联系来降低分割结果的差异性. OANet[14]提出一种空间排序模块来融合两种分割结构, 改善目标遮挡问题. 除此之外, 全景分割还存在一种编解码网络架构, 以 DeeperLab[16]和 Panoptic-DeepLab[17]为代表, 主要表现为双路并行结构, 多用多路并行空洞卷积网络合并特征以拓宽感受野, 提高语义分割精度.

Transformers: Transformers 常用于机器翻译和自然语言处理. 近年来, 在图像识别任务中, Transformer 被视为卷积神经网络的可行替代方案. 一些工作单纯使用 Transformer 的结构来完成图像任务. ViT[40]是第一个完全采用 Transformer 来完成图像分类任务的工作. LRNet[41]和 SAN[42]在减轻由于全局自注意力机制带来的繁重计算上面做出了探索. Axial-DeepLab[43]将全局注意力分解为两个单独的轴向注意力, 从而大大减少了计算量. 另一些工作则将 Transformer 与 CNN 结合. STTR[44]和 LSTR[45]分别运用 Transformer 进行视差估计和车道形状预测. DETR[19]和 VisTR[46]都是以 CNN 为网络主干将输入图像处理为一系列图像特征, 送入 Transformer 网络, 然后搭配不同的功能网络分别实现目标检测和实例分割的任务. SETR[18]使用线性映射和多层注意力结构作为编码器学习图像特征, 利用 CNN 作为解码器, 实现语义分割功能.

除了计算复杂度的增加, Transformer 在精度上的性能表现还存在一些缺陷, 这主要是由 Transformer 网络结构所决定的. 在图像任务中, Transformer 需要先将输入图像映射为众多小的特征图, 而后利用注意力和 MLP 等结构, 为这些特征图嵌入位置信息并使用注意力模型提取特征. 然而, 由于 projection 和 MLP 会对图像本身的空间关系造成破坏; 再加上, 注意力网络和 MLP 结构对图像空间关系和局部细节特征的提取能力与 CNN 相比相对较弱, 这些由结构本身带来的不利条件使得 Transformer 在图像领域的应用仍然面临阻碍.

为增强 Transformer 对图像空间关系的感知能力, CvT[21]在 Transformer 中引入卷积来模拟局部空间环境并完成特征映射变换, 使得 Transformer 能够提取到对解决视觉任务十分重要的特征属性, 弥补 Transformer 在图像识别任务上的局限. 在 CvT 的启发下, 本文提出基于卷积神经网络的 Transformer 模型, 完全使用 CNN 构建 Transformer 网络, 充分利用 CNN 在图像识别上的有利条件, 采用 Transformer 变换思想学习图像任务. 为实现 Transformer 功能, 我们构造了映射器来进行特征映射 (替代 projection)、提取器进行特征提取 (替代注意力网络), 图 1 展示了 Transformer 和基于卷积神经网络的 Transformer 的不同.

3 基于卷积神经网络的 Transformer 模型

本节分为 3 部分来论述基于卷积神经网络的 Transformer 模型的构建过程. 第 3.1 节介绍模型的构成单元, 即映射器和提取器的设计; 第 3.2 节给出模型的框架结构, 即如何组建映射器和提取器; 第 3.3 节使用该模型设计全景分割网络.

3.1 构成单元

基于卷积神经网络的 Transformer 模型的构成单元是映射器和提取器, 映射器负责特征域变换的操作, 提取器则承担着变换前后特征的提取工作, 如图 2 所示.

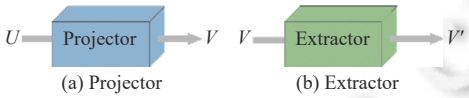


图 2 映射器和提取器

3.1.1 映射器

定义 1. 现在图像处理任务都是在单一域内进行的, 其研究方法和手段具有局限性. 于是本文提出域变换方法, 将特征放置到不同域内进行处理, 从而采取新的视角和研究方法来解决该任务. 映射器就是来实现域变换的, 它可被定义为:

$$V = P(U) \tag{1}$$

其中, P 是由卷积网络构成的映射函数; U 是域变换过程的输入; V 是域变换过程的输出.

映射器的本质是一种特征变换, 将输入信息从当前域 (U 域) 转换到特定的域 (V 域) 内, 这一变换过程是由卷积网络来实现的, 更进一步可归结于卷积滤波器对输入图像空间性质的改变. 然而, 现存卷积滤波器的构型十分有限, 使得卷积网络的映射能力受到很大局限, 为开发卷积滤波器的映射能力, 为卷积变换网络提供更多的映射工具, 我们从卷积滤波器的空间构型出发设计了一系列卷积映射器.

卷积映射器的设计有两个方面的考虑, 一是以空间利用率为线索来挖掘现有卷积滤波器物理设计上的潜能, 二是通过滤波器内部空间关系的建模优化特征映射方法. 根据空间利用率由大到小的顺序, 本文给出 4 种具有不同空间关系的卷积滤波器 P_{L1}, P_{L2}, P_{L3}, P_{L4}. 因其形似“Lattice”, 故将其称为 Lattice 卷积模型, 图 3 展示了稀疏率分别为 S_{25}^{0}、S_{25}^{8}、S_{25}^{12} 和 S_{25}^{16} 的 4 种 Lattice 卷积模型.

