©中国图象图形学报版权所有

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2024)05-1119-27

论文引用格式:Gao C X, Xu Z Z, Wu D Y, Yu C Q and Sang N. 2024. Deep learning-based real-time semantic segmentation: a survey. Journal of Image and Graphics, 29(05):1119-1145(高常鑫, 徐正泽, 吴东岳, 余昌黔, 桑农. 2024. 深度学习实时语义分割综述. 中国图象图形学报, 29(05):1119-1145)[DOI:10.11834/jig.230659]

深度学习实时语义分割综述

高常鑫[1,2],徐正泽[1,2],吴东岳[1,2],余昌黔[3],桑农[1*]

1. 华中科技大学人工智能与自动化学院,武汉　430074; 2. 类脑智能系统湖北省重点实验室,武汉　430074;
3. 北京三快科技有限公司(美团),北京　100102

摘　要:语义分割是计算机视觉领域的一项像素级别的感知任务,目的是为图像中的每个像素分配相应类别标签,具有广泛应用。许多语义分割网络结构复杂,计算量和参数量较大,在对高分辨率图像进行像素层次的理解时具有较大的延迟,这极大限制了其在资源受限环境下的应用,如自动驾驶、辅助医疗和移动设备等。因此,实时推理的语义分割网络得到了广泛关注。本文对深度学习中实时语义分割算法进行了全面论述和分析。1)介绍了语义分割和实时语义分割任务的基本概念、应用场景和面临问题;2)详细介绍了实时语义分割算法中常用的技术和设计,包括模型压缩技术、高效卷积神经网络(convolutional neural network,CNN)模块和高效Transformer模块;3)全面整理和归纳了现阶段的实时语义分割算法,包括单分支网络、双分支网络、多分支网络、U型网络和神经架构搜索网络5种类别的实时语义分割方法,涵盖基于CNN、基于Transformer和基于混合框架的分割网络,并分析了各类实时语义分割算法的特点和局限性;4)提供了完整的实时语义分割评价体系,包括相关数据集和评价指标、现有方法性能汇总以及领域主流方法的同设备比较,为后续研究者提供统一的比较标准;5)给出结论并分析了实时语义分割领域仍存在的挑战,对实时语义分割领域未来可能的研究方向提出了相应见解。本文提及的算法、数据集和评估指标已汇总至https://github.com/xzz777/Awesome-Real-time-Semantic-Segmentation,以便后续研究者使用。

关键词:实时语义分割;模型轻量化;高效模块设计;计算机视觉;深度学习

Deep learning-based real-time semantic segmentation: a survey

Gao Changxin[1,2], Xu Zhengze[1,2], Wu Dongyue[1,2], Yu Changqian[3], Sang Nong[1*]

1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan　430074, China;
2. Hubei Key Laboratory of Brain-Inspired Intelligent Systems, Wuhan　430074, China;
3. Beijing Sankuai Technology Co., Ltd., (Meituan), Beijing　100102, China

Abstract: Semantic segmentation is a fundamental task in the field of computer vision, which aims to assign a category label to each pixel in the input image. Many semantic segmentation networks have complex structures, high computational costs, and massive parameters. As a result, they introduce considerable latency when performing pixel-level scene understanding on high-resolution images. These limitations greatly restrict the applicability of these methods in resource-constrained scenarios, such as autonomous driving, medical applications, and mobile devices. Therefore, real-time semantic segmentation methods, which produce high-precision segmentation masks with fast inference speeds, receive widespread attention. This study provides a systematic and critical review of real-time semantic segmentation algorithms

收稿日期:2023-09-12;**修回日期**:2023-12-26;**预印本日期**:2024-01-02

*通信作者:桑农　nsang@hust.edu.cn

基金项目:国家自然科学基金项目(62176097);湖北省自然科学基金项目(2022CFA055)

Supported by: National Natural Science Foundation of China(62176097); Natural Science Foundation of Hubei Province, China(2022CFA055)

based on deep learning techniques to explore the development of real-time semantic segmentation in recent years. Moreover, it covers three key aspects of real-time semantic segmentation: real-time semantic segmentation networks, mainstream datasets, and common evaluation indicators. In addition, this study conducts a quantitative evaluation of the real-time semantic segmentation methods discussed and provides some insights into the future development in this field. First, semantic segmentation and real-time semantic segmentation tasks and their application scenarios and challenges are introduced. The key challenge in real-time semantic segmentation mainly lies on how to extract high-quality semantic information with high efficiency. Second, some preliminary knowledge for studying real-time semantic segmentation algorithms is introduced in detail. Specifically, this study introduces four kinds of general model compression methods: network pruning, neural architecture search, knowledge distillation, and parameter quantification. It also introduces some popular efficient CNN modules in real-time semantic segmentation networks, such as MobileNet, ShuffleNet, EfficientNet, and efficient Transformer modules, such as external attention, SeaFormer, and MobileViT. Then, existing real-time semantic segmentation algorithms are organized and summarized. In accordance with the characteristics of the overall network structure, existing works are categorized into five categories: single-branch, two-branch, multibranch, U-shaped, and neural architecture search networks. Specifically, the encoder of a single-branch network is a single-branch hierarchical backbone network, and its decoder is usually lightweight and does not involve complex fusion of multiscale features. The two-branch network adopts a two-branch encoder structure, using one branch to capture spatial detail information and the other branch to model semantic context information. Multibranch networks are characterized by a multibranch structure in the encoder part of the network or a network with multiresolution inputs, where the input of each resolution passes through a different subnetwork. The U-shaped network has a contracting encoder and an expansive decoder, which are roughly symmetrical to the encoder. Most works of these aforementioned four categories are manually designed, while the neural architecture search networks are obtained using network architecture search technology based on the four types of architectures. These five categories of real-time semantic segmentation methods cover almost all real-time semantic segmentation algorithms based on deep learning, including CNN-based, Transformer-based, and hybrid-architecture-based segmentation networks. Moreover, commonly used datasets and evaluation indicators of accuracy, speed, and model size are introduced for real-time segmentation. We divided popular datasets into the autonomous driving scene and general scene datasets, and the evaluation indicators are divided into accuracy indicators and efficiency descriptors. In addition, this study quantitatively evaluates various real-time semantic segmentation algorithms mentioned on multiple datasets by using relevant evaluation indicators. To avoid the interference of different devices when conducting a quantitative comparison between real-time semantic segmentation algorithms, this study compares the performance of advanced methods of each category with the same devices and configuration and establishes a fair and integral real-time semantic segmentation evaluation system for subsequent research, thereby contributing to a unified standard for comparison. Finally, current challenges in real-time semantic segmentation are discussed, and possible future directions for improvements are envisioned (e.g., utilization of Transformers, applications on edge devices, knowledge transfer of visual foundation models, diversity of evaluation indicators, variety of datasets, utilization of multimodal data and weakly supervised methods, combination with incremental learning). The algorithms, datasets, and evaluation indicators mentioned in this paper are summarized at https://github.com/xzz777/Awesome-Real-time-Semantic-Segmentation for the convenience of subsequent researchers.

Key words: real-time semantic segmentation; lightweight model design; efficient module design; computer vision; deep learning

0 引 言

语义分割任务是计算机视觉长期以来一项基础的、像素级别的感知与理解任务，旨在为输入图像的每个像素分配到对应的类别标签。语义分割在很多实际应用中都起着至关重要的作用，如医学图像处理、机器人视觉、遥感图像分类、增强现实、图像的压缩与传送、自动驾驶视觉以及智能视频分析等领域。

在深度学习出现之前，语义分割算法主要基于

传统图像分割，包括多种基于区域和基于边界的算法，如OTSU法（Otsu，1979）、K均值聚类（Dhanachandra等，2015）、分水岭（Najman和Schmitt，1994）、区域生长法（Nock和Nielsen，2004）、活动轮廓（Kass等，1988）、图割法（Boykov等，2001）、条件随机场和马尔可夫随机场（Plath等，2009）等。深度学习的提出，使语义分割算法的性能得到显著提高，有力地促进了语义分割相关应用的发展。全卷积网络（fully convolutional network，FCN）（Long等，2015）首先被提出用于语义分割。在FCN的启发下，卷积神经网络（convolutional neural network，CNN）逐渐变为图像分割算法的新范式。然而，由于语义分割需要恢复下采样中损失的信息，因此多尺度的特征信息、长距离的上下文对语义分割的精度有着显著提升，因此许多语义分割模型提出了多种多样的方法来捕获丰富的上下文信息，如扩大感受野、多尺度融合和自注意力机制等。研究者或者提出更好的特征提取网络，如VGG（Visual Geometry Group）（Simonyan和Zisserman，2015）、GoogLeNet（Szegedy等，2015）、ResNet（residual network）（He等，2016）、HRNet（high-resolution network）（Sun等，2019）等；或者提出更好的上下文捕获方法，如U-Net（Ronneberger等，2015）、PSPNet（pyramid scene parseing network）（Zhao等，2017）、RefineNet（refinement networks）（Lin等，2017）；或者提出各种添加了注意力模块的CNN网络（Hu等，2019；Yuan等，2021；Huang等，2019）等。

Dosovitskiy等人（2021）提出视觉Transformer的概念，将自然语言处理（natural language processing，NLP）中使用的Transformer机制引入到计算机视觉领域后，许多研究者致力于将Transformer机制应用于语义分割领域。研究者在这方面进行了许多工作，SETR（segmentation Transformer）（Zheng等，2021）首次将视觉Transformer直接应用到图像分割，PVT（pyramid vision Transformer）（Wang等，2021）将CNN中常见的特征金字塔架构引入基于Transformer图像分割的模型，SegFormer（Xie等，2021）提出了一个简洁、高效且多尺度的Transformer图像分割模型，Liu等人（2021）提出的Swin Transformer网络更是在提出时代替CNN成为包括图像分割在内的许多计算机视觉任务上最先进的主干网络。这些工作不断提高图像分割算法在各种分割数据集上的最高水平。

尽管这些方法带来了显著的分割精度提升，但同时带来了高额的计算代价；尤其是自注意力机制和以自注意力机制为核心的Transformer网络，虽然具有全局建模能力，已被证明非常适合捕获长距离上下文，但是与图像分辨率呈平方复杂度，显著增大了语义分割模型的推理延迟。

然而现实应用中许多场景都需要语义分割模型具有实时性，例如移动端应用、自动驾驶和人机交互等。在这些情况下，上述繁重的语义分割网络往往具有不可接受的秒级延迟。为解决上述问题，一些基于深度学习的高效的实时语义分割方法被相继提出，能够以较小的延迟代价获得较高的分割精度。

一般而言，实时语义分割网络是指在指定设备上，推理时的帧率能够达到30帧/s及以上（即人眼对视频流畅的最低帧率要求）的语义分割网络。

为了得到更高的分割精度，语义分割模型同时需要丰富的空间细节信息和多尺度上下文信息。然而，一方面，丰富的空间细节信息需要保留高分辨率的底层特征图，这会极大地增加计算代价；另一方面，多尺度上下文的捕获和融合又需要设计复杂的模块和交互，这会增加推理的延迟。如何以更少的计算代价，保留更丰富的空间信息、捕获更有效的多尺度上下文，以得到更好的模型速度—精度平衡，是实时分割领域一直以来的挑战和领域内研究者们一直以来的追求。此外，在一些资源受限的移动设备和边缘设备上，模型的大小和内存占用量也显得至关重要，在这些设备上的实时语义分割网络如何进行优化设计也是实时分割领域面临的一项挑战。

本文结构框架如图1所示，分为5个部分：1）引言，简要介绍实时语义分割的定义、应用以及面临的挑战；2）前置知识，主要介绍实时语义分割网络设计中常用的一些方法，包括模型压缩的通用方法，以及轻量卷积神经网络和轻量Transformer网络设计中常用的高效模块；3）方法分类，主要按照网络结构类型将截止到目前的实时语义分割方法归纳为五大类来介绍，包括单分支网络、双分支网络、多分支网络、U型网络和神经架构搜索网络。另外，还从网络框架类型、应用场景类型对实时语义分割网络进行了分类。4）评价体系，对实时语义分割任务提供了系统

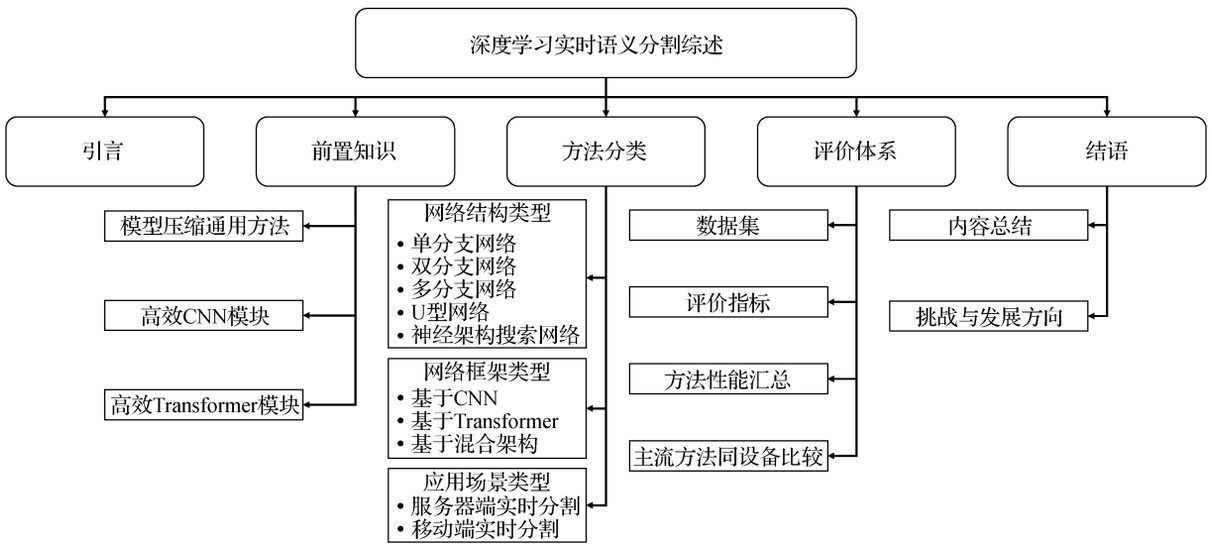


图1　本文结构框架图

Fig. 1　Overall framework of this review

的评估方法，包括数据集、评估指标、方法性能汇总和当前主流方法同设备比较。5)结语，总结目前实时分割语义分割网络的设计趋势以及面临的挑战和未来可能的发展方向。

1　前置知识

1.1　模型压缩通用方法

模型压缩是指将原本的大网络模型通过一些技术手段，压缩成为具有更好的实时性或参数量更小的模型。常见的模型压缩技术包括网络剪枝、神经架构搜索、知识蒸馏和量化。

网络剪枝(network pruning)是指去掉网络模型中不必要的参数。网络剪枝的一般步骤是：训练一个大网络、评估每个参数的重要性、去掉不重要的参数以及微调去掉参数后的网络以恢复剪枝损失的部分精度。剪枝可以利用大模型本身容易训练到较高精度的优势，以最小的精度损失代价来获得更小的模型。