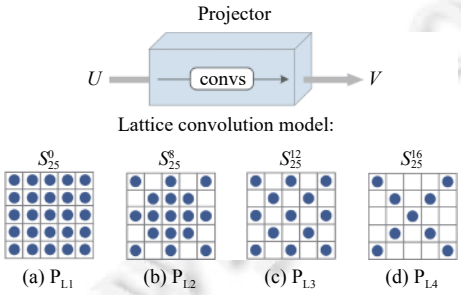


图 3 Lattice 卷积模型

Lattice 卷积模型通过标准卷积和空洞卷积的组合来构建滤波器内部的空间关系模型. 定义 Lattice 卷积滤波器的稀疏率为 $S = S_{k\times k}^{n}$, 其中, k 为滤波器大小, n 为滤波器中数值为 0 的区域. 假设用 $F_{k\times k}^{r}$ 表示卷积核大小为 $k\times k$, 扩张率为 r 的标准卷积, 则稀疏率分别为 P_{L2}、P_{L3}、P_{L4} 和 P_{L1} 的 Lattice 卷积模型的构造方法如下.

• P_{L2} 型的 Lattice 卷积的构造方法为:

$$F_1^1+\sum_{i=1}^{r}F_{3\times3}^r \tag{2}$$

• P_{L3} 型的 Lattice 卷积的构造方法为:

$$F_{1\times1}^1+\sum_{i=2}^{k}F_{k\times k}^2 \tag{3}$$

• P_{L4} 型的 Lattice 卷积的构造方法为:

$$F_{1\times1}^1+\sum_{i=2}^{2r}F_{2\times2}^r \tag{4}$$

• 特别地, P_{L1} 型的 Lattice 卷积表现为标准卷积滤波器, 是全映射.

图 4 展示了 k=5 时 Lattice 卷积模型的构造法.

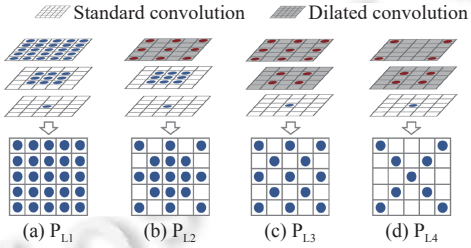


图 4 Lattice 卷积模型的构造方法 (k=5)

Lattice 卷积模型通过对滤波器稀疏性的控制构造了不同的空间关系模型, 从而形成多种特征映射方法. P_{L1} 型的 Lattice 卷积表现为特征的全映射, 通过卷积空间的全部利用来全面且诚实地反映输入的信息; 剩下的 3 种卷积空间关系模型是为提取图像边缘、细节特征而设计的, 通过卷积模板的路径来拟合图像中各线条间的逻辑关系. P_{L2} 型和 P_{L3} 型的 Lattice 卷积通过抑制部分无意义像素的表达来强化边缘特征等线条的表现; 稀疏率最大的 P_{L4} 型 Lattice 卷积的分布是对角线位置来抓取图像全局的主要信息.

3.1.2 提取器

定义 2. 在域变换后, 需要在新域内执行功能操作来完成具体任务, 于是提取器应运而生, 其作用是特征提取——从输入信息中有效地获取所需结果. 提取器可被定义为:

$$V'=E(V) \tag{5}$$

其中, E 是由卷积网络构成的提取函数, 和 V 维度和大小相同; V 是提取过程的输入, V'是提取过程的输出.

提取器是由卷积网络构建的, 其计算过程是在同一个特征域内完成的. 提取器的本质是不断地卷积计算, 通过参数的学习, 使得卷积具备过滤知识的能力, 继而提取有用特征, 提升网络性能. 特征提取能力的高低取决于网络架构的优化设计, 于是我们将映射器用作卷积滤波器, 在网络架构的层面上开展提取器的设计工作.

网络架构的改进主要有网络深度和宽度两个方面. 在网络宽度上, 通常采用多路并行处理结构来拓宽网络提高特征的丰富性. 在网络深度方面则通过网络层数的增加来实现特征质量的提升. 得益于 ResNet 的提出, 深度卷积神经网络得以迅速发展, 而基本单元复制堆叠的思想也得到了广泛传播和使用. 受此启发, 本文提出链式网络作为提取器, 该网络通过一种精巧的单元结构——链式单元来完成特征的层次化提取, 利用这一单元结构的复制堆叠深化网络的特征提取能力. 链式单元由映射器及其参考支路构成, 映射器执行特征映射过程, 参考支路通过 shortcut 来保持原始特征信息, 使得映射变换过程在原始信息的参考下实现特定目的的提取功能.

根据构造方法的不同, 链式网络有 E_{C1} 和 E_{C2} 两种类型. 如图 5 所示.

E_{C1} 型链式网络中链式单元采取顺序堆叠方式, 单元中的映射器位于链式网络同侧. 这种结构设计能够层次化地提取特征, 丰富特征体系.

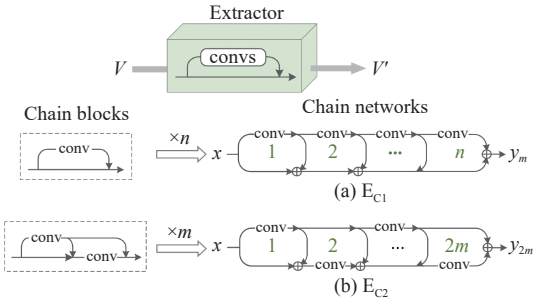


图 5 链式网络及其构造方法

• 假定 n ($n>0$, $n\in z$) 为链式网络包括链式单元的个数, 那么 E_{C1} 型链式网络可定义为:

$$y_n = x + C(x) + C(C(x)) + \ldots + \underbrace{C\ldots(C(C(x)))}_{n} \tag{6}$$

其中, x 为 E_{C1} 型链式网络输入, $C(x)$ 是映射网络函数, y_n 是第 n 个链式单元的输出, 即 E_{C1} 型链式网络输出.

从公式 (6) 可见, E_{C1} 网络能够收获多层次特征, 随着网络深度的延长, 特征信息的丰富性和层次性会逐渐提高. 能够拥有这样的特征提取效果与网络结构的设计是分不开的, 一方面, E_{C1} 型网络中映射器同侧排布的结构特点能够对特征进行逐层提取, 不断前进, 获取深层次特征; 另一方面, 捷径结构的使用使得提取到不同的层次特征得以保留, 维持了特征层次系统的完整性. 这样的网络结构为 E_{C1} 网络的特征提取能力提供了坚实的基础, 保证了网络在图像识别任务上的性能表现.

与 E_{C1} 型不同, E_{C2} 型链式网络以链式单元及其翻转结构的组合作为基本单元进行复制堆叠, 单元中的映射器位于网络异侧. E_{C2} 型网络通过函数位置的变换来激发自身的学习潜能, 避免提取方式的固化, 深层挖掘特征信息.

• 假定 $2m$ ($m>0$, $m\in z$) 为链式网络包括链式单元的个数, 则 E_{C2} 型链式网络可定义为:

$$\begin{cases} y_{2m-1} = y_{2m-2} + \ldots + \underbrace{C\ldots(C(C(x)))}_{m} \\ y_{2m} = C(y_{2m-1}) + \ldots + \underbrace{C\ldots(C(C(x)))}_{m} \end{cases} \tag{7}$$

其中, x 为 E_{C2} 型链式网络输入, $C(x)$ 是映射网络函数, y_{2m-1} 和 y_{2m} 分别是第 $(2m-1)$ 个和第 $2m$ 个链式单元的输出. 特别地, 网络中没有链式单元时, $y_0=x$.

在 E_{C1} 型网络基础上提出 E_{C2} 型网络的目的是为了给特征层次化提取的过程赋予不确定的因素, 以避免固定提取方法可能会引起的效率降低问题. E_{C2} 型链式网络将链式单元和它的翻转结构作为基本构成单元, 构造映射器异侧分布的结构形态, 映射函数的位置变换打乱了传统的提取模式, 给网络创造了学习空间, 通过结构上的调整开发网络的学习潜力, 实现提取能力的提升.

3.2 框架结构

映射器 (可缩写为 P) 是基于 CNN 的 Transformer 模型的核心, 它的功能是处理空间级别的特征, 空间变换器网络[47]是典型的一种. 而提取器 (可缩写为 E) 则凭借自身的特征提取能力, 决定了 Transformer 模型的性能, 其功能类似于残差网络.

在基于 CNN 的 Transformer 网络中, 位于映射器前的提取器能够对要变换的特征进行预处理, 形成一个“E-P”型结构, 这种结构常用于以 CNN 为骨干的 Transformer 模型, 如 DETR[19]和 VisTR[46]. 其中, CNN 充当提取器来执行预处理操作. 3DSTN[48]也清楚地反映了“E-P”结构的使用; 位于映射器后面的提取器可以对转换后的特征进行提纯, 形成“P-E”型的结构. 这种结构常见于采用 CNN 作为解码器的 Transformer 模型, 比如 SETR[18], 其中 CNN 起到了特征提纯的作用, 使用空间变换器网络的算法[49]也体现了“P-E”结构; 面对复杂的任务, 可以将提取器置于映射器结构的两侧, 形成“E-P-E”型结构, 这种类型的网络适用于需要高精度的精确预测任务. 表 1 显示了映射器和提取器的组合形式.

表 1 映射器和提取器的组合形式

结构	描述
E-P	预处理, 域变换
P-E	域变换, 特征提纯
E-P-E	预处理, 域变换, 特征提纯

3.3 全景分割网络

要用基于 CNN 的 Transformer 网络来完成全景分割, 要先对全景分割任务进行功能分析. 全景分割需实现对实例和填充物两类目标的分割, 因此, 其网络通常由主干、实例分割和语义分割这 3 个模块构成. 自然地, Transformer 网络也必须具备实现这 3 种功能的结构来完成全景分割任务, 根据这一需求, 将提取器作为全景分割的主干, 执行特征预处理操作; 针对两种不同的分割任务, 利用映射器实现特征域的变换, 并再次使用提取器以确保分割质量, 通过映射器和提取器的组合来完成实例分割和语义分割的任务. 基于 CNN 的全景分割 Transformer 网络的构成如表 2 所示.

表 2 基于 CNN 的全景分割 Transformer 的网络框架

结构	主干	实例分割	语义分割
映射器	—	P_{L1}	P_{L2}
提取器	E_{C1}	E_{C2}	E_{C2}
网络框架		E-P-E	

基于 CNN 的全景分割 Transformer 网络由网络主干, 映射器和提取器这 3 部分构成. 我们以 E_{C1} 型网络设计网络主干, 对输入信息进行统一层次化的特征提取; 根据面向对象的不同, 实例分割和语义分割采用的映射器是不同的, 我们选定不同的 Lattice 卷积模型以实现通用特征到前景目标域和背景填充域的变换; 而后使用 E_{C2} 型链式网络对映射后的特征进行精炼萃取和精密预测; 最后融合两种分割结果, 形成全景分割预测结果. 基于 CNN 的全景分割 Transformer 网络结构如图 6 所示. 下面将对网络主干、映射器和提取器这 3 部分展开详细论述.