神经架构搜索(neural architecture search, NAS)是一种利用强化学习方法同时学习模型架构和相应参数的方法。简单来说，就是在一个定义好的搜索空间内，通过一定的搜索策略，得到最终表现最好的网络。通过加入准确率、推理延迟等指标，网络架构搜索产生的网络结构在轻量化应用中能获得更高的竞争力。

知识蒸馏(knowledge distillation)利用大型教师模型网络参数包含的知识监督小型学生模型，使其能够在一定程度上拟合大的教师模型的输出，从而提高学生模型的精度，以得到更高精度的紧凑小模型。

量化(quantization)是指通过一定技术手段降低模型的数字精度以达到压缩模型、加快推理速度的效果，是模型部署常用的技术之一。

1.2　高效CNN模块设计

高效CNN模块设计一直是卷积网络设计的热点之一。本小节主要介绍实时分割中最常用的几种高效CNN模块，包括ShuffleNet、MobileNet和EfficientNet系列。

Krizhevsky等人(2012)提出了分组卷积(group convolution)，先将特征通道分组，再卷积操作限制在对应组内，减少了浮点运算量和参数量。然而，分组阻碍了通道之间的信息流，削弱了网络的表达能力。为此，ShuffleNet(Zhang等，2018)提出使用通道洗牌(channel shuffle)操作可以用较小的代价来实现组卷积通道间的信息交互，得到可观的精度提升。ShuffleNet的模块如图2所示。图2左侧为具有逐点组的ShuffleNet单元，其中GConv(group convolution)表示分组点卷积；图2右侧为下采样步长为2的ShuffleNet模块，其中残差边使用平均池化(AVG pool)进行快速下采样。ShuffleNetV2在ShuffleNet的基础上进行了一系列网络延迟分析实验，以通道划分操作

代替了分组卷积，并调整了通道洗牌的操作的位置，以通道拼接和通道洗牌操作代替了像素级别的相加；这些优化设计使得 ShuffleNetV2 相比 ShuffleNet 获得了更高的运行效率。

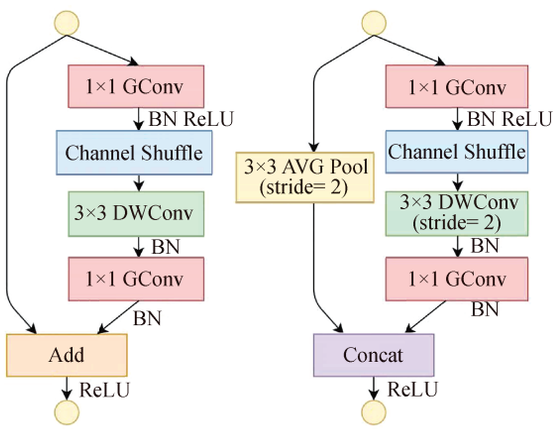


图 2　ShuffleNet（Zhang 等，2018）模块结构示意图

Fig. 2　Schematic diagram of the ShuffleNet module (Zhang et al. , 2018)

MobileNet（Howard 等，2017）广泛使用深度可分离卷积来提高网络效率。深度可分离卷积将标准卷积拆分为深度卷积和点卷积，其中深度卷积为分组数等于通道数的组卷积。MobileNetV2（Sandler 等，2018）引入了倒置残差模块（inverted residual block）。倒置残差模块的结构如图 3（a）所示，由一个升维的点卷积、一个深度卷积、一个降维的点卷积以及残差连接组成。同时，由于 ReLU（rectified linear unit）激活函数应用于低维特征会带来较大的信息损失，因此倒置残差模块中降维点卷积后没有加入激活函数，这称为线性瓶颈块（linear bottleneck）。此外，MobileNetV2 使用的激活函数为 ReLU6，ReLU6 相比 ReLU 在移动端低比特的量化模型上具有更好的鲁棒性。截止到目前，MobileNetV2 模块仍然被许多实时语义分割网络广泛作为基础模块。MobileNetV3（Howard 等，2019）在 MobileV2 的倒置残差块中加入了 SE（squeeze and excitation）模块（Hu 等，2018），如图 3（b）所示。另外，MobileNetV3 引入了 h-swish 作为非线性激活函数，同时引入了网络搜索技术来进一步提高网络效率。

EfficientNet（Tan 和 Le，2020）分析了模型在深度、宽度和分辨率上缩放的影响，基于 MobileNetV2 模块使用网络搜索技术，EfficientNet 获得了不同大小的一系列紧凑高效的轻量化网络。EfficientNetV2（Tan 和 Le，2021）从充分利用优化器的角度改进了 MobileNetV3 模块，并将其加入搜索空间；另外，EfficientNetV2 采用了非均匀模型缩放策略和渐进式输入。

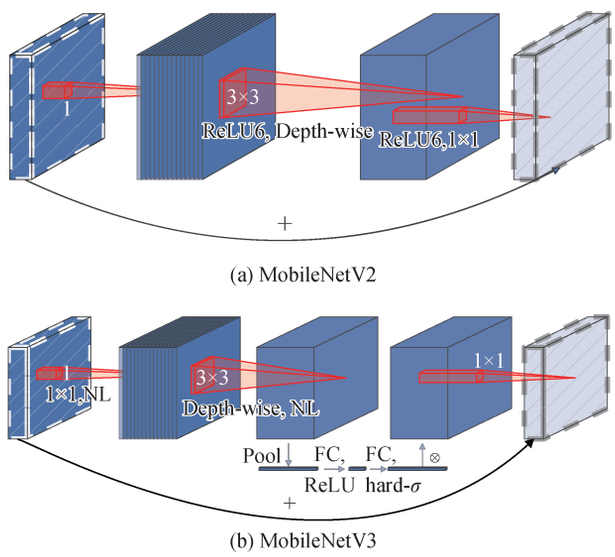


图 3　MobileNetV2 和 MobileNetV3（Howard 等，2019）模块结构示意图

Fig. 3　Schematic diagram of MobileNetV2 and MobileNetV3 module (Howard et al. , 2019)

((a)MobileNetV2;(b)MobileNetV3)

1.3　高效 Transformer 模块设计

本小节主要介绍在实时分割领域使用较多的高效的 Transformer 模块设计，包括外部注意力（external attention，EA）、轴向压缩增强 Transformer（squeeze-enhance axial Transformer，SeaFormer）以及 MobileViT 等。

EA（Guo 等，2023）将标准注意力的 query 仍保留原始特征图，而 key 和 value 修改为与图像无关的可学习参数 $\boldsymbol{M}_{\mathrm{k}}$ 和 $\boldsymbol{M}_{\mathrm{v}}$，如式（1）所示。

$$\boldsymbol{F}_{\mathrm{out}} = DoubleNorm(\boldsymbol{F}_{\mathrm{in}}\boldsymbol{M}_{\mathrm{k}}^{\mathrm{T}})\boldsymbol{M}_{\mathrm{v}} \tag{1}$$

式中 $\boldsymbol{M}_{\mathrm{k}}$，$\boldsymbol{M}_{\mathrm{v}} \in \mathbf{R}^{B\times M\times C}$，$\boldsymbol{F}_{\mathrm{in}} \in \mathbf{R}^{B\times N\times C}$，$B$ 为 batch 数量、C 为通道数、M 和 N 分别为特征点数量和外部参数数量。另外，外部注意力将产生注意力图时使用的激活函数修改为双重归一化 *DoubleNorm*，即在维度 N 上使用 softmax 激活函数，维度 M 上使用 L1 激活函数。外部注意力引入外部参数使得注意力的计算复杂度从 $O(H^2W^2)$ 降低至 $O(HW)$，且仅由两个额外的线性层实现，故推理延迟较低。

SeaFormer（Wan 等，2023）结构如图 4 所示，其

相比标准Transformer模块的主要改变是提出了轴向压缩增强注意力(squeeze-enhance axial attention),以解决原始注意力操作的大计算量和高延迟的问题。类似Ho等人(2019)提出的轴向注意力,SeaFormer将key和value在图像宽和高的方向上分别进行压缩。此外,SeaFormer进一步将query也在图像宽高方向分别压缩。为了弥补压缩造成的局部信息的丢失,SeaFormer将原始query、key、value通过深度可分离卷积获得特征图,并使用轴向压缩增强注意力的输出作为权重增强上述特征图。

除了上述高效Transformer模块,还有其他许多高效Transformer模块被提出,如MobileViT(Mehta和Rastegari,2022a,b;Wadekar和Chaurasia,2022)混合Transformer和CNN架构,在保持图像空间结构的前提下进行注意力操作;MobileFormer(Chen等,2022)利用Transformer进行全局信息的提取,并和卷积分支进行反复跨注意力交互;EdgeViT(Pan等,2022)使用局部卷积和代表token交互的稀疏注意力结合来模拟标准Transformer。这些方法都取得了不错的速度—精度平衡,但在实时语义分割领域应用较少,因此不在本节进行详细说明。

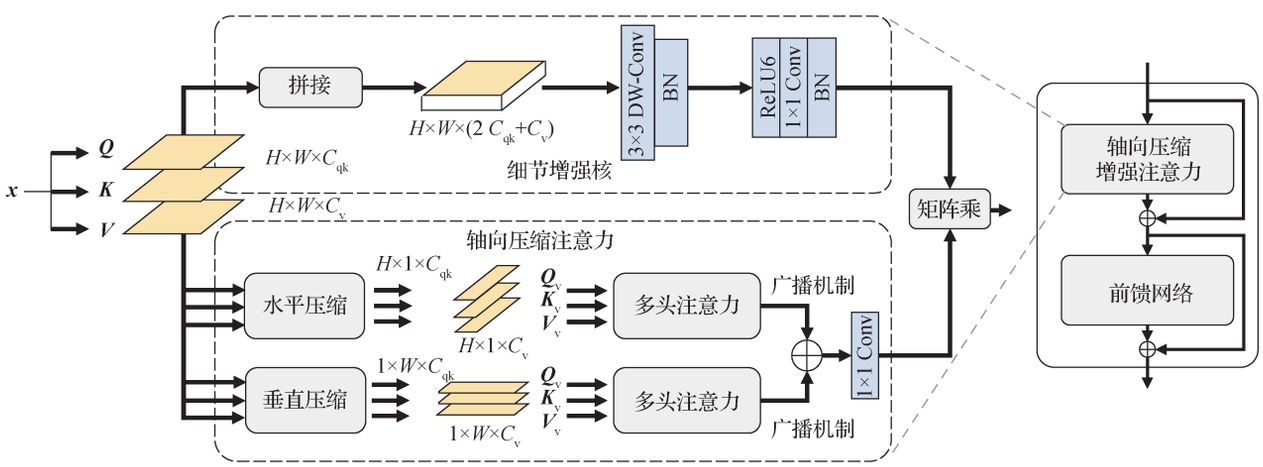


图4 SeaFormer(Wan等,2023)结构示意图

Fig. 4 Schematic diagram of the SeaFormer(Wan et al., 2023)

2 方法分类

本节归纳讨论了从提出实时语义分割的概念开始到目前为止,具有代表性的一些实时语义分割方法。

2.1 网络结构类型

从网络结构类型上将现有的实时语义分割网络大致分为5种类型:单分支网络、双分支网络、多分支网络、U型网络和神经架构搜索网络。这5类方法各自特点的简要介绍见表1。值得注意的是,在进行介绍时,每个子类中的具体方法并不按照时间顺序排列,而是按照其发展历程的相关性排列。有个别方法可能由多种类别衍生而出,本文根据其最重要的贡献来将其归到其中一类。

2.1.1 单分支网络

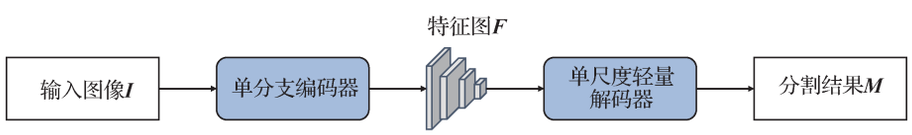
单分支网络共同的核心特点是编码器为单支层级式的主干网络。除此之外,单分支网络的解码器通常设计得十分轻量。根据轻量解码器的作用不同,单分支网络可以细分为3种:1)单尺度轻量解码器,解码器仅利用编码器最后的输出进行特征恢复和解码,如ENet;2)多尺度轻量解码器,解码器利用编码器的多尺度特征进行简单拼接融合,如SegFormer和SegNext;3)无解码器,彻底舍弃了解码器,直接对编码器特征进行分类输出,如DABNet(depthwise asymmetric bottleneck net)和AFFormer(adaptive frequency Transformer)。这3种类型的单分支网络的结构示意图如图5所示。因为轻量解码器承担的功能有限,单分支网络的编码器需要能够提取到表征能力较强的特征,这是单分支网络的设计重点,也是具体区分各单分支网络算法的关键。

1)单尺度轻量解码器。单尺度轻量解码器单分支网络的结构示意图如图5(a)所示,其中最具代

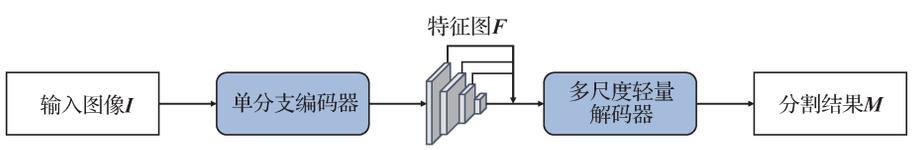
表 1　不同网络结构实时语义分割方法归纳

Table 1　Overview of different architectures of real-time semantic segmentation networks

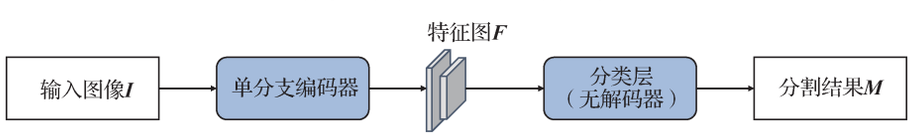
网络结构类别	结构特点	设计理念
单分支网络	编码器为单分支结构，解码器轻量	减少多支路和繁重解码器带来的延迟
双分支网络	编码器为解耦的空间、上下文双分支结构	解耦空间、上下文信息，以双分支分别提取
多分支网络	编码器为多分支结构或具有不同分辨率输入的多个子网络结构	利用多分支/多分辨率输入含有的不同信息
U 型网络	编码器、解码器为对称的 U 型结构，解码器逐步向上恢复特征	设计解码器以充分利用多尺度信息恢复图像特征
神经架构搜索网络	利用 NAS、剪枝技术优化网络，具体网络结构取决于用于搜索的超网结构	利用训练技术得到更优的网络结构



(a) 单尺度轻量解码器单分支结构



(b) 多尺度轻量解码器单分支结构



(c) 无解码器单分支结构

图 5　单分支实时分割网络结构示意图

Fig. 5　Schematic diagram of the single-branch real-time semantic segmentation networks ((a) single-scale light-weight decoder; (b) multi-scale light-weight decoder; (c) without decoder)

表性的是 ENet。

ENet（Paszke 等，2016）是最早探索实时性的语义分割网络。从网络结构上来看，ENet 的总体架构使用与 SegNet（Badrinarayanan 等，2017）相似的结构，编码器和解码器均为瓶颈层的堆叠，保存最大池化层的索引信息来辅助解码器的上采样操作。ENet 的核心贡献主要在于提出使用早期下采样和轻量化解码器。ENet 认为初始的网络层不应该直接有助于分类，而应该充当良好的特征提取器，只是对网络后续部分的输入进行预处理。因此，ENet 选择在前两个模块进行下采样，使用更小的特征图进行后续操作。另外，ENet 认为解码器的作用是对编码器的输出进行上采样，只对细节进行微调，编码器本身应具有信息处理和过滤的作用，因此 ENet 使用了小型解码器，得到了一个编解码不对称的单分支网络结构。后续的单分支网络均延续了 ENet 的这两个特点。

单尺度轻量解码器尽管带来推理速度的显著提升，但在使用轻量解码器时，低分辨率的单尺度特征输入使得图象细节难以恢复，导致该类实时分割网络的精度优势不大。

2）多尺度轻量解码器。多尺度轻量解码器相比单尺度轻量解码器而言，使用了编码器输出的更多尺度的特征，如图 5（b）所示，但其解码器仍然保持轻量级，例如，只对多尺度特征进行简单的通道数映射和通道拼接，如 SegFormer 和 SegNext。

SegFormer（Xie 等，2021）的结构十分简洁，是第 1 个在实时分割领域应用的 Transformer 网络。SegFormer 首先采用重叠的分块操作，即使用 3 × 3 卷积先将输入图像分块。之后是一个单分支层级式

的Transformer编码器，即典型的4阶段，每层为由Transformer模块和下采样模块堆叠的结构。与普通Transformer不同的是，SegFormer中使用的自注意力模块中的Key和Value是经过下采样的。即存在一个下采样因子R，使得Key和Value的维度从本身的$N \times C$变为$(N/R) \times C$，因此，注意力的复杂度从$O(N^2)$减小为$O(N^2/R)$。SegFormer的解码器十分轻量，是由几个线性层组成的多层感知机。编码器中4个阶段的特征通过线性层调整通道后统一上采样到输入图像的1/4，然后经过通道拼接后，直接经过线性层压缩通道后分类输出。SegFormer能够使用如此轻量的解码器受益于其编码器相比CNN网络更大的感受野和更强的特征建模能力。

SegNext（Guo等，2022）的整体网络架构与SegFormer非常相似，主要不同在于SegNext只使用阶段2—阶段4的特征进行通道拼接。Guo等人（2022）认为阶段1的特征包含太多过于底层的特征，会对精度造成负面影响。另外，SegNext的组成模块均为相比Transformer更加高效的多尺度卷积注意力（multi-scale convolutional attention，MSCA），其结构如图6所示。在图6（a）中，MSCAN（multi-scale convolutional attention network）每个阶段的结构与Transformer的结构相似，不同的是，归一化函数使用的是BN而非LN，且注意力为MSCA。MSCA的结构如图6（b）所示，主要组件包括：一个提取局部信息的深度卷积、多支不同尺度的深度条纹卷积和一个混合通道信息的点卷积。通过上述组件生成的特征图被作为注意力图和原始图像进行元素乘法操作。Guo等人（2022）证明了这种具有多尺度交互的卷积模块能够很好地模拟自注意力，达到高精度的语义分割。SegNext在不同规模的模型上都超越了SegFormer，成为目前性能最优秀的单分支实时分割网络之一。

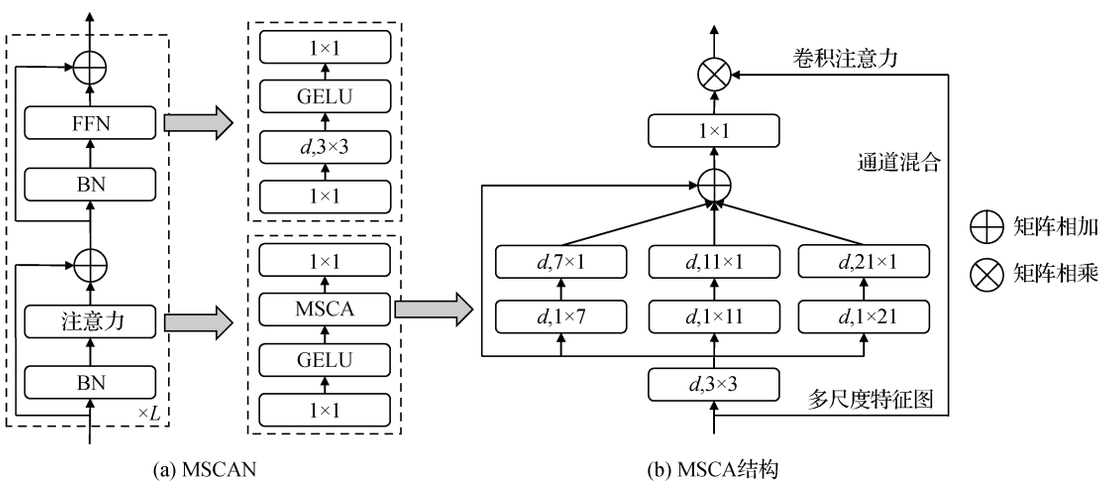


图6　SegNext中MSCAN的结构（Xie等，2021）

Fig. 6　The structure of the MSCAN in SegNext (Xie et al., 2021)((a)MSCAN;(b)MSCA)

多尺度轻量解码器引入了不同尺度的特征，能更好地进行特征解码和恢复。尽管引入多尺度特征会导致推理延迟的增加，但多尺度轻量解码器通常相比单尺度轻量解码器有着更高的速度—精度平衡。

3）无解码器。图5（c）所示的无解码器网络则直接舍弃了普通的分割解码器，直接使用分类头得到输出结果，其中的代表网络为DABNet和AFFormer。

DABNet（Li等，2019c）将轻量解码器推向了极致。具体来说，Li等人（2019c）认为解码器虽然有利于提升精度，但其拖慢网络推理效率的效应则更为严重；因此，所提出的DABNet完全放弃了解码器结构。输出经过一个1×1卷积和上采样即得到分割图。DABNet只进行了3次下采样，最小分辨率为原分辨率的1/8，这使得图像的恢复更加容易，无需解码操作。另外，也是由于只有3次下采样，DABNet需要更强大的模块来获得更大的感受野，以捕获长距离上下文。为此，DABNet设计了DAB模块。DAB模块由多个堆叠的3×3卷积、3×1和1×3的条纹卷积和条纹空洞卷积以及1×1卷积组成。空洞卷积使DAB模块能够有效扩大感受野，同时条纹

卷积的使用使其效率进一步提高。此外，DABNet模块将上述堆叠的卷积模块的输出和输入采用通道拼接的方式融合，以模拟空间、上下文双分支的效果。

王囡等人(2022)基于空洞卷积和注意力模块提出了一种单分支无解码器实时语义分割网络。为了获得更好的上下文，该方法使用空间注意力、通道注意力以及全局平均池化来增强上下文信息；同时，该方法利用空洞卷积，在获得更大感受野的同时，避免下采样倍数过大导致细节信息丢失。

Dong 等人(2023)提出的频率自适应 Transformer(adaptive frequency Transformer，AFFormer)，是最新的单分支实时分割网络。AFFormer的网络结构与DABNet相似，均舍弃了解码器，只保留简单的分类层，且图像分辨率只降至输入分辨率的1/8。AFFormer的每个阶段包括聚类操作、PL(prototype learning)和PD(pixel descriptor)。PL将标准Transformer中的自注意力模块替换为自适应频率滤波器(adaptive frequency filter，AFF)。具体来说，AFF包含频率相似度内核、动态低通滤波和动态高通滤波器3个组件，输入特征通过这3个组件的输入经过通道拼接后得到AFF的输出。PL的作用是动态地提取特征图的上下文信息；而PD的作用是保存输入图像(聚类前)的局部空间细节，与PL的输出进行融合，从而达到同时建模空间细节和上下文信息的目的。利用PL和PD双支的架构，AFFormer的编码器能够自适应地建模图像从局部到全局的信息，从而达到只需要一个简单的分类头即可完成语义分割，提高了推理效率。

无解码器单分支分割网络的整体网络结构简单，同时由于无需下采样到低分辨率，能更好地保留图像的细节信息。但也由于下采样倍数较小，导致特征感受野较小，特征分辨率较大，需要在主干网络设计时采用更加轻量化的模块以减小延迟、同时采用更强的长距离信息提取模块。

2.1.2 双分支网络

语义分割作为一项像素级别的稠密预测任务，除了需要具有长距离上下文外，也需要精细的局部空间信息。在通用的语义分割网络中，这两方面的需求是在多尺度的特征的交互、融合中耦合实现的。这种耦合的实现可能并非效率上的最优选择，因此研究者们提出了双分支网络。双分支网络建议使用一个分支进行空间细节信息的捕获，而用另外一个分支进行语义上下文信息的建模，以获得更好的分割精度和推理速度的平衡。如图7所示，双分支实时语义分割网络大致可以归为两个子类别：1)解耦双分支网络；2)特征共享双分支网络。

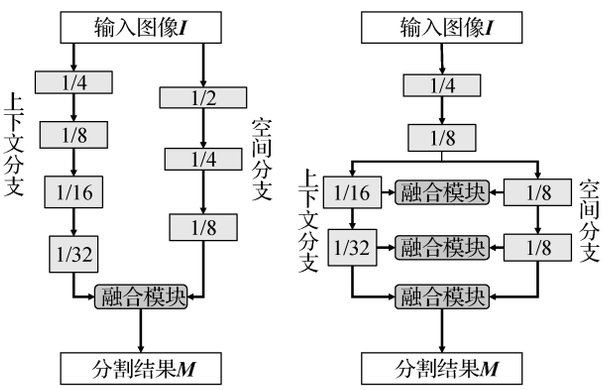


图7 双分支实时语义分割网络结构示意图

Fig. 7 Schematic diagram of the two-branch real-time semantic segmentation networks ((a) decoupled two-branch network; (b) feature sharing two-branch network)

1) 解耦双分支网络。如图7(a)所示，解耦双分支网络从输入图像开始，直接分为两个不同支路，分别提取图像的空间信息和上下文信息，并在网络的最后(解码器部分)进行两个分支输出特征的融合。解耦的双分支网络由BiSeNet首先提出，之后的BiSeNetV2、STDC则是在BiSeNet网络基础上的改进。

Yu 等人(2018)首次提出双边分割网络(bilateral segmentation network，BiSeNet)。BiSeNet的网络结构由空间路径分支与上下文路径分支两个并行分支组成，其中空间路径分支使用浅层结构保持较高特征分辨率，以保留丰富的底层细节信息，而上下文路径分支则使用深层结构与大感受野对特征进行迅速下采样，以较低计算代价提取准确的上下文语义信息。由于两分支提取的特征之间有较大差异，简单相加或通道拼接无法有效利用二者各自所含的信息，因此BiSeNet设计了特征融合模块(feature fusion module，FFM)对二者进行低计算代价融合。

在BiSeNet的基础上，BiSeNetV2(Yu 等，2021)为上下文路径分支设计了Stem模块、聚合拓展模块与上下文表征模块。其中Stem模块作为上下文路径分支的第1阶段，使用卷积与最大池化两种方式缩小特征分辨率，将二者输出拼接结合。聚合拓展

模块利用深度卷积拓展通道数量。该模块利用两个3×3卷积代替5×5卷积，在保持感受野的同时节省计算成本与内存访问成本。上下文路径分支的最后还具有一个上下文表征模块，使用全局池化获得更大的感受野捕获高级语义信息。另外，BiSeNetV2设计了双向引导聚合层、双向融合两分支特征。

虽然BiSeNet与BiSeNetV2已在实时分割领域依靠双分支结构取得了显著的精确度与推理速度提升，但相比单分支结构，其额外的分支，尤其是空间路径分支对空间信息的特征提取仍增加了不容忽视的计算量。另外，空间路径分支也同样缺乏底层信息监督。为了解决以上问题，Fan等人(2021)提出短期密集级联(short-term dense concatenate，STDC)网络。STDC网络主要由STDC模块组成，该模块通过精心调整卷积核的深度和个数，得到不同感受野的特征图，这些特征图经过通道维度的拼接，可充分利用大感受野与多尺度信息，产生高质量特征图。另外，STDC的空间路径分支仅使用stage 3的特征，通过对该特征上添加细节头(detailed head)进行边缘预测任务增强特征中的空间细节信息。该边缘预测任务在推理阶段不进行计算，STDC的空间路径分支以极小的代价提供了较为丰富的空间信息，极大削减了双分支结构带来的额外计算量。

2) 特征共享双分支网络。特征共享双分支网络本质上是解耦双分支网络的一种改进和优化，其结构示意图如图7(b)所示。特征共享双分支网络并不从网络起始阶段就将网络分为两个不同支路，而是在早期共享特征图，表现为单分支结构；在网络的深层分为空间和上下文双分支。代表性的特征共享双分支网络包括Fast-SCNN(fast segmentation convolutional neural network)、DDRNet(deep dual-resolution networks)、RTFormer(real-time Transformer)和SeaFormer。

受到BiSeNet提出的解耦双分支网络结构的启发，Fast-SCNN(Poudel等，2019)提出了一种改进的双分支网络来降低双分支网络早期重复的计算量。具体来说，Poudel等人(2019)认为，深度神经网络在早期主要提取底层特征，如边缘和角点；双分支网络中无论是空间分支还是上下文分支，其前几层卷积都是在提取这种底层特征和进行下采样。因此Fast-SCNN将双分支网络双分支的早期几层卷积融合为单支，称做学习下采样；在后面更深的层再分为上下文分支和空间分支。其中，上下文分支使用全局特征提取器，空间分支则简单保留特征；最终再将双分支的特征进行融合。Fast-SCNN提出的网络结构是特征共享双分支网络的雏形。

Pan等人(2023)提出的DDRNet则将BiSeNet的双分支结构修改为深层对偶分辨率网络，这一修改进一步探索了Fast-SCNN(Poudel等，2019)中提出的早期特征共享的网络结构。即DDRNet在网络的浅层(1/8分辨率之前)使用共享的单分支，而在网络的深层分为高、低分辨率双分支，分别提取空间信息和上下文信息，这种变化在一定程度上降低了双分支网络额外分支带来的推理延迟。同时，与BiSeNet仅在最后进行双分支特征融合相比，DDRNet选择了在双分辨率分支的每个阶段进行双边融合，这种稠密的双边融合相比后期的单次融合能够达到更好的效果，有助于提高分割精度。同时，DDRNet的双边融合模块仅由卷积调整通道、上下采样和相加等简单操作组成，尽可能降低了推理延迟。此外，DDRNet在上下文分支还采用了深度聚合池化金字塔模块(deep aggregation pyramid pooling module，DAPPM)进一步扩大感受野和整合多尺度上下文。DAPPM将特征图进行多尺度的池化，获得更深层的不同尺度的特征，并进行这些特征间的融合，形成多尺度特征。由于DAPPM使用在1/32分辨率的图像上，且池化会进一步缩小分辨率，因此DAPPM带来的延迟代价较小，但有明显的分割精度收益。