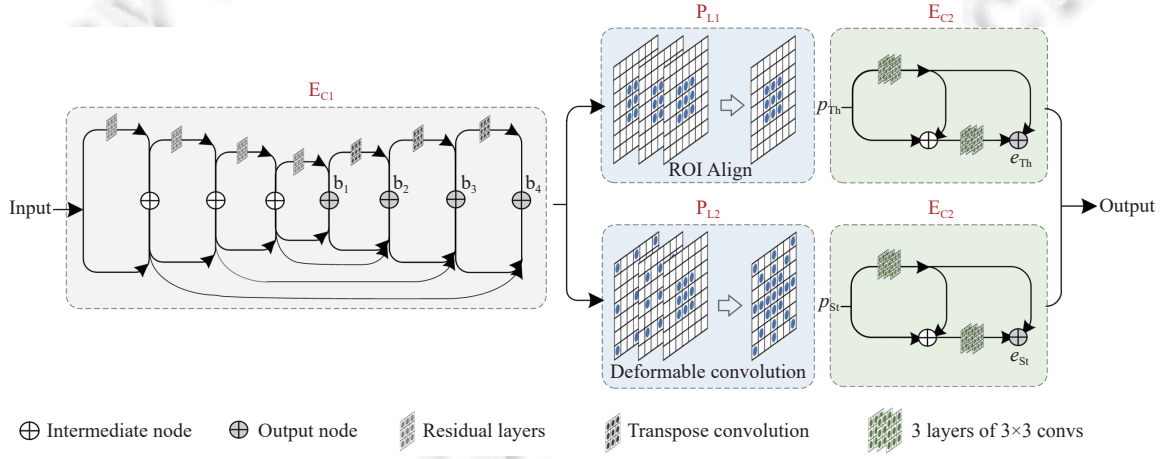


图 6 基于 CNN 的全景分割 Transformer 网络

网络主干: 基于功能需求分析, 主干网络需为后面分割任务的执行提供通用且丰富的特征, 因此, 我们使用了 E_{C1} 型网络, 利用链式学习原理将卷积映射器置于主干网络同侧, 逐层提取输入图像的特征信息, 丰富特征层次. 并且, 将标准卷积作为映射器, 保证信息的全面性. 网络主干分为下采样和上采样两部分, 下采样阶段以残差模块充当映射器, 对输入图像进行特征编码; 上采样阶段则利用反卷积恢复空间关系完成解码过程. 我们沿用了 FPN

的跨层连接, 在编码和解码网络间建立了信息联系, 以提升主干网络的性能. 最终由主干网络的上采样部分输出 4 层特征$\{b_1, b_2, b_3, b_4\}$以供后面分割网络使用.

映射器: 针对两种识别对象, 本文选择了 P_{L1} 和 P_{L2} 两个映射器分别进行映射变换.

实例目标通常是体积较小且具有固定形状的目标, 实例分割的重点是区分不同的实例个体, 于是挖掘实例的细节特征以确定各个实例目标的独特性便成了重要事宜. 细节特征往往是对实例目标的深层提取获得的, 因此, 细节特征可看作是实例目标深层特征的映射. 所以, 我们堆叠了 3 层稀疏率等于 1 的 Lattice 卷积作为实例分割的映射器 P_{L1}^{Th}, 完成特征映射过程, 通过提取深度来保证提取特征的质量. 具体过程为, 首先利用 ROI Align[2]来处理特征图$\{b_1, b_2, b_3, b_4\}$, 然后便进入映射器 P_{L1}^{Th}, 经过 Lattice 卷积的映射变换后得到特征图 p_{Th}.

填充目标往往体积较大且没有固定形状, 对这类目标进行分割的关键在于对填充区域像素的准确分类, 为区域像素确定类别归属的首要问题是找到区域边界, 因此, 填充目标边缘特征的获取对语义分割这一任务至关重要, 故我们使用稀疏率 s<1 的 Lattice 模型对填充目标特征信息进行域变换, 将输入映射为包含丰富轮廓、线条信息的边缘特征, 同时采用可变形卷积作为 Lattice 模型的基底, 聚集上下文语义特征, 形成语义分割的映射器 P_{L2}^{St}. 接收到主干网络输出的 4 层特征图$\{b_1, b_2, b_3, b_4\}$, 我们设计了 4 支不同参数的 P_{L2}^{St} 并行支路分别对每个特征图进行独立操作, 并利用上采样调整特征尺寸来合并 4 个输出结果, 得到合并特征图 p_{St}.

提取器: 特征映射变换后, 需使用特征提取网络对映射后的特征质量提供保证, 此处我们使用了 E_{C2} 型网络强化特征的表现力, 并与 P_{L1}^{St} 和 P_{L2}^{Th} 配合共同完成语义分割和实例分割两项任务. E_{C2} 具有两个链式单元, 分别对特征图 p_{St} 和 p_{Th} 进行特征提取生成提取结果 e_{St} 和 e_{Th}, 并融合两种结果得到全景分割输出结果.

4 实 验

4.1 数据集和指标

本文使用 MS COCO[50]和 Cityscapes[51]数据集进行仿真测试. MS COCO 是用于场景理解的公共数据集. 我们使用的 COCO2017 版本在训练集中有 118 000 张图像, 在验证集中有 5 000 张图像, 其中包括 80 个实例类别和 53 个填充物类别. Cityscapes 是城市交通场景的数据集, 由于交通驾驶环境的复杂性, 该数据集是全景分割最具挑战性的数据集之一. Cityscapes 数据集一共拥有 5 000 张图片, 包括 3 000 张训练图片、500 张验证图片和 1 500 张测试图片, 可识别 8 种实例类目标和 11 种填充物. 实验采取全景分割评价指标, 即: 全景质量 (PQ)、分割质量 (SQ) 和识别质量 (RQ) 来评估全景分割结果. PQ、SQ 和 RQ 表示如下:

$$PQ=\underbrace{\frac{\sum_{(p,g)}IoU(p,g)}{|TP|}}_{SQ}\times\underbrace{\frac{|TP|}{|TP|+\frac{1}{2}|FP|+\frac{1}{2}|FN|}}_{RQ} \tag{8}$$

其中, $IoU(p, g)$ 表示预测对象 p 与真值 g 的交并比 IoU, TP、FP 和 FN 分别表示正确匹配的分割图、错误匹配的分割预测图和错误匹配的分割真值图. 另外我们也比较了实例目标和填充物两个种类的具体表现, 分别用 PQ^{Th}、PQ^{St} 来表示.