RTFormer(Wang等，2022)是对DDRNet的一种改进，其整体架构沿用了DDRNet的网络结构。主要区别在网络的深层，将对偶分支替换为RTFormer模块。RTFormer模块也是一种双分辨率的模块，分为高分辨率(1/8分辨率)和低分辨率(1/32分辨率)分支。其中低分辨率分支使用自注意力，即自身的特征图作为query、key和value；而高分辨率分支使用跨注意力，自身的特征图作为query，低分辨率的特征图通过自适应池化层下采样后的输出作为key和value。RTFormer使用的注意力模块和外部注意力模块(Guo等，2023)相似，但将外部注意力的分头操作转移到了激活函数中，使得矩阵乘法时仍保留完整的大矩阵；这样的注意力对GPU设备更加优化，推理延迟更低，被RTFormer称为GPU友好的注意力。另外，RTFormer将DDRNet的双边融合方式修改为逐步融合方式，即首先低分辨率融合至高分

辨率分支，通过高分辨率的模块后再进行和低分辨率特征的融合，之后再通过低分辨率的模块，提高了分割精度。

SeaFormer（Wan 等，2023）提出了一种混合Transformer-CNN的双分支实时语义分割网络结构，其总体的网络结构仍然采用DDRNet类似的特征共享双分支网络结构框架，但不同的是SeaFormer采用的是单边融合模块，即只是在每个阶段将上下文分支的特征融合到空间分支，而非上下文分支和空间分支的双向融合交互。SeaFormer的上下文分支使用SeaFormer层，这是一种部署友好的轻量化Transformer模块，在1.3节中已详细介绍过。另外，SeaFormer中大量使用了MobileNetV2模块等轻量化CNN模块，使得SeaFormer在移动端的部署更加高效。Transformer具有强大的长距离语义建模能力、缺乏偏置等特点，而CNN具有良好的局部建模能力、部署友好等优势；SeaFormer同时利用了Transformer和CNN的优势，具有当前移动端实时分割网络中最好的速度—精度平衡。

双分支网络通过对图像的双路解耦建模，能够高效地捕获语义分割所需的空间细节和长距离上下文信息，提高分割效率。但其额外的分支以及分支间的交互也带来了额外的计算代价和推理延迟。

2.1.3　多分支网络

多分支网络的特点是网络的编码器部分具有多个独立的分支结构，或由多个不完全重叠的子网结构组成。其网络结构的核心是利用多分支不同的结构来捕获图像不同层次的特征信息，或者利用不同分辨率图像的特点来设计更合适子网，以达到更好的速度与精度的平衡。常见的多分支实时语义分割网络包括ICNet（image cascade network）、ESPNet（efficient spatial pyramid net）、DFANet（deep feature aggregation net）和PIDNet（proportional integral derivative net），这些方法的结构特点以及其分支/子网的输入类型总结见表2。

ICNet（Zhao等，2018）首次采用了多分支结构来构建实时语义分割网络。ICNet利用低分辨率图像前向推理较快、高分辨率图像预测质量较高的特点，设计多分支网络，使低分辨率（1/4分辨率）图像经过深层网络获得粗糙的预测结果，中分辨率（1/2分辨率）与高分辨率（原始分辨率）图像被引入以逐渐提升预测结果的细节质量。ICNet提出了级联特征融合单元（cascaded fusion unit，CCF Unit），将较低分辨率图像产生的特征上采样，与较高分辨率的特征相融合。另外，ICNet将真值标签采样至多种分辨率，对同分辨率的预测输出进行监督。

ESPNet（Mehta等，2018）引入高效空间金字塔（efficient spatial pyramid，ESP）。先使用1 × 1点卷积对特征图进行降维，再将降维后的图像输入不同空洞率的空洞卷积；为了避免空洞卷积中的网格效应，ESPNet使用了分层图像融合（hierarchical feature fusion，HFF），即逐层将不同空洞率的空洞卷积的输出相加并向下传递，最后在通道维度将所有的输出进行通道拼接。ESPNet在网络结构上也采用了多分支的结构，除了原始分辨率的图像输入，1/2分辨率和1/4分辨率的图像也会在网络对应位置输入和特征图进行通道拼接。这种逐渐引入不同分辨率的图像以提升预测结果质量的方法和ICNet采用的方法相似。不同的是，ICNet的不同分辨率的图像由深度不同的单独的网络分支处理，而ESPNet的大分辨率图像和更小分辨率的图像拼接后会继续进入更深层的子网络。由于ESPNet的每个子网络共享其他网络的浅层部分的参数和结构，因此ESPNet的网络参数量较小。

DFANet（Li等，2019a）使用了额外子网络的结构，融合多尺度特征对高层特征进行精炼（refine）。相比于ICNet和ESPNet使用多分辨率的输入图像，DFANet利用将每个子网输出的多尺度特征重新输

表2　多分支实时语义分割方法归纳

Table 2　Overview of multi-branch real-time semantic segmentation methods

网络	结构特点	分支/子网输入类型
ICNet（Zhao等，2018）	独立的多个分支结构	多分辨率输入图像
ESPNet（Mehta等，2018）	多个共享的子网	多分辨率输入图像
DFANet（Li等，2019a）	多个独立的子网	多尺度特征图
PIDNet（Xu等，2023）	浅层共享、深层独立的多个分支结构	多尺度特征图

入新的子网络进行精炼。为了防止单一路径、大感受野和小尺度的高维特征缺少空间细节信息可能会导致的精度下降，DFANet还在解码器部分逐层使用较高分辨率特征，在大分辨率上进行特征精炼，最终恢复得到大尺度、具有丰富空间结构与细节信息的特征。

PIDNet（Xu等，2023）可以视为对多分支网络ICNet在分支结构共享和分支设计上的改进，也可以视为双分支网络DDRNet的分支结构的拓展。但从本质上来讲，Xu等人（2023）从PID控制算法的角度来思考实时语义分割网络的构建，最终得到一个三分支语义分割网络，PIDNet首先将输入图像输入一个单分支卷积网络逐步下采样到1/8分辨率的特征，然后使用3个分支来分别解析图像的细节信息（I分支）、上下文信息（P分支）和边界信息也即语义的导数（D分支），并在最后阶段使用边界注意力指导融合模块（boundary-attention-guided fusion module，BAG）来进行三分支信息的融合。具体来说，BAG分别用细节特征（I）和上下文特征（P）填充图像的高频和低频区域，其中填充的概率是由边界信息（D）的特征概率来确定的。PIDNet还设计了一种像素注意力引导的融合模块（pixel-attention-guided fusion module，Pag module），其作用是通过像素相似度，使P分支有选择地从I分支中学习有用的语义信息。在监督训练方面，PIDNet还使用真值标签以及真值标签经过Canny算子和膨胀操作所得的边界标签，分别去监督P分支和D分支，进一步提高了P分支和D分支提取对应特征的能力。PIDNet在推理速度和准确度之间实现了目前实时语义分割网络的最佳平衡。

多分支网络由于能够使用多个分支/子网处理不同分辨率的输入和不同尺度的特征，在分割精度方面具有优势。但多分支网络设计复杂，且额外的分支和子网会带来较大的延迟，需要进行特殊设计以加速网络，在推理速度方面的优势较小。

2.1.4　U型网络

U型网络使用编码器—解码器的架构，其结构如图8所示。通过不同尺度特征的梯次上采样融合，能够获得丰富的空间信息与不同尺度的上下文信息。实时分割中的U型结构网络，主要在U型网络基础结构上进行网络结构与特征融合方法的改进。

SwiftNet（Oršic等，2019）率先在实时分割领域应用U型结构。相比基础U型结构，SwiftNet添加了另一个输入图像为原始分辨率1/2的编码器，两编码器共享参数，二者的同尺度特征通过拼接方式进行融合，从而可以从不同分辨率输入图像中获得不同尺度的语义信息。另外，SwiftNet还提出利用图像分类任务上的轻量级网络结构及预训练部分作为编码器、解码器（简单的1×1分类器），增强了模型编码语义信息的能力。

ShelfNet（Zhuang等，2019）提出多分支编码—解码分支，将两个U型基础结构堆叠，并在每个分辨率相同的特征间添加跳层连接，形成了类似“多行置物架”的整体结构。这种U型堆叠结构可视为多个深层与浅层分支的集成，因此可以提升预测精度。虽然为了减少计算量，ShelfNet减少了通道数，但仍能凭借此特殊网络结构设计取得较高精度。另外，ShelfNet在残差块中应用的参数共享策略，提出了共享权重残差（shraed-weight residual block，S-Block），进行不同尺度特征间的融合，可在不牺牲精度表现的同时减少卷积层参数的数量。

SFNet（Li等，2020）在U型基础结构的基础上在解码器部分又额外添加了一条梯次上采样路径，进行语义特征间的融合。这使得SFNet能够在各尺度整合进语义信息后再次进行语义信息交互。另外，为了更好地融合与传递不同尺度间的语义信息，受到光流网络的启发，SFNet提出语义流，将同图像不同分辨率的特征之间的关系也用像素间的“运动变化”形式表示。获得语义流后，网络可以以较低的信息损失在不同尺度间传递精确的语义信息。基于语义流提出的流对齐模块（flow alignment module，FAM）可以有效地将高层语义信息传递至浅层高分辨率特征，使其同时具有高层语义信息与精细的空间结构信息，在提升预测精度的同时保持高效计算。

Nirkin等人（2021）认为，由于分割物体可能存在于图像中的所有位置，语义分割网络应该具有较强的局部适应性。因此，Nirkin等人（2021）提出的HyperSeg（hypernetwork for real-time semantic segmentation）模型使用的U型编码器结构，不仅编码图像信息，也基于编码之后的特征生成动态的网络权重。HyperSeg在进行不同尺度特征融合时，对该权重会根据不同空间位置进行动态变化。具体而言，该模块将patch级别的动态权重与输入特征进行点乘并参与特征融合，以此提升网络对于不同局部位置中独有特征的适应性，带来可观的分割精度提升。

TopFormer（Zhang 等，2022）探索了 Transformer 在移动设备上的语义分割任务中的应用。TopFormer 提出了一种 U 型结构卷积编码器，以生成不同尺度的 tokens。TopFormer 提出了尺度感知语义提取器，将 U 型结构不同尺度的 tokens 下采样到输入图像分辨率的 1/64 后按通道拼接，接着使用 Transformer 进行语义特征提取。通过自注意力的不同尺度的 token 交互了不同空间、不同尺度的信息，具有全局语义。在特征融合方面，TopFormer 按照通道将生成的 tokens 拆分，通过语义注入模块分别将不同尺度 tokens 的信息与对应尺度的特征进行交互，以增强不同尺度特征的语义信息。之后，每个增强过的特征图再进行梯次上采样融合后输入分割头。因为 TopFormer 只在输入图像 1/64 的分辨率下使用 Transformer，而且 Transformer 中进行了一些加速操作，如降低 query、key 的维度和降低前馈网络的膨胀率等；因此 TopFormer 并没有因为引入 Transformer 而带来大量的推理延迟。TopFormer 最终得到了一个在移动设备上能够进行高效推理的语义分割架构，具有较好的速度—精度平衡。

U 型网络能够提取和融合多个不同尺度的特征信息，解码器能力强大，能有效提高分割精度；但多尺度特征的梯次融合和解码会带来更大的解码计算代价和推理延迟。因此 U 型实时语义分割网络需要设计更高效的解码融合方式，或者尽可能地去除冗余的特征。

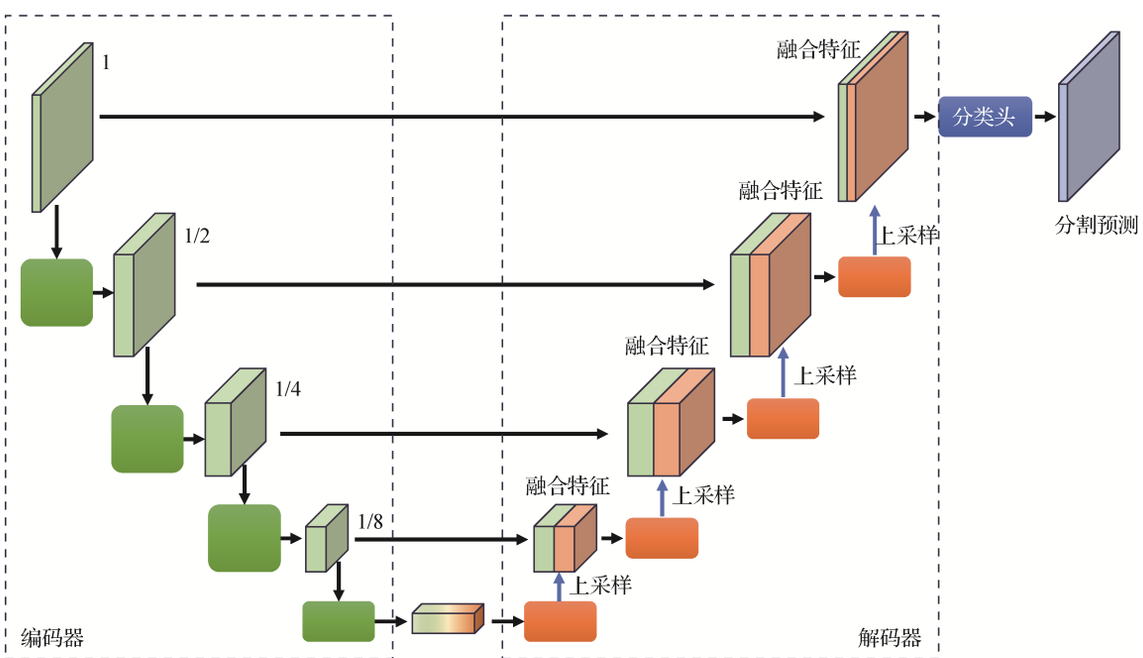


图 8　U 型实时语义分割网络结构示意图

Fig. 8　Schematic diagram of the U-shape real-time semantic segmentation network

2.1.5　神经架构搜索网络

神经架构搜索技术广泛应用于神经网络设计中，用以得到搜索空间中更加适合特定任务的网络结构。当加入推理延迟为优化指标时，神经架构搜索技术也能用于设计实时语义分割网络。不同于前述 4 种人工设计的网络结构在对应类型的实时分割结构设计方面进行改进和贡献，网络搜索相关工作的贡献主要集中在搜索空间和搜索策略的设计改进上。而且，搜索得到的实时分割网络的结构主要取决于搜索空间的结构，即超网使用的分割网络结构。由于超网结构的不同，最终所得的具体实时分割网络可能采用与前述 4 种不同结构的子类。