基于 PyTorch 平台[52], 我们使用了 4 个 GPU 来训练模型, 采用了与文献 [15] 中类似的训练技巧, 并将 SGD 设置为具有 0.9 动量和 0.000 1 权重衰减的优化算法. 在 COCO 数据集上, 训练的迭代次数为 400 000 次, 学习率在 240 000 和 320 000 次迭代时降低了 10 倍. 在 Cityscapes 数据集上, 训练的迭代次数为 48 000 次, 在迭代次数为 36 000 次时降低了学习率. 对于其他实验细节, 我们采用了文献 [15] 中的经典超参数配置. P_{L1} 由 ROI Align 组成, E_{C2} 充当掩码分支, 其中链式单元中的映射器函数使用了 3 层 3×3 卷积层. 可变形卷积被更改为 Lattice 卷积架构, 从而得到 P_{L2}. 一个 1×1 的卷积结构与 Lattice 卷积并联进行加法融合, 为 Lattice 卷积提取的边缘特征提供参考信息.

4.2 实验结果对比

我们在 MS COCO 和 Cityscapes 数据集上将提出方法与其他全景分割方法进行了比较. 如表 3 和表 4 所示,

比较方法为基于 CNN 的全景分割传统算法, 包括以 Xception71 为主干的 DeeperLab 和 Panoptic-DeepLab 和以 ResNet50 为主干的 JSIS-Net[53]、AdaptIS[54]、Panoptic FPN[38]、OANet[14]、AUNet[13]、TASCNet[12]、SpatialFlow[55] 和 UPSNet[15].

表 3 MS COCO 全景分割实验结果对比

Method	Backbone	*PQ*	PQ^{Th}	PQ^{St}	*SQ*	*RQ*
DeeperLab[16]	Xception71	34.3	37.5	29.6	77.1	43.1
Panoptic-DeepLab[17]	Xception71	39.7	43.9	33.2	—	—
JSIS-Net[53]	ResNet50	26.9	29.3	23.3	72.4	35.7
AdaptIS[54]	ResNet50	35.9	40.3	29.3	—	—
Panoptic FPN[38]	ResNet50	39.0	45.9	28.7	—	—
OANet[14]	ResNet50	39.0	48.3	24.9	77.1	47.8
AUNet[13]	ResNet50	39.6	**49.1**	25.2	—	—
TASCNet[12]	ResNet50	40.7	47.0	31.0	**78.5**	50.1
SpatialFlow[55]	ResNet50	40.9	46.8	31.9	—	—
UPSNet[15]	ResNet50	42.5	48.5	33.4	78.0	52.4
CBT	ResNet50	**42.9**	48.8	**33.8**	78.1	**52.8**

表 4 Cityscapes 全景分割实验结果对比

Method	Backbone	*PQ*	PQ^{Th}	PQ^{St}	*SQ*	*RQ*
DeeperLab[16]	Xception71	56.5	37.5	29.6	77.1	43.1
Panotic-DeepLab[17]	Xception71	63.0	—	—	—	—
TASCNet[12]	ResNet50	55.9	50.5	59.8	—	—
AUNet[13]	ResNet50	56.4	52.7	59.0	—	—
Panoptic FPN[38]	ResNet50	57.7	51.6	62.2	—	—
SpatialFlow[55]	ResNet50	58.6	54.9	61.4	—	—
AdaptIS[54]	ResNet50	59.0	**55.8**	61.3	—	—
UPSNet[15]	ResNet50	59.3	54.6	62.7	79.7	73.0
CBT	ResNet50	**59.4**	55.0	**62.7**	**79.7**	**73.2**

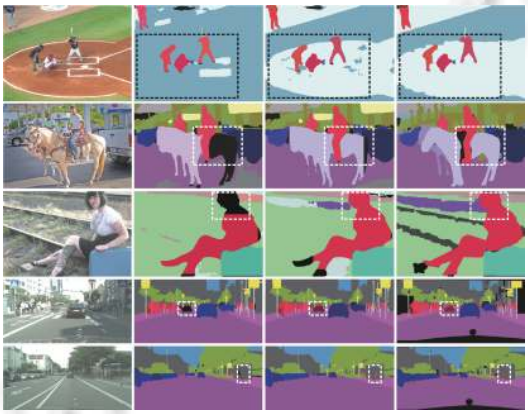
总体上, 基于 CNN 的全景分割 Transformer 网络的性能优于现存全景分割算法, *PQ* 值比 UPSNet 最高提升了 0.4. 在 MS COCO 数据集上, 与其他算法相比, CBT 的 *PQ* 和 *RQ* 达到了最高值, 体现了基于 CNN 的 Transformer 模型的有效性. 对于 Cityscapes 数据集, CBT 则实现了在 *PQ*、*SQ* 和 *RQ* 这 3 个指标上的全面领先, 说明了该算法在应对复杂交通场景时的巨大潜力. 另外需要说明的是, 与其他全景分割方法相比, CBT 在 PQ^{Th} 和 *SQ* 上的优势并不十分明显, 这是因为映射器 Lattice 卷积的空间大小在一定程度上限制了网络对距离较远的小目标的识别能力.

表 5 给出了提出算法与 DeeperLab、Panoptic-DeepLab 和 UPSNet 在 MS COCO 和 Cityscapes 数据集上的运行时间的对比. 实验是在单个 GPU 上完成的, 表 5 中还列出了每个算法的主干结构、*PQ* 值以及输入图像尺寸的大小. 由表 5 可见, 在同样使用 ResNet50 作为主干的算法中, 与 UPSNet 相比, CBT 在运行时间损失较小的情况下实现了全景分割精度的显著提升.

图 7 比较了提出算法和传统 CNN 算法[15]的分割结果, 分割结果表明前者的性能优于后者. 图 7 中的前 3 行反映了 CBT 在全局语义分析中的准确性. 即使在第 2 行的遮挡场景中, CBT 依然可以准确地分析空间关系并分割一个完整的对象. CBT 对图像前景和背景的分割表现也十分优越, 如图中第 1 行的网球场和第 3 行的前景人物. 最后两行显示了 CBT 在小物体识别方面的优势. 它可以检测街道场景中 CNN 算法遗漏或错误分割的小物体, 说明了 CBT 提取物体细节的能力.

表 5 运行时间对比

数据集	Method	Backbone	Input size	*PQ*	Speed (ms)
MS COCO	DeeperLab[16]	Xception71	641×641	34.3	119
	Panoptic-DeepLab[17]	Xception71	641×641	39.7	132
	UPSNet[15]	ResNet50	800×1300	42.5	167
	CBT	ResNet50	800×1300	42.9	174
Cityscapes	DeeperLab[16]	Xception71	1025×2049	34.3	463
	Panotic-DeepLab[17]	Xception71	1025×2049	63.0	175
	UPSNet[15]	ResNet50	1024×2048	59.3	202
	CBT	ResNet50	1024×2048	59.4	208



(a) Image (b) CNN (c) CBT (d) Ground truth

图 7 CBT 和 CNN 的全景分割结果对比图

4.3 消融实验

在本节中, 我们开展映射器和提取器的消融实验, 分析它们在基于 CNN 的 Transformer 网络中的功能和作用. 全部消融实验均在 MS COCO 验证集上使用单个 GPU 进行验证.

4.3.1 映射器

表 6 显示了 P_{L1} (实例分割所用映射器) 和 P_{L2} (语义分割所用映射器) 的消融实验结果. 从表中可以看出, 映射器的应用使得评价指标 PQ^{St} 和 PQ^{Th} 具有明显提高, 在 *PQ* 上带来了 0.9% 的绝对改进, 这归功于映射器的域转换能力. P_{L1} 将输入映射为细节特征, 增强了特征显著性, 提高了网络区分不同个体的能力. P_{L2} 使得网络对全局信息和目标边缘的提取变得更加容易.

表 6 MS COCO 映射器消融实验

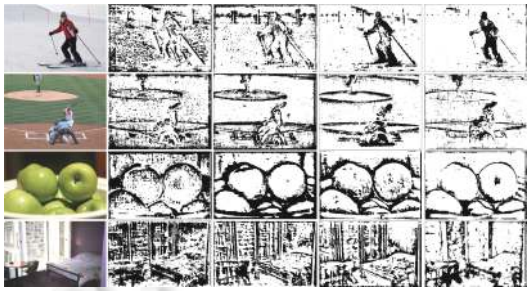
Model	Backbone (E_{C1})	Projector		Extractor (E_{C2})	*PQ*	PQ^{Th}	PQ^{St}
		Thing (P_{L1})	Stuff (P_{L2})				
M1	√			√	38.1	44.0	29.1
M2	√		√	√	38.5	44.3	29.8
M3	√	√		√	38.2	44.2	29.1
M4	√	√	√	√	**39.0**	**44.8**	**30.1**

Lattice 卷积模型: 表 7 给出了用作映射器的 Lattice 卷积模型的实验结果. 本实验的基线网络是以残差网络为骨干, 采用 Lattice 卷积模型, 并连接实例[2]和语义[3]分割结构的标准网络. 我们使用没有映射器的算法作为性能比较的基准.

表 7 说明, 随着 Lattice 卷积稀疏率的增加, 分割质量逐渐提高, 反映 Lattice 卷积模型对图像空间关系建模是有效的. 尤其是稀疏率最大的 P_{L4} 表现出了最好的性能, 说明对卷积滤波器空间利用率的研究具有重要意义. 本研究旨在设计卷积滤波器进行边缘特征提取以实现准确分割. 图 8 显示了 Lattice 卷积模型的可视化结果. Lattice 卷积模型提取到的边缘特征随着稀疏率的增加变得更加显著, 验证了 Lattice 卷积在边缘特征提取上的有效性.

表 7 Lattice 卷积模型消融实验

Model	S	PQ	SQ	RQ
None	—	38.0	76.4	47.7
P_{L1}	S_{25}^{0}	38.2	76.7	47.7
P_{L2}	S_{25}^{8}	38.3	76.7	47.9
P_{L3}	S_{25}^{12}	38.3	76.6	47.7
P_{L4}	S_{25}^{16}	38.4	76.8	47.9



(a) Image (b) P_{L1} (c) P_{L2} (d) P_{L3} (e) P_{L4}

图 8 Lattice 卷积模型的可视化对比

4.3.2 提取器

我们在提取器中进行了链式网络的消融实验. 表 8 列出了包含不同链式单元个数的 E_{C2} 网络的分割结果. 消融实验的基线网络由 E_{C1}、P_{L1} 和 P_{L2} 组成, 并连接到包含不同链式单元的 E_{C2} 网络. 表 6 可见, 当 E_{C2} 包含两个链式单元时, 分割性能最好, 带来 0.6% 的提升, 体现了链式网络在特征提取方面的能力. 然而, 当堆叠链式单元的数量增加时, 由于特征信息冗余, 算法质量有所下降. 因此, 在实际应用中, 应根据具体任务谨慎确定链式单元的数量. 出于保守的原因, 初始实验可选择两个链式单元.