因此，在介绍了上述 4 类结构后，本文将实时语义分割领域中的神经架构搜索网络单独列为一个类别以便读者理解。目前使用网络搜索技术进行实时语义分割网络设计的方法主要包括 DF-Seg（dongfeng segmentation）、FasterSeg（faster real-time segmentation）、RT-Seg（real-time segmentation）和 PPB-Seg（pruning parameterization with bi-level segmentation）。这 4 种网络使用的超网结构和进行的方法改进可见表 3。

表3　实时分割中的神经架构搜索网络归纳

Table 3　Overview of neural architecture search networks in real-time segmentation

网络	超网结构	方法改进
DF-Seg	U型	首次引入推理延迟搜索指标、编序剪枝缩小搜索空间
Faster Seg	ICNet(多分支)、BiseNet(双分支)	为不同搜索敏感度的归一化系数、引入超网蒸馏
RT-Seg	BiSeNet(双分支)	缩小搜索空间、引入ViT模块、延迟感知正则化
PPB-Seg	TopFormer(U型)	引入soft mask剪枝、Bi-level优化

早期研究多用每秒浮点运算次数(floating-point operations per second, FLOPs)评价实时模型的轻量程度,但FLOPs只反映浮点运算量的大小,不包含内存读取,并不真实反映实际推理速度的快慢。对于不同平台,即便不同模型具有同样的FLOPs,也可能因为不同硬件性能的区别在推理速度上具有较大差异。为了在实际设备上获得更好的推理速度与精度,Li等人(2019b)提出了在实际设备上估计网络结构的推理延迟以评价实时性,并提出偏序剪枝算法缩小搜索空间。搜索空间内不同网络结构可以基于通道数、网络层数等构建偏序关系。在网络层数较小时,延迟与精度大致呈正相关。将已训练的准确率最高同时速度最快的模型的延时作为延时下界,该算法根据偏序关系丢弃精度与延时都较差的模型集合。使用该方法搜索出的DF-Seg模型具有低推理延迟,与其他方法相比在相同延时下具有更高的推理精度。

尽管DF-Seg的搜索方法取得了成功,但在设计搜索空间时并没有将人工设计分割网络的成功经验吸收进来,例如,人工设计的结构往往使用对语义分割极为重要的多个不同尺度分支结构。另外,因为对不同基础操作层、下采样倍数和通道扩张比例的敏感程度不同,NAS方法往往容易得到推理延迟低但表现不佳的搜索结果。为了解决以上问题,FasterSeg(Chen等,2020)设计了多尺度的搜索空间,其中包含ICNet与BiseNet的结构;同时通过为敏感度不同的搜索网络基础操作层、下采样倍数和通道扩张比例添加不同的归一化系数,避免了搜索早期就落入虽然推理延时很低但精度表现不佳的搜索空间中。另外,FasterSeg创新性地首次提出可以在一轮搜索中同时搜索复杂大模型作为教师网络、低推理延迟的轻量网络作为学生网络,之后进行知识蒸馏进一步提升轻量化网络的性能表现。

RT-Seg(Li等,2023)旨在利用NAS技术得到更好的双分支网络。具体来说,RT-Seg将搜索空间局限在BiSeNet型的双分支网络的模块的搜索和网络宽度的搜索,这与密集连接网格上的搜索相比(Chen等,2020;Zhang等,2021;Liu等,2019),节省了相当大的搜索成本。RT-Seg网络宽度搜索是通过通道剪枝实现的,而模块的搜索集中在混合模块的选择上。具体来说,对于搜索的超网,RT-Seg构建了一种混合模块,如图9所示。混合模块结合了ViT(vision Transformer)和CNN的优势,具有全局感受野和高质量的局部特征,克服了纯Transformer的密集计算,子网通过基于梯度的gamble softmax采样进行块搜索。RTFormer的超网结构是将上下文分支的模块替换为混合块的BiSeNet。RT-Seg没有使用代理搜索,还设计了一种新的延迟感知正则化,直接评估候选的推理速度/延迟。另外,RT-Seg还利用自蒸馏和辅助损失,在没有外部知识的情况下充分挖掘子网的精度潜力。

Yang等人(2023)提出了一种通过soft mask进行剪枝参数化的方法,使用mask阈值化方法在训练和推理过程中实现剪枝,同时利用直通估计器(straight-through estimator, STE)来传播阈值化中无法计算的梯度。此外,通过Bi-level的优化方法,证明了隐式梯度中的二阶导数可以通过一阶导数有效地得到,节省了计算量和内存。最后,Yang等人(2023)分析了TopFormer结构中延迟占比,对TopFormer中卷积层加上了其提出的剪枝方法。通过该剪枝方法得到的TopFormer优化网络,相比TopFormer原网络,在速度—精度平衡上得到了全面的提升,在移动端GPU上达到了最好的速度—精度平衡。需要说明的是,由于Yang等人(2023)未给其提出的实时分割方法起名,为了表述方便,本文皆以

PPB-Seg(pruning parameterization with bi-level optimization for efficient semantic segmentation on the edge)来指代该方法。

相比人工设计的实时语义分割网络，神经架构搜索网络能够得到冗余度更小、效率更高的实时语义分割网络，在精度和速度上都能带来增强。但另一方面，神经架构搜索网络会使训练过程更加复杂，增大训练成本。

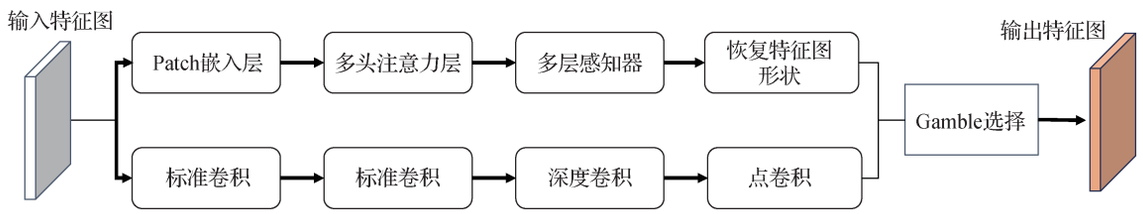


图 9　混合模块的结构图

Fig. 9　An overview of the mix block

2.2　基础框架类型

基于深度学习的实时语义分割方法按照基础框架类型可以分为 CNN 框架、Transformer 框架和基于 CNN-Transformer 混合框架的 3 种方法。这 3 种框架实时语义分割方法包括的具体算法见图 10。其中，由于 CNN 框架具有关于图像分辨率的线性复杂度以及经过长期发展具备的良好硬件优化，故目前实时语义分割领域绝大部分方法仍为 CNN 框架。基于 CNN 框架的实时语义分割方法的结构多样，涵盖单分支、双分支、多分支、U 型、神经架构搜索网络。

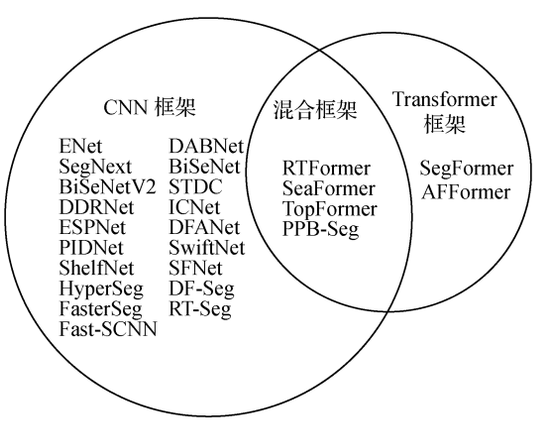


图 10　不同基础框架的实时语义分割方法归纳

Fig. 10　Overview of different frameworks of real-time semantic segmentation methods

标准 Transformer 模块关于图像分辨率的平方复杂度带来了巨大的推理延迟，这在实时语义分割任务中是难以接受的。因此，目前基于 Transformer 框架的实时语义分割方法较少，仅有 SegFormer 和 AFFormer。而且，为了减小推理延迟，SegFormer 和 AFFormer 都使用了轻量 Transformer 模块，并且只保留了单分支网络结构。

基于 CNN-Transformer 混合框架的实时语义分割方法的核心思想是利用 CNN 框架和 Transformer 框架各自的优点，实现更好的实时语义分割网络。基于 CNN-Transformer 混合框架的实时语义分割方法通常只在网络深层使用 Transformer 提取上下文信息，如 Topformer 和 PPB-Seg；或者使用双分支网络结构，将 Transformer 当做上下文分支、将 CNN 当做空间分支，如 RTFormer 和 SeaFormer。通过上述设计，基于 CNN-Transformer 混合框架的算法能够同时利用 Transformer 的长距离上下文建模能力和 CNN 的局部建模能力，同时避免了 Transformer 在高分辨率图像上的操作带来的高推理延迟。

2.3　应用场景类型

基于深度学习的实时语义分割方法按照应用场景类型可以分为服务器端的实时语义分割方法和移动端的实时语义分割方法。

服务器端的实时语义分割方法的部署设备通常为桌面 GPU，例如 GTX 1080 Ti、RTX2080 Ti、RTX 3090 等设备。桌面 GPU 具有较大的显存和算力，且对高并行度的计算密集型操作优化更好。因此，针对桌面 GPU 端的实时语义分割方法通常使用并行度更高的操作。RTFormer(Wang 等，2022)将 Transformer 的注意力分头操作限制在归一化中，使得注意力操作的并行程度更高。

移动端的实时语义分割算法分为移动端 GPU 实时语义分割算法和移动端 CPU 实时语义分割算法。其中移动端 GPU 与桌面 GPU 架构类似，但显存更小，因此还需要考虑模型的最大显存占用限制。而移动端 CPU 由于算力和内存更小，且对并行操作

的支持较差，因此，设计针对移动端CPU的实时语义分割方法时，需要追求更小的参数量和浮点运算数。

表4展示了本文提及的实时语义分割方法所属的应用场景类型，目前绝大部分的实时语义分割方法的应用场景为桌面GPU和移动端GPU，仅有TopFormer（Zhang等，2022）和SeaFormer（Wan等，2023）二者聚焦在移动端CPU分割。

表4　实时语义分割方法归纳

Table 4　Overview of real-time semantic segmentation methods

类型	方法	基础框架	应用场景	发表期刊或会议
单分支网络	ENet（Paszke等，2016）	CNN	Desk/Mobile GPU	Arxiv
	DABNet（Li等，2019c）	CNN	Desk GPU	BMVC
	SegFormer（Xie等，2021）	Transformer	Desk GPU	NeurIPS
	SegNext（Guo等，2022）	CNN	Desk GPU	NeurIPS
	AFFormer（Dong等，2023）	Transformer	Desk GPU	AAAI
双分支网络	BiSeNet（Yu等，2018）	CNN	Desk GPU	ECCV
	Fast-SCNN（Poudel等，2019）	CNN	Desk GPU	BMVC
	BiSeNetV2（Yu等，2021）	CNN	Desk GPU	IJCV
	STDC（Fan等，2021）	CNN	Desk GPU	CVPR
	DDRNet（Pan等，2023）	CNN	Desk GPU	TIP
	RTFormer（Wang等，2022）	Hybrid	Desk GPU	NeurIPS
	SeaFormer（Wan等，2023）	Hybrid	Mobile CPU	ICLR
多分支网络	ICNet（Zhao等，2018）	CNN	Desk GPU	ECCV
	ESPNet（Mehta等，2018）	CNN	Desk/Mobile GPU	ECCV
	DFANet（Li等，2019a）	CNN	Desk GPU	CVPR
	PIDNet（Xu等，2023）	CNN	Desk GPU	CVPR
U型网络	SwiftNet（Oršic等，2019）	CNN	Desk/Mobile GPU	CVPR
	ShelfNet（Zhuang等，2019）	CNN	Desk GPU	ICCV（workshop）
	SFNet（Li等，2020）	CNN	Desk GPU	ECCV
	HyperSeg（Nirkin等，2021）	CNN	Desk GPU	CVPR
	TopFormer（Zhang等，2022）	Hybrid	Mobile CPU	CVPR
神经架构搜索网络	DF-Seg（Li等，2019b）	CNN	Desk/Mobile GPU	CVPR
	FasterSeg（Chen等，2020）	CNN	Desk GPU	ICLR
	RT-Seg（Li等，2023）	CNN	Mobile GPU	AAAI
	PPB-Seg（Yang等，2023）	Hybrid	Mobile GPU	CVPR

3　评价体系

3.1　数据集

语义分割任务需要依赖于大量的标注数据，实时语义分割任务常用的数据集如表5所示。

3.1.1　Cityscapes

Cityscapes（Cordts等，2016）数据集是一个大规模的城市街景语义理解数据集（https://www.cityscapes-dataset.com/）。Cityscapes有5 000幅精细注释图像（其中，2 975幅用于训练，500幅用于验证，1 525幅用于测试）和20 000幅粗糙注释图像，所

表 5　实时语义分割的数据集总结

Table 5　Summary of the datasets in real-time semantic segmentation

数据集	年份	类别	规模	分辨率/像素	训练集	验证集	测试集	类型
Cityscapes	2012	19	5 000	2 048 × 1 024	2 975	500	1 525	自动驾驶
CamVid	2009	11	701	960 × 720	367	101	233	自动驾驶
ADE20K	2017	150	22 210	不固定	20 210	2 000	–	通用场景
COCOStuff-10K	2018	171	10 000	不固定	9 000	–	1 000	通用场景
PASCAL VOC 2012	2012	20	13 487	不固定	10 582	1 449	1 456	通用场景
PascalContext	2015	59	10 103	不固定	4 998	–	5 105	通用场景

注：“–”表示无此项数据。

有图像分辨率均为2 048 × 1 024像素，涵盖了30个不同的类别，其中19类用于语义分割。实时语义分割方法通常只使用其中的精细注释图像，只有少量方法使用粗糙数据对网络性能进行进一步增强。Cityscapes涵盖了各种复杂的城市街景场景，具有图像分辨率高、标注质量高和时空跨度范围广等特点，因此已成为实时分割领域最常用的数据集。

3.1.2　CamVid

CamVid（Brostow等，2009）是第1个稠密标注的自动驾驶数据集（http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/）。CamVid数据集包含701幅车辆行驶视角的图像，来自英国剑桥的10 min驾驶拍摄序列；其中包括367幅训练图像、101幅验证图像、233幅测试图像，分辨率均为960 × 720像素。CamVid包含32个类别，其中11个类别用于语义分割。CamVid数据集时空多样性较弱且规模较小，用于CamVid的网络往往需要进行Cityscapes预训练。