表 8 链式网络消融实验 (m 表示链式单元的个数)

m	PQ	SQ	RQ
0	38.4	76.4	48.1
2	39.0	76.8	48.6
4	37.9	76.0	47.2
6	37.7	76.0	47.0

5 结 论

本文提出了一种基于 CNN 的 Transformer 网络, 利用 CNN 在图像特征建模方面的优势来完成视觉任务. 我们创建了基于 CNN 的 Transformer 的网络框架, 该框架包含两个基本结构, 即用于特征域变换的映射器和用于特征提取的提取器. 映射器由模拟图像空间关系的 Lattice 卷积模型实现, 而提取器由具有深度堆叠能力的链式网络实现. 在框架下, 根据全景分割任务的功能需求, 有效组织映射器和提取器, 形成基于 CNN 的全景分割 Transformer 网络.

基于 CNN 的 Transformer 模型为 Transformer 在视觉任务中的应用开辟了一条新途径, 该模型对图像特征的敏感性使其自然适用于图像处理任务. CNN 构建的 Lattice 卷积设计了各种图像空间关系模型, 给卷积滤波器的配置设计和空间利用带来了很多思考. 基于 CNN 的链式网络也提供了一种新的特征提取方法.

References:

[1] Kirillov A, He KM, Girshick R, Rother C, Dollár P. Panoptic segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9396−9405. [doi: 10.1109/CVPR.2019.00963]

[2] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE,

2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[3] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431−3440. [doi: 10.1109/CVPR.2015.7298965]

[4] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495. [doi: 10.1109/TPAMI.2016.2644615]

[5] Zhao HS, Shi JP, Qi XJ, Wang XG, Jia JY. Pyramid scene parsing network. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230−6239. [doi: 10.1109/CVPR.2017.660]

[6] Li HC, Xiong PF, Fan HQ, Sun J. DFANet: Deep feature aggregation for real-time semantic segmentation. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9514–9523. [doi: 10.1109/CVPR.2019.00975]

[7] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597, 2015.

[8] Zhang ZL, Zhang XY, Peng C, Cheng DZ, Sun J. ExFuse: Enhancing feature fusion for semantic segmentation. arXiv:1804.03821, 2018.

[9] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

[10] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proc. of the 2015 Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1520−1528. [doi: 10.1109/ICCV.2015.178]

[11] Peng C, Zhang XY, Yu G, Luo GM, Sun J. Large kernel matters—Improve semantic segmentation by global convolutional network. arXiv:1703.02719, 2017.

[12] Li J, Raventos A, Bhargava A, Tagawa T, Gaidon A. Learning to fuse things and stuff. arXiv:1812.01192, 2019.

[13] Li YW, Chen XZ, Zhu Z, Xie LX, Huang G, Du DL, Wang XG. Attention-guided unified network for panoptic segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7019−7028. [doi: 10.1109/CVPR.2019.00719]

[14] Liu HY, Peng C, Yu CQ, Wang JB, Liu X, Yu G, Jiang W. An end-to-end network for panoptic segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6165−6174. [doi: 10.1109/CVPR.2019.00633]

[15] Xiong YW, Liao RJ, Zhao HS, Hu R, Bai M, Yumer E, Urtasun R. UPSNet: A unified panoptic segmentation network. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8810−8818. [doi: 10.1109/CVPR.2019.00902]

[16] Yang TJ, Collins MD, Zhu YK, Hwang JJ, Liu T, Zhang X, Sze V, Papandreou G, Chen LC. DeeperLab: Single-shot image parser. arXiv: 1902.05093, 2019.

[17] Cheng BW, Collins MW, Zhu YK, Liu T, Huang TS, Adam H, Chen LC. Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12472−12482. [doi: 10.1109/CVPR42600.2020.01249]

[18] Zheng SX, Lu JC, Zhao HS, Zhu XT, Luo ZK, Wang YB, Fu YW, Feng JF, Xiang T, Torr PHS, Zhang L. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers. arXiv:2012.15840, 2021.

[19] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. arXiv: 2005.12872, 2020.

[20] Lee-Thorp J, Ainslie J, Eckstein I, Ontanon S. FNet: Mixing tokens with Fourier transforms. arXiv:2105.03824, 2021.

[21] Wu HP, Xiao B, Codella N, Liu MC, Dai XY, Yuan L, Zhang L. CvT: Introducing convolutions to vision Transformers. arXiv:2103.15808, 2021.

[22] Dai JF, He KM, Sun J. Instance-aware semantic segmentation via multi-task network cascades. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3150–3158. [doi: 10.1109/CVPR.2016.343]

[23] Liu S, Qi L, Qin HF, Shi JP, Jia JY. Path aggregation network for instance segmentation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768. [doi: 10.1109/CVPR.2018.00913]

[24] Li Y, Qi HZ, Dai JF, Ji XY, Wei YC. Fully convolutional instance-aware semantic segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4438–4446. [doi: 10.1109/CVPR.2017.472]

[25] Liang XD, Lin L, Wei YC, Shen XH, Yang JC, Yan SC. Proposal-free network for instance-level object segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2978–2991. [doi: 10.1109/TPAMI.2017.2775623]

[26] Bai M, Urtasun R. Deep watershed transform for instance segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2858–2866. [doi: 10.1109/CVPR.2017.305]

[27] Uhrig J, Cordts M, Franke U, Brox T. Pixel-level encoding and depth layering for instance-level semantic labeling. In: Rosenhahn B,

Andres B, eds. Proc. of the 2016 German Conf. on Pattern Recognition. Hannover: Springer, 2016. 14–25. [doi: 10.1007/978-3-319-45886-1_2]

[28] Chen LC, Zhu YK, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611, 2018.