3.1.3　ADE20K

ADE20K（Zhou等，2017）数据集（https://groups.csail.mit.edu/vision/datasets/ADE20K/）是目前最全面的用于场景解析和语义分割数据集之一。ADE20K数据集一共包含22 210幅图像（其中，20 210幅用于训练，2 000幅用于验证），其图像尺寸大小不一，涵盖了各种场景，包括室内、室外、自然和城市等。这些图像都有像素级别的注释，涵盖了150个类别的物体。这些类别包括常见的物体类别，如人、车、树，以及一些罕见的类别，如飞机、船等。ADE20K具有丰富的语义类别和长尾分布的问题，对于轻量的实时语义分割网络是一项挑战。

3.1.4　COCOStuff-10K

COCOStuff（Caesar等，2018）数据集（https://github.com/eulersantana/cocostuff）通过在大规模的场景理解数据集COCO（common objects in context）的基础上添加大量的“stuff”类别扩展得到。COCO-Stuff-10K（Caesar等，2018）数据集是COCOStuff中带有人工稠密标注的10 000幅复杂图像，其中9 000幅用于训练，1 000幅用于测试。COCOStuff-10K包含关于182个类别，包括91个thing类和91个stuff类，但其中11个thing类没有标注，因此实际使用的只有171个类别。对于实时语义分割来说，这也是一个具有挑战性的数据集，因为它有更复杂的类别和更多变的场景。

3.1.5　PASCAL VOC 2012

PASCAL VOC 2012（pattern analysis，statistical modeling and computational learning visual object classes）（Everingham等，2015）数据集（http://host.robots.ox.ac.uk/pascal/VOC/voc2012/）包含了20个不同的前景类别和1个背景类别，有10 582幅训练图像、1 449幅验证图像、1 456幅测试图像。

3.1.6　PascalContext

PascalContext（Mottaghi等，2014）数据集（https://www.cs.stanford.edu/~roozbeh/pascal-context/）包含了4 998幅训练图像和5 104幅测试图像，覆盖了与Pascal VOC相同的20个对象类别，但每个像素都被分配到更详细的语义类别，共有59个前景类别和1个背景类别。

3.2　评价指标

本节介绍实时语义分割中常用的4个指标：平均交并比、每秒帧数、推理延迟和参数量。如图11所示。另外，虽然部分实时分割网络汇报了浮点运

算数(floating point operations, FLOPs),但由于浮点运算数不能直接估计推理速度,且不能直接反映内存占用,仅仅是二者的一种参考,因此本文没有选择FLOPs作为评价指标。

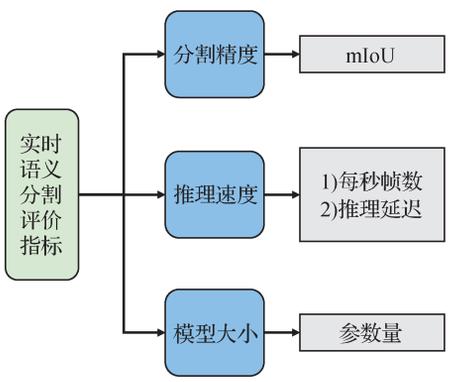


图11　实时分割评价指标总结

Fig. 11　Summary of the metrics of real-time segmentation

3.2.1　平均交并比

平均交并比(mean IoU, mIoU)是一种广泛应用于语义分割任务的评价指标。IoU通过计算两组像素的交集与并集的比率来量化分割掩码预测值与真实值之间的重叠程度。mIoU是数据集中所有类的IoU的平均值,是分割模型整体准确性的度量。

3.2.2　每秒帧数

每秒帧数(frames per second, FPS)即帧速率,是网络模型每秒可以处理的图像帧的数量,大小受具体设备性能的影响。实时分割要求网络的FPS大于等于30帧/s。在同设备下比较时,FPS可以直观地衡量网络的推理速度和吞吐量。

3.2.3　推理延迟

延迟(latency)是FPS的倒数,表示网络模型处理一帧图像所需要的时间。通常在移动端语义分割模型中使用较多。GPU设备的算力较强,且使用Tensor-RT加速后能获得更快的推理速度,许多实时分割网络在GPU上的FPS远超30帧/s,此时对比延迟不直观。而移动端CPU算力资源有限,延迟大而FPS小,使用延迟能更直观地比较网络的推理速度。

3.2.4　参数量

参数量(parameters)是指在模型训练过程中需要训练学习的参数总数。参数量通常用于度量模型的大小。在资源受限的边缘设备上,参数量是一个需要考虑的关键因素。

3.3　方法性能汇总

本节所有的精度和速度均来自原论文,注明了测速设备和加速设置。

3.3.1　Cityscapes性能汇总

Cityscapes数据集上性能的汇总如表6所示,mIoU(val)和mIoU(test)分别表示验证集和测试集的mIoU。其中,加*的帧速率表明其使用了加速技术(通常为Tensor-RT加速),加*的mIoU表明该论文只汇报了多尺度的精度。

由表6结果可知,对于Cityscapes验证集和测试集,在精度指标上表现最好的是多分支网络PIDNet,其PIDNet-L版本在RTX 3090上以实时运行的最低水平(30帧/s)在Cityscapes验证集上达到了80.9% mIoU,在Cityscape测试集上达到了80.6% mIoU。值得注意的是,目前Cityscapes上高精度分割网络(无延迟、计算量、参数量约束)的最高分割精度为InternImage-H (Wang等, 2023),其在Cityscapes验证集上达到了87.0% mIoU,在Cityscape测试集上达到了86.1% mIoU。另一方面,从速度角度来看,未加速的网络中推理速度最快的是Fast-SCNN,Tensor-RT加速后最快的网络为FasterSeg。由于速度受到设备和加速设置的影响,推荐参考3.4节主流方法同设备比较部分以进行推理速度上的直观比较。

3.3.2　CamVid性能汇总

CamVid数据集上性能的汇总如表7所示,其中RTFormer-B达到了82.5% mIoU,RTFormer-S达到了81.4% mIoU,展现出远超其他方法的精度;值得一提的是,大多数方法在CamVid数据集上训练时使用了Cityscapes数据集上的预训练权重,而RTFormer没有使用Cityscapes数据集上的预训练权重,这更加凸显了RTFormer在CamVid数据集上的高效性。另一方面,FastSeg则达到了最快的推理速度,帧速率FPS高达398.1帧/s,推理速度第2快的方法STDC1-Seg的帧速率的两倍多。

3.3.3　ADE20K和COCOStuff-10K性能汇总

许多方法只汇报ADE20K和COCOStuff-10K上的FLOPs,而不汇报帧速率FPS。由于FLOPs无法直观地反映推理速度的大小,这不利于直观地比较方法的推理效率。为此,本文在RTX 3090 GPU上测量了相关方法在ADE20K和COCOStuff-10K上的FPS,对空缺的FPS进行了补全。

表 6　Cityscapes 数据集上性能比较

Table 6　Comparison of the performance on Cityscapes dataset

类型	方法	参数量/M	分辨率/像素	GPU	FPS/(帧/s)	mIoU (val)/%	mIoU (test) /%
单分支网络	ENet	0.4	1 024 × 512	TitanX	76.9	–	58.3
	DABNet	0.76	2 048 × 1 024	GTX 1080Ti	27.7	–	70.1
	SegFormer-B0	3.8	1 536 × 768	V100	26.3	75.3	–
	SegNext-T	4.3	1 536 × 768	RTX 3090	25.0	78.0	–
	AFFormer-B	3.0	2 048 × 1 024	V100	22	78.7	–
双分支网络	BiSeNet-ResNetR18	49.0	1 536 × 768	GTX 1080Ti	65.5	74.8	74.7
	Fast-SCNN	1.1	2 048 × 1 024	TitanXp	123.5	68.6	68.0
	BiSeNetV2-L	–	1 024 × 512	GTX 1080Ti	47.3*	75.8	75.3
	STDC1-Seg75	14.2	1 536 × 768	GTX 1080Ti	<u>126.7*</u>	74.5	75.3
	STDC2-Seg75	22.2	1 536 × 768	GTX 1080Ti	97.0*	77.0	76.8
	DDRNet-23-S	5.7	2 048 × 1 024	RTX 2080Ti	101.6	77.8	77.4
	DDRNet-23	20.1	2 048 × 1 024	RTX 2080Ti	37.1	79.5	79.4
	RTFormer-S	4.8	2 048 × 1 024	RTX 2080Ti	110.0*	76.3	–
	RTFormer-B	16.8	2 048 × 1 024	RTX 2080Ti	39.1*	79.3	–
多分支网络	ICNet	26.5	2 048 × 1 024	TitanX	30.3	–	69.5
	ESPNet	0.4	1 024 × 512	TitanX	113	–	60.3
	DFANet A	7.8	1 024 × 1 024	TitanX	100.0	–	71.3
	PIDNet-S	7.6	2 048 × 1 024	RTX 3090	93.2	78.8	78.6
	PIDNet-M	34.4	2 048 × 1 024	RTX 3090	39.8	<u>80.1</u>	<u>80.1</u>
	PIDNet-L	36.9	2 048 × 1 024	RTX 3090	31.1	**80.9**	**80.6**
U 型网络	SwiftNetRN-18	11.8	2 048 × 1 024	GTX 1080Ti	39.9	75.5	75.4
	ShelfNet18-lw	–	2 048 × 1 024	GTX 1080Ti	36.9	–	74.8*
	SFNet-ResNet18	12.9	2 048 × 1 024	GTX 1080Ti	18	–	78.9
	HyperSeg-M	10.1	1 024 × 512	GTX 1080Ti	36.9	76.2	75.8
	HyperSeg-S	10.2	1 536 × 768	GTX 1080Ti	16.1	78.2	78.1
神经架构搜索网络	DF2-Seg1	–	1 536 × 768	GTX 1080Ti	67.2*	75.9	74.8
	DF2-Seg2	–	1 536 × 768	GTX 1080Ti	56.3*	76.9	75.3
	FasterSeg	4.4	2 048 × 1 024	GTX 1080Ti	**163.9***	73.1	71.5

注：加粗、下划线字体分别表示各列最优、次优结果，“–”表示未提供相关数据。

如表 8 和表 9 所示，ADE20K 上和 COCOSTuff-10K 数据集上，RTFormer-B 显示出超越其他实时分割网络的速度—精度平衡，在维持更快分割速度的同时达到了更高的分割精度。

3.4　主流方法同设备比较

由于实时语义分割网络的推理速度和设备以及加速手段息息相关，为了更直观地对比现有的实时语义分割方法的分割精度和推理速度性能，本文对其中一些具有代表性的主流方法进行了统一设备比较。网络实现均采用作者源码或 MMSegmentation (Contributors，2020)库官方代码，测速代码和加速设置统一。数据集选取实时语义分割领域使用最广泛

表7 CamVid数据集上性能比较

Table 7 Comparison of performance on CamVid dataset

方法	mIoU/%	FPS/(帧/s)	GPU
ENet	68.3	61.2	TitanX
DABNet	66.4	146	GTX 1080Ti
BiSeNet-ResNetR18	68.7	116	GTX 1080Ti
BiSeNetV2-L	78.6	33*	GTX 1080Ti
STDC1-Seg	73.0	197.6*	GTX 1080Ti
STDC2-Seg	73.9	152.2	GTX 1080Ti
DDRNet-23-S	78.6	230	RTX 2080Ti
DDRNet-23	80.6	94	RTX 2080Ti
RTFormer-S	81.4	190.7	RTX 2080Ti
RTFormer-B	**82.5**	94.0	RTX 2080Ti
ICNet	67.1	34.5	TitanX
DFANet A	64.7	120	TitanX
PIDNet-S	80.1	153.7	RTX 3090
PIDNet-S-Wider	82.0	85.6	RTX 3090
SwiftNetRN-18	72.6	–	GTX 1080Ti
SFNet-ResNet18	73.8	36	GTX 1080Ti
HyperSeg-S	78.4	38	GTX 1080Ti
HyperSeg-L	79.1	16.6	GTX 1080Ti
FasterSeg	71.1	**398.1***	GTX 1080Ti

注：加粗、下划线字体分别表示各列最优、次优结果，“–”表示未提供相关数据。

表8 ADE20K数据集上性能比较

Table 8 Comparison of performance on ADE20K dataset

方法	mIoU/%	FPS/(帧/s)	GPU
SegFormer-B0	37.4	84.4	RTX 3090
SegNext-T	41.1	60.3	RTX 3090
AFFormer-B	41.8	49.6	RTX 3090
RTFormer-B	**42.1**	93.4	RTX 3090
SeaFormer-B	41.0	44.5	RTX 3090
TopFormer-B	39.2	**96.2**	RTX 3090

注：加粗、下划线字体分别表示各列最优、次优结果。

的Cityscapes。

3.4.1 实验设置

本文在2种设备上对目前的SOTA（state-of-the-art）方法进行比较：RTX 3090 GPU、Snapdragon 865 CPU。这两种设备分别代表了实时语义分割中常见的桌面GPU（服务器端）和边缘CPU（移动端）的两种应用。

表9 COCOStuff-10K数据集上性能比较

Table 9 Performance comparison on COCOStuff-10K dataset

方法	mIoU/%	FPS/(帧/s)	GPU
ICNet	29.1	35.7	Titan X
BiSeNetV2-L	28.7	65.1	RTX 3090
AFFormer-B	35.1	46.5	RTX 3090
DDRNet-23	32.1	**108.8**	RTX 3090
RTFormer-B	**35.3**	90.9	RTX 3090
SeaFormer-B	34.1	41.9	RTX 3090
TopFormer-B	33.4	94.7	RTX 3090

注：加粗、下划线字体分别表示各列最优、次优结果。

对于GPU测速，本文分别测量Torch速度和Tensor-RT。Torch速度使用代码本身的数位精度进行测量，测速代码可见https://github.com/xzz777/Awesome-Real-time-Semantic-Segmentation，测速过程只记录模型推理时间，不记录数据读取和预处理时间。对于Tensor-RT加速，本文采用MMDeploy（Contributors，2021）实现。采用Tensor-RT 8.2.1进行加速，推理时使用FP16精度进行测量。

对于CPU测速，使用Mnn（Jiang等，2020）框架，将Torch模型转换为Mnn模型，在Snapdragon 865 CPU测量推理延迟，线程设置为1，推理时使用FP32精度进行测量。

3.4.2 桌面GPU同设备比较

具有代表性的一些先进实时语义分割方法在RTX 3090 GPU上的详细性能比较见表10。在2 048×1 024像素、1 536×768像素和1 024×512像素3种分辨率下验证集mIoU指标最高的网络分别为PIDNet-L、SegNext-T-75和BiSeNet-L。在不使用Tensor-RT加速时，在2 048×511 024像素、1 536×51 768像素和1 024×51 512像素分辨率下速度最快的网络分别为DDRNet-23-S、BiSeNet-ResNetR18和STDC1-Seg50。使用Tensor-RT加速时，在2 048×1 024像素、1 536×768像素和1 024×512像素分辨率下速度最快的网络分别为TopFomrer-B-100、BiSeNet-ResNetR18和TopFomrer-B-50。

表 10　同设备比较(RTX 3090 GPU)

Table 10　Comparison on the same device(RTX 3090 GPU)

类型	方法	参数量/M	分辨率/像素	FPS (Tensor-RT) /(帧/s)	FPS (Torch) /(帧/s)	mIoU (val) /%	mIoU (test) /%
单分支网络	SegFormer-B0	3.8	1 536 × 768	60.3	39.6	75.3	–
	SegNext-T-75	4.3	1 536 × 768	78.3	47.3	78.0	–
	SegNext-T-100	4.3	2 048 × 1 024	46.5	28.1	79.8	–
	AFFormer-B-50	3.0	1 024 × 512	148.4	49.5	73.5	–
	AFFormer-B-75	3.0	1 536 × 768	96.4	38.6	76.5	–
	AFFormer-B-100	3.0	2 048 × 1 024	58.3	28.4	78.7	–
双分支网络	BiSeNet-ResNetR18	49.0	1 536 × 768	182.9	112.3	74.8	74.7
	BiSeNetV2-L	–	1 024 × 512	102.3	67.6	75.8	75.3
	STDC1-Seg75	14.2	1 536 × 768	209.5	101.9	74.5	75.3
	STDC2-Seg75	22.2	1 536 × 768	139.2	84.3	77.0	76.8
	STDC1-Seg50	14.2	1 024 × 512	397.6	146.2	72.2	71.9
	STDC2-Seg50	22.2	1 024 × 512	279.7	94.6	74.2	73.4
	DDRNet-23-S	5.7	2 048 × 1 024	138.9	106.7	77.8	77.4
	DDRNet-23	20.1	2 048 × 1 024	101.9	56.7	79.5	79.4
	RTFormer-S	4.8	2 048 × 1 024	–	89.6	76.3	–
	RTFormer-B	16.8	2 048 × 1 024	–	50.2	79.3	–
	SeaFormer-B-50	8.7	1 024 × 512	231.6	44.1	72.2	–
	SeaFormer-B-100	8.7	2 048 × 1 024	103.6	37.5	77.7	–
多分支网络	PIDNet-S	7.6	2 048 × 1 024	127.1	74.2	78.8	78.6
	PIDNet-M	34.4	2 048 × 1 024	90.7	41.0	80.1	80.1
	PIDNet-L	36.9	2 048 × 1 024	76.9	31.2	80.9	80.6
U 型网络	TopFormer-B-50	8.7	1 024 × 512	410.9	81.4	70.7	–
	TopFormer-B-100	8.7	2 048 × 1 024	128.4	95.7	76.3	–

注：“–”表示未提供相关数据。

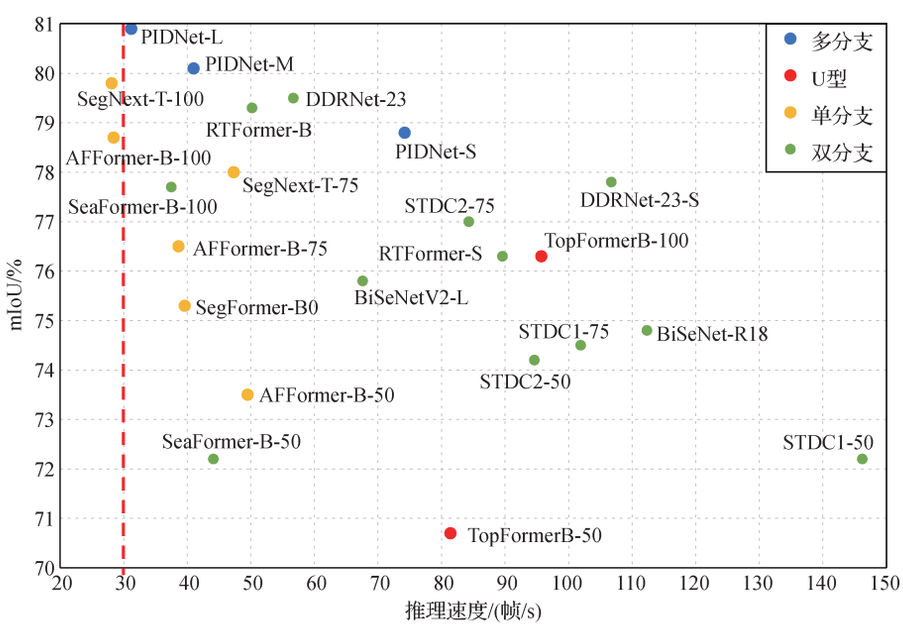


图 12　RTX 3090 上的速度—精度平衡图(未加速)

Fig. 12　Speed and accuracy trade-off on RTX 3090(without acceleration)

为了进行更直观地比较，图12展示了这些方法在Torch上的精度—速度平衡图，其中纵轴选择验证集mIoU，以覆盖更多方法。可以看出，从实时标准线30～100帧/s以上，多分支网络PIDNet和双分支网络DDRNet共同维持着目前实时语义分割在Cityscapes验证集上最好的速度—精度平衡。而且，双分支/多分支网络普遍有着比单分支网络更快的推理速度。这些现象充分说明了分支设计在实时语义分割中发挥的巨大作用。

图13展示了经过Tensor-RT加速后的精度—速度平衡图。大部分方法经过Tensor-RT加速后和加速前的相对关系保持一致。经过加速后，在140帧/s以下的区间，多分支网络PIDNet和双分支网络DDRNet共同维持着最好的速度—精度平衡；而在140帧/s以上的高速区间，STDCNet表现更加优秀。值得注意的是，DDRNet在Tensor-RT加速前后的加速比较小，而TopFormer的加速比相当大。不同网络的加速比不同可能是由于网络中各模块采用的算子的优化程度不同，当网络中的算子更倾向于访存密集型而非计算密集型时，加速比可能相对较小。

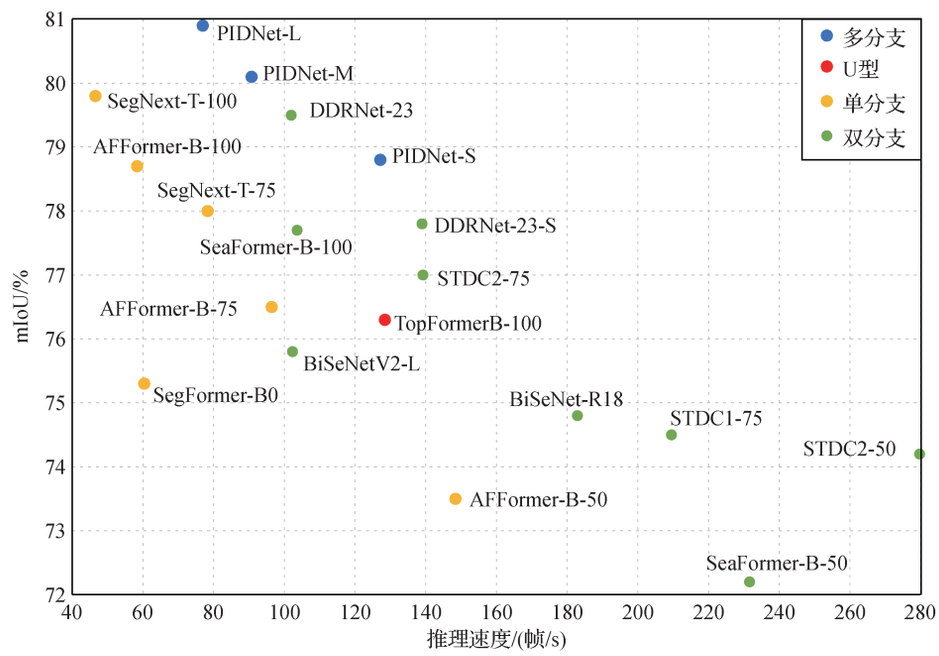


图13　RTX 3090上的速度—精度平衡图（Tensor-RT加速）

Fig. 13　Speed and accuracy trade-off on RTX 3090 (with Tensor-RT acceleration)

3.4.3　边缘CPU同设备比较

一些具有代表性实时语义分割算法在Snapdragon 865上的详细性能比较见表11。

由于计算资源受限，CPU上的高效算子和GPU上的高效算子不相同，一些GPU上最先进的实时语义分割算法在Snapdragon 865上有相当大的延迟。PIDNet-M在Torch上的速度略快于SeaFormer-B-100，但在Snapdragon 865上速度为SeaFormer-B-100的1/5。由于针对桌面GPU设计，RTFormer在Snapdragon 865上的速度也非常慢。值得注意的是，虽然STDC和DDRNet针对桌面GPU设计，但在Snapdragon 865上也表现出相对优秀的速度—精度平衡。总体来看，SeaFormer维持着目前边缘CPU上最好的速度—精度平衡。

由于边缘设备Snapdragon 865 CPU的算力远小于桌面GPU，目前大部分实时语义分割的算法在Snapdragon 865的运行速度都较慢。其中最快的TopFormer-B-50的推理延迟仍有118 ms，换算成每秒帧数仍小于10，远远不到30帧/s的实时标准。边缘设备上的实时语义分割算法仍有较大发展空间。

4　结　语

语义分割作为一项重要的图像感知与理解任务，在医学图像处理、场景分析理解、自动驾驶感知以及智能视频分析等领域中得到广泛应用。而在实

表 11　同设备比较（Snapdragon 865）

Table 11　The comparison on the same device (Snapdragon 865)

方法	参数量/M	分辨率/像素	延迟/ms	mIoU/%
STDC1-Seg50	14.2	1 024 × 512	259	72.2
STDC2-Seg50	22.2	1 024 × 512	340	74.2
TopFormer-B-50	8.7	1 024 × 512	118	70.7
SeaFormer-B-50	8.7	1 024 × 512	142	72.2
BiSeNet-R18	49.0	1 536 × 768	907	74.8
STDC1-Seg75	14.2	1 536 × 768	585	74.5
STDC2-Seg75	22.2	1 536 × 768	771	77.0
DDRNet-23-S	5.7	2 048 × 1 024	625	77.8
DDRNet-23	20.1	2 048 × 1 024	1 780	79.5
RTFormer-S	4.8	2 048 × 1 024	1 839	76.3
RTFormer-B	16.8	2 048 × 1 024	3 288	79.3
PIDNet-S	7.6	2 048 × 1 024	1 027	78.8
PIDNet-M	34.4	2 048 × 1 024	2 821	80.1
PIDNet-L	36.9	2 048 × 1 024	3 472	80.9
TopFormer-B-100	8.7	2 048 × 1 024	498	76.3
SeaFormer-B-100	8.7	2 048 × 1 024	569	77.7

际应用中，受限于计算资源、交互需求和成本，实时语义分割则具有更多的应用场景。随着深度学习技术的不断发展和进步，实时语义分割算法也在频繁更新迭代。致力于使相关领域的研究者能快速把握实时语义分割算法的应用与设计，本文对实时分割领域内常用的模型压缩技术、常用的 CNN 和 Transformer 高效模块的设计进行了分析。而且，为了使研究者能更清晰地把握实时语义分割算法的发展，本文从网络架构和模块设计细节、基础框架和应用场景等多个角度讨论了各类算法等基本思想和基本特点。针对现有各类算法的局限性，本文提出了未来的改进方向。除此之外，为了方便研究者开展相关研究，本文详细介绍了实时语义分割任务中常用的数据集和评价指标，且全面汇总了各个类别实时语义分割算法的性能。对于需要快速对比各种实时语义分割算法效率的研究者和从业人员，本文还提供了统一设备的定量评估以及直观的速度—精度平衡比较。

尽管实时语义分割领域经过多年的发展，已经取得了巨大的成功，但仍存在一些挑战和未充分探索的难题：

1）Transformer 在实时分割领域的应用。由表 4 可以看出，现有的实时语义分割算法，以 Transformer 架构或 Transformer-CNN 混合框架为主要架构的网络较少，大部分网络还是以 CNN 为主要的模型框架，Transformer 在实时分割领域的应用还没有被充分挖掘。这可能是由于 CNN 在部署时具有更好的优化和加速效果，且在模型和数据量都较小时，CNN 更加容易训练。然而 Transformer 的高质量上下文提取能力（Liu 等，2021；Fang 等，2022）和灵活的掩码生成方式（Cheng 等，2021，2022）已经有力推动了语义分割领域的发展。如何将广泛的轻量化 Transformer 设计的研究结果（Mehta 和 Rastegari，2022b；Chen 等，2022；Pan 等，2022）应用到实时语义分割领域，进一步提高实时语义分割任务的性能依然是目前的一项挑战。

2）更多样的数据集。目前实时语义分割算法使用的数据集主要是两个自动驾驶场景的数据集 Cityscapes 和 CamVid。在类别更多、更具挑战性的通用场景数据集上或其他类型的专用场景分割数据集上的探索可以进一步验证实时语义分割算法的泛化性和通用性。

3）边缘设备上的应用。目前大部分实时语义分割算法还是在桌面 GPU 进行通用的实时语义分割算法的设计。然而在实际应用领域，如针对移动设备和边缘设备设计和验证的实时语义分割算法仍然留有较大的空缺。这些移动设备和边缘设备面临着内存资源有限、计算效率低等问题。在移动设备和边缘设备上的实时语义分割网络的精度—速度平衡显著落后于桌面 GPU 上的实时语义分割网络。推动实时语义分割网络更好地与具体应用领域的结合将是实时分割领域未来重要的发展方向。

4）大模型知识迁移。视觉大模型在近期取得了巨大的进展，推动了包括语义分割在内的多项视觉任务的发展。专注于小模型的实时语义分割任务如何从视觉大模型中受益，仍是目前有待探索的问题。知识迁移是其中一种可能的手段，例如，Mobile-SAM（Zhang 等，2023）利用知识蒸馏学习视觉大模型 SAM（segment anything model）（Kirillov 等，2023）具有的知识，最终得到能在移动端平稳运行的轻量化 SAM 模型。

5)更丰富的评价指标。目前实时语义分割算法的评价指标主要是精度上的mIoU和速度上的FPS,比较单一。引入例如运行的功耗、每秒请求数量(query per second,QPS)等指标能够更加多元地刻画实时语义分割的应用性能。

6)多模态与弱监督应用。从训练范式来看,目前绝大部分实时语义分割网络都是在有监督的闭集上训练,需要较大的标注成本。如何将实时语义分割与其他模态的数据和更弱的监督形式结合也是未来可能的发展方向之一。

7)增量学习的应用。实时语义分割的目的是部署到实际的应用场景,而实际应用中的数据集往往会随着时间不断扩大。这种情况需要实时语义分割算法能快速适应新类别、新分布的数据,利用新的标注样本、甚至在线未标注样本快速更新模型,避免重新训练模型,这涉及增量学习相关的技术,目前在实时语义分割领域探索较少,是该领域未来值得探索的方向之一。