[29] Naresh YG, Little S, O'Connor NE. A residual encoder-decoder network for semantic segmentation in autonomous driving scenarios. In: Proc. of the 26th European Signal Processing Conf. Rome: IEEE, 2018. 1052–1056. [doi: 10.23919/EUSIPCO.2018.8553161]

[30] Wang Y, Zhou Q, Liu J, Xiong J, Gao GW, Wu XF, Latecki LJ. Lednet: A lightweight encoder-decoder network for real-time semantic segmentation. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing. Taipei: IEEE, 2019. 1860–1864. [doi: 10.1109/ICIP.2019.8803154]

[31] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.

[32] Dai JF, Qi HZ, Xiong YW, Li Y, Zhang GD, Hu H, Wei YC. Deformable convolutional networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 764–773. [doi: 10.1109/ICCV.2017.89]

[33] Li HC, Xiong PF, An J, Wang LX. Pyramid attention network for semantic segmentation. arXiv:1805.10180, 2018.

[34] Zhong ZL, Lin ZQ, Bidart R, Hu XD, Daya IB, Li ZF, Zheng WS, Li J, Wong A. Squeeze-and-attention networks for semantic segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13062–13071. [doi: 10.1109/CVPR42600.2020.01308]

[35] Fu J, Liu J, Tian HJ, Li Y, Bao YJ, Fang ZW, Lu HQ. Dual attention network for scene segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149. [doi: 10.1109/CVPR.2019.00326]

[36] Huang ZL, Wang XG, Wei YC, Huang LC, Shi H, Liu WY, Huang TS. CCNet: Criss-cross attention for semantic segmentation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 603–612. [doi: 10.1109/TPAMI.2020.3007032]

[37] Zhao HS, Zhang Y, Liu S, Shi JP, Loy CC, Lin DH, Jia JY. PSANet: Point-wise spatial attention network for scene parsing. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. Proc. of the 2018 European Conf. on Computer Vision. Munich: Springer, 2018. 270–286. [doi: 10.1007/978-3-030-01240-3_17]

[38] Kirillov A, Girshick R, He KM, Dollár P. Panoptic feature pyramid networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6392–6401. [doi: 10.1109/CVPR.2019.00656]

[39] Mohan R, Valada A. EfficientPS: Efficient panoptic segmentation. arXiv:2004.02307, 2021.

[40] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2021.

[41] Hu H, Zhang Z, Xie ZD, Lin S. Local relation networks for image recognition. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3463–3472. [doi: 10.1109/ICCV.2019.00356]

[42] Zhao HS, Jia JY, Koltun V. Exploring self-attention for image recognition. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10073–10082. [doi: 10.1109/CVPR42600.2020.01009]

[43] Wang HY, Zhu YK, Green B, Adam H, Yuille A, Chen LC. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In: Vedaldi A, Bischof H, Brox T, Frahm JM, eds. Proc. of the 2020 European Conf. on Computer Visio. Glasgow: Springer, 2020. 108–126. [doi: 10.1007/978-3-030-58548-8_7]

[44] Li ZS, Liu XT, Drenkow N, Ding A, Creighton FX, Taylor RH, Unberath M. Revisiting stereo depth estimation from a sequence-to-sequence perspective with Transformers. arXiv:2011.02910, 2021.

[45] Liu RJ, Yuan ZJ, Liu T, Xiong ZL. End-to-end lane shape prediction with Transformers. arXiv:2011.04233, 2020.

[46] Wang YQ, Xu ZL, Wang XL, Shen CH, Cheng BS, Hao S, Xia HX. End-to-end video instance segmentation with Transformers. arXiv:2011.14503, 2021.

[47] Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial Transformer networks. arXiv:1506.02025, 2016.

[48] Bhagavatula C, Zhu CC, Luu K, Savvides M. Faster than real-time facial alignment: A 3D spatial Transformer network approach in unconstrained poses. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 4000–4009. [doi: 10.1109/ICCV.2017.429]

[49] Kasem HM, Hung KW, Jiang JM. Revised spatial Transformer network towards improved image super-resolutions. In: Proc. of the 24th IEEE Int'l Conf. on Pattern Recognition. Beijing: IEEE, 2018. 2688–2692. [doi: 10.1109/ICPR.2018.8546080]

[50] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick LC. Microsoft COCO: Common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. Proc. of the 2014 European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[51] Cordts M, Omran M, Ranos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth F, Schiele B. The Cityscapes dataset for semantic

urban scene understanding. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3213–3223. [doi: 10.1109/CVPR.2016.350]

[52] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin ZM, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. In: Proc. of the 31st Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1–4.

[53] de Geus D, Meletis P, Dubbelman G. Panoptic segmentation with a joint semantic and instance segmentation network. arXiv: 1809.02110, 2019.

[54] Sofiiuk K, Sofiyuk K, Barinova O, Konushin A, Barinova O. AdaptIS: Adaptive instance selection network. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7354–7362. [doi: 10.1109/ICCV.2019.00745]

[55] Chen Q, Cheng AD, He XY, Wang PS, Cheng J. SpatialFlow: Bridging all tasks for panoptic segmentation. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 31(6): 2288–2300. [doi: 10.1109/TCSVT.2020.3020257]

毛琳(1977－), 女, 博士, 副教授, 主要研究领域为目标跟踪, 多传感器信息融合.

杨大伟(1978－), 男, 博士, 副教授, 主要研究领域为图像处理, 计算机视觉.

任凤至(1995－), 女, 硕士生, 主要研究领域为计算机视觉, 图像分割.

张汝波(1962－), 男, 博士, 教授, 主要研究领域为智能机器人技术, 智能信息处理技术.