通过本文的回顾和展望,研究者可以快速把握实时语义分割领域目前的发展现状和未来的发展趋势。可以看出,无论是结构、性能,还是评估体系和具体应用,实时语义分割算法都还有很广阔的发展空间。

参考文献(References)

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

Boykov Y, Veksler O and Zabih R. 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(11): 1222-1239 [DOI: 10.1109/34.969114]

Brostow G J, Fauqueur J and Cipolla R. 2009. Semantic object classes in video: a high-definition ground truth database. Pattern Recognition Letters, 30(2): 88-97 [DOI: 10.1016/j.patrec.2008.04.005]

Caesar H, Uijlings J and Ferrari V. 2018. COCO-stuff: thing and stuff classes in context//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1209-1218 [DOI: 10.1109/CVPR.2018.00132]

Chen W Y, Gong X Y, Liu X M, Zhang Q, Li Y and Wang Z Y. 2020. FasterSeg: searching for faster real-time semantic segmentation [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1912.10917.pdf

Chen Y P, Dai X Y, Chen D D, Liu M C, Dong X Y, Yuan L and Liu Z C. 2022. Mobile-former: bridging mobilenet and Transformer// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5260-5269 [DOI: 10.1109/CVPR52688.2022.00520]

Cheng B W, Misra I, Schwing A G, Kirillov A and Girdhar R. 2022. Masked-attention mask Transformer for universal image segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1280-1289 [DOI: 10.1109/CVPR52688.2022.00135]

Cheng B W, Schwing A G and Kirillov A. 2021. Per-pixel classification is not all you need for semantic segmentation [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2107.06278.pdf

Contributors M. 2020. OpenMMLab Semantic Segmentation Toolbox and Benchmark [EB/OL]. [2023-08-27]. https://github.com/open-mmlab/mmsegmentation

Contributors M. 2021. OpenMMLab's Model Deployment Toolbox [EB/OL]. [2023-08-27]. https://github.com/open-mmlab/mmdeploy

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Dhanachandra N, Manglem K and Chanu Y J. 2015. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54: 764-771 [DOI: 10.1016/j.procs.2015.06.090]

Dong B, Wang P C and Wang F. 2023. Head-free lightweight semantic segmentation with linear Transformer//Proceedings of the 37th AAAI Conference on Artificial Intelligence and the 35th Conference on Innovative Applications of Artificial Intelligence and the 13th Symposium on Educational Advances in Artificial Intelligence. Washington, USA: AAAI: 516-524 [DOI: 10.1609/aaai.v37i1.25126]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2010.11929.pdf

Everingham M, Eslami S M A, Van Gool L, Williams C K I, Winn J and Zisserman A. 2015. The PASCAL Visual Object Classes challenge: a retrospective. International Journal of Computer Vision, 111(1): 98-136 [DOI: 10.1007/s11263-014-0733-5]

Fan M Y, Lai S Q, Huang J S, Wei X M, Chai Z H, Luo J F and Wei X L. 2021. Rethinking BiSeNet for real-time semantic segmentation// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9711-9720 [DOI: 10.1109/CVPR46437.2021.00959]

Fang J M, Xie L X, Wang X G, Zhang X P, Liu W Y and Tian Q. 2022. MSG-Transformer: exchanging local spatial information by

manipulating messenger tokens//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12053-12062 [DOI: 10.1109/CVPR52688.2022.01175]

Guo M H, Liu Z N, Mu T J and Hu S M. 2023. Beyond self-attention: external attention using two linear layers for visual tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5): 5436-5447 [DOI: 10.1109/TPAMI.2022.3211006]

Guo M H, Lu C Z, Hou Q B, Liu Z N, Cheng M M and Hu S M. 2022. SegNeXt: rethinking convolutional attention design for semantic segmentation[EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2209.08575.pdf

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Ho J, Kalchbrenner N, Weissenborn D and Salimans T. 2019. Axial attention in multidimensional Transformers [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1912.12180.pdf

Howard A, Sandler M, Chen B, Wang W J, Chen L C, Tan M X, Chu G, Vasudevan V, Zhu Y K, Pang R M, Adam H and Le Q. 2019. Searching for MobileNetV3//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1314-1324 [DOI: 10.1109/ICCV.2019.00140]

Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, Andreetto M and Adam H. 2017. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1704.04861.pdf

Hu H, Zhang Z, Xie Z D and Lin S. 2019. Local relation networks for image recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3463-3472 [DOI: 10.1109/ICCV.2019.00356]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Huang Z L, Wang X G, Huang L C, Huang C, Wei Y C and Liu W Y. 2019. CCNet: criss-cross attention for semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 603-612 [DOI: 10.1109/ICCV.2019.00069]

Jiang X T, Wang H, Chen Y L, Wu Z Q, Wang L C, Zou B, Yang Y F, Cui Z Y, Cai Y, Yu T H, Lv C F and Wu Z H. 2020. MNN: a universal and efficient inference engine [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2002.12418.pdf

Kass M, Witkin A and Terzopoulos D. 1988. Snakes: active contour models. International Journal of Computer Vision, 1(4): 321-331 [DOI: 10.1007/BF00133570]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2304.02643.pdf

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc.: 1097-1105

Li G, Yun I, Kim J and Kim J. 2019c. DABNet: depth-wise asymmetric bottleneck for real-time semantic segmentation[EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1907.11357.pdf

Li H C, Xiong P F, Fan H Q and Sun J. 2019a. DFANet: deep feature aggregation for real-time semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9514-9523 [DOI: 10.1109/CVPR.2019.00975]

Li X, Zhou Y M, Pan Z and Feng J S. 2019b. Partial order pruning: for best speed/accuracy trade-off in neural architecture search//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9137-9145 [DOI: 10.1109/CVPR.2019.00936]

Li X T, You A S, Zhu Z, Zhao H L, Yang M K, Yang K Y, Tan S H and Tong Y H. 2020. Semantic flow for fast and accurate scene parsing//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 775-793 [DOI: 10.1007/978-3-030-58452-8_45]

Li Y Y, Yang C D, Zhao P, Yuan G, Niu W, Guan J X, Tang H, Qin M H, Jin Q, Ren B, Lin X and Wang Y Z. 2023. Towards real-time segmentation on the edge//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 1468-1476 [DOI: 10.1609/aaai.v37i2.25232]

Lin G S, Milan A, Shen C H and Reid I. 2017. RefineNet: multi-path refinement networks for high-resolution semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5168-5177 [DOI: 10.1109/CVPR.2017.549]

Liu C X, Chen L C, Schroff F, Adam H, Hua W, Yuille A L and Li F F. 2019. Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 82-92 [DOI: 10.1109/CVPR.2019.00017]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Mehta S, Rastegari M, Caspi A, Shapiro L and Hajishirzi H. 2018. ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 561-580 [DOI: 10.1007/978-3-030-01249-6_34]

Mehta S and Rastegari M. 2022a. Mobilevit: light-weight, general-purpose, and mobile-friendly vision Transformer [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2110.02178.pdf

Mehta S and Rastegari M. 2022b. Separable self-attention for mobile vision Transformers [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2206.02680.pdf

Mottaghi R, Chen X J, Liu X B, Cho N G, Lee S W, Fidler S, Urtasun R and Yuille A. 2014. The role of context for object detection and semantic segmentation in the wild//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 891-898 [DOI: 10.1109/CVPR.2014.119]

Najman L and Schmitt M. 1994. Watershed of a continuous function. Signal Processing, 38(1): 99-112 [DOI: 10.1016/0165-1684(94)90059-0]

Nirkin Y, Wolf L and Hassner T. 2021. HyperSeg: patch-wise hypernetwork for real-time semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4060-4069 [DOI: 10.1109/CVPR46437.2021.00405]

Nock R and Nielsen F. 2004. Statistical region merging. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(11): 1452-1458 [DOI: 10.1109/TPAMI.2004.110]

Oršic M, Krešo I, Bevandic P and Šegvic S. 2019. In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12599-12608 [DOI: 10.1109/CVPR.2019.01289]

Otsu N. 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, Cybernetics, 9(1): 62-66 [DOI: 10.1109/tsmc.1979.4310076]

Pan H H, Hong Y D, Sun W C and Jia Y S. 2023. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. IEEE Transactions on Intelligent Transportation Systems, 24(3): 3448-3460 [DOI: 10.1109/TITS.2022.3228042]

Pan J T, Bulat A, Tan F W, Zhu X T, Dudziak L, Li H S, Tzimiropoulos G and Martinez B. 2022. EdgeViTs: competing light-weight cnns on mobile devices with vision Transformers//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 294-311 [DOI: 10.1007/978-3-031-20083-0_18]

Paszke A, Chaurasia A, Kim S and Culurciello E. 2016. ENet: a deep neural network architecture for real-time semantic segmentation [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1606.02147.pdf

Plath N, Toussaint M and Nakajima S. 2009. Multi-class image segmentation using conditional random fields and global classification//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM: 817-824 [DOI: 10.1145/1553374.1553479]

Poudel R P K, Liwicki S and Cipolla R. 2019. Fast-SCNN: fast semantic segmentation network [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1902.04502.pdf

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Sandler M, Howard A G, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1409.1556.pdf

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5686-5696 [DOI: 10.1109/CVPR.2019.00584]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tan M X and Le Q V. 2020. EfficientNet: rethinking model scaling for convolutional neural networks [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/1905.11946.pdf

Tan M X and Le Q. 2021. EfficientNetV2: smaller models and faster training [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2104.00298.pdf

Wadekar S N and Chaurasia A. 2022. MobileViTv3: mobile-friendly vision Transformer with simple and effective fusion of local, global and input features [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2209.15159.pdf

Wan Q, Huang Z L, Lu J C, Yu G and Zhang L. 2023. SeaFormer: squeeze-enhanced axial Transformer for mobile semantic segmentation [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2301.13156.pdf

Wang J, Gou C H, Wu Q M, Feng H C, Han J Y, Ding E R and Wang J D. 2022. RTFormer: efficient design for real-time semantic segmentation with Transformer [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2210.07124.pdf

Wang N, Hou Z Q, Pu L, Ma S G and Cheng H H. 2022. Real-time semantic segmentation analysis based on cavity separable convolution and attention mechanism. Journal of Image and Graphics,

27(4)：1216-1225（王囡，侯志强，蒲磊，马素刚，程环环. 2022. 空洞可分离卷积和注意力机制的实时语义分割. 中国图象图形学报，27(4)：1216-1225）[DOI：10.11834/jig.200729]

Wang W H，Dai J F，Chen Z，Huang Z H，Li Z Q，Zhu X Z，Hu X W，Lu T，Lu L W，Li H S，Wang X G and Qiao Y. 2023. InternImage：exploring large-scale vision foundation models with deformable convolutions//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：14408-14419 [DOI：10.1109/CVPR52729.2023.01385]

Wang W H，Xie E Z，Li X，Fan D P，Song K T，Liang D，Lu T，Luo P and Shao L. 2021. Pyramid vision Transformer：a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：548-558 [DOI：10.1109/ICCV48922.2021.00061]

Xie E Z，Wang W H，Yu Z D，Anandkumar A，Alvarez J M and Luo P. 2021. SegFormer：simple and efficient design for semantic segmentation with Transformers[EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2105.15203.pdf

Xu J C，Xiong Z X and Bhattacharyya S P. 2023. PIDNet：a real-time semantic segmentation network inspired by PID controllers//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：19529-19539 [DOI：10.1109/CVPR52729.2023.01871]

Yang C D，Zhao P，Li Y Y，Niu W，Guan J X，Tang H，Qin M H，Ren B，Lin X and Wang Y Z. 2023. Pruning parameterization with Bi-level optimization for efficient semantic segmentation on the edge//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：15402-15412 [DOI：10.1109/CVPR52729.2023.01478]

Yu C，Gao C，Wang J，Yu G，Shen C and Sang N. 2021. BiSeNet V2：bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision，129(11)：3051-3068 [DOI：10.1007/s11263-021-01515-2]

Yu C Q，Wang J B，Peng C，Gao C X，Yu G and Sang N. 2018. BiSeNet：bilateral segmentation network for real-time semantic segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：334-349 [DOI：10.1007/978-3-030-01261-8_20]

Yuan Y H，Huang L，Guo J Y，Zhang C，Chen X L and Wang J D. 2021. OCNet：object context for semantic segmentation. International Journal of Computer Vision，129(8)：2375-2398 [DOI：10.1007/s11263-021-01465-9]

Zhang C N，Han D S，Qiao Y，Kim J U，Bae S H，Lee S and Hong C S. 2023. Faster segment anything：towards lightweight SAM for mobile applications [EB/OL]. [2023-08-27]. https://arxiv.org/pdf/2306.14289.pdf

Zhang W Q，Huang Z L，Luo G Z，Chen T，Wang X G，Liu W Y，Yu G and Shen C H. 2022. TopFormer：token pyramid Transformer for mobile semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：12073-12083 [DOI：10.1109/CVPR52688.2022.01177]

Zhang X，Xu H M，Mo H，Tan J C，Yang C，Wang L and Ren W Q. 2021. DCNAS：densely connected neural architecture search for semantic image segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：13951-13962

Zhang X Y，Zhou X Y，Lin M X and Sun J. 2018. ShuffleNet：an extremely efficient convolutional neural network for mobile devices//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：6848-6856 [DOI：10.1109/CVPR.2018.00716]

Zhao H S，Qi X J，Shen X Y，Shi J P and Jia J Y. 2018. ICNet for real-time semantic segmentation on high-resolution images//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：418-434 [DOI：10.1007/978-3-030-01219-9_25]

Zhao H S，Shi J P，Qi X J，Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：6230-6239 [DOI：10.1109/CVPR.2017.660]

Zheng S X，Lu J C，Zhao H S，Zhu X T，Luo Z K，Wang Y B，Fu Y W，Feng J F，Xiang T，Torr P H S and Zhang L. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：6877-6886 [DOI：10.1109/CVPR46437.2021.00681]

Zhou B L，Zhao H，Puig X，Fidler S，Barriuso A and Torralba A. 2017. Scene parsing through ADE20K dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：5122-5130 [DOI：10.1109/CVPR.2017.544]

Zhuang J T，Yang J L，Gu L and Dvornek N. 2019. Shelfnet for fast semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul，Korea (South)：IEEE：847-856 [DOI：10.1109/ICCVW.2019.00113]

作者简介

高常鑫，男，教授，主要研究方向为计算机视觉和图像/视频理解。E-mail：cgao@hust.edu.cn

桑农，通信作者，男，教授，主要研究方向为模式识别和计算机视觉。E-mail：nsang@hust.edu.cn

徐正泽，男，硕士研究生，主要研究方向为语义分割和模型轻量化。E-mail：zhengzexu@hust.edu.cn

吴东岳，男，博士研究生，主要研究方向为语义分割。E-mail：dongyue_wu@hust.edu.cn

余昌黔，男，算法工程师，主要研究方向为计算机视觉和自动驾驶。E-mail：yuchangqian@meituan.com