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

《计算机工程与应用》网络首发论文

题目：全卷积神经网络图像语义分割方法综述
作者：张鑫，姚庆安，赵健，金镇君，冯云丛
网络首发日期：2021-12-20
引用格式：张鑫，姚庆安，赵健，金镇君，冯云丛．全卷积神经网络图像语义分割方法综述[J/OL]．计算机工程与应用．
https://kns.cnki.net/kcms/detail/11.2127.TP.20211220.0930.002.html

网络首发时间：2021-12-20 14:00:27
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20211220.0930.002.html

全卷积神经网络图像语义分割方法综述

张鑫，姚庆安，赵健，金镇君，冯云丛

长春工业大学 计算机科学与工程学院，长春 130102

摘 要：图像语义分割是计算机视觉领域的热点研究课题，随着全卷积神经网络的迅速兴起，图像语义分割和全卷积神经网络的融合发展取得了非常卓越的成绩。通过对近年来高质量文献的收集，重点对全卷积神经网络图像语义分割方法进行总结。将收集的文献，按照应用场景的不同，划分为经典语义分割、实时性语义分割和 RGBD 语义分割，然后对具有代表性的分割方法进行阐述。同时归纳了常用的公共数据集和性能的评价指标，并对常用数据集上的实验进行分析总结，最后对全卷积神经网络未来可能的研究方向进行展望。

关键词：图像语义分割；计算机视觉；全卷积神经网络

文献标志码：A 中图分类号：TP391 doi：10.3778/j.issn.1002-8331.2109-0091

Image Semantic Segmentation Based on Fully Convolutional Neural Network

ZHANG Xin, YAO Qing’an, ZHAO Jian, JIN Zhenjun, FENG Yuncong
School of Computer Science and Engineering, Changchun University of Technology, Changchun 130102, China

Abstract: Image semantic segmentation is a hot research topic in the field of computer vision. With the rapid rise of fully convolutional neural networks, the development of fusion of image semantic segmentation and fully convolutional networks has shown very bright results. Through the collection of high-quality literature in recent years, the focus is on the summary of full convolutional neural network image semantic segmentation methods. The collected literature is divided into classical semantic segmentation, real-time semantic segmentation and RGBD semantic segmentation according to the application scenarios, and then the representative segmentation methods are described. Commonly used public datasets and evaluation metrics for performance are also summarized, and experiments on commonly used datasets are analyzed and summarized. Finally, the possible future research directions of fully convolutional neural networks are prospected.

Key words: image semantic segmentation; computer vision; fully convolutional neural network

语义分割是将场景图像分割为若干个有意义的图像区域，并对不同图像区域分配指定标签的过程。然而语义分割的难点主要体现在两个方面：一是类内实例间的相异性和类间物体的相似性；二是复杂的背景大幅度提高了语义分割的难度。

图像语义分割的传统方法是利用图片中边缘、颜色、纹理等特征将图片分割成不同的区域。如基于阈值[1-4]、边缘[5-8]、聚类[9-12]、图论[13-16]等常用的经典分

基金项目：吉林省科技发展规划重点研发项目（20200401076GX）；吉林省教育厅“十三五”科学技术研究规划项目（JJKH20200678KJ）；符号计算与知识工程教育部重点实验室 2020 年度开放基金项目（93K172020K05）。

作者简介：张鑫(1995-)，女，硕士研究生，CCF 会员，主要研究方向为图像分割；姚庆安(1975-)，通信作者，男，硕士，副教授，CCF 会员，主要研究方向为图像处理、智能数据处理、深度学习，E-mail: yao@ccut.edu.cn；赵健（1997-），男，硕士研究生，主要研究方向为图像分割；金镇君（1980-），男，博士，讲师，CCF 会员，主要研究方向为云计算、边缘计算、机器学习、智能机器人；冯云丛（1987-），女，博士，讲师，CCF 会员，主要研究方向为图像分割、深度学习。

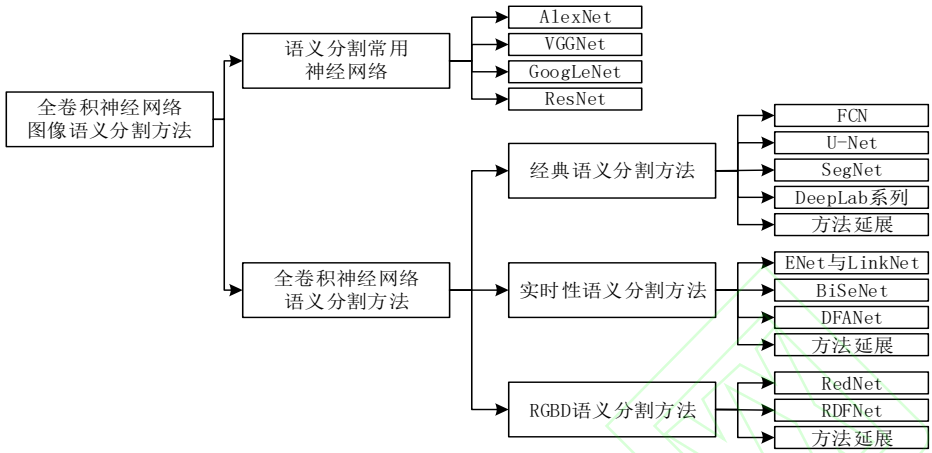

图 1　全卷积神经网络图像语义分割方法分类

Fig.1　Classification of semantic segmentation methods for fully convolutional neural network images

割方法。由于计算机的硬件设备限制，图像分割技术仅能对灰度图像进行处理，后期才逐渐发展到对 RGB 图像进行处理的阶段。随着 GPU 的飞速发展，深度学习（Deep Learing,DL）[17]技术为语义分割技术的发展提供有效的支撑。研究人员使用卷积神经网络（Convolutional Neural Network, CNN），通过端到端的训练方式推理每个像素的语义信息并实现有意义图形区域的分类。由于 CNN 特征学习和表达能力的优势明显，使其成为图像语义分割领域优先考虑的方法。

2015 年 IEEE 国际计算机视觉与模式识别会议(IEEE Conference on Comper Vision and Pattern Recognition)，Long 等人提出了全卷积神经网络(Fully Convolutional Networks for semantic segmentation, FCN)[18]，至此图像语义分割进入了全卷积神经网络时期。全卷积神经网络在深度学习中表现出强大的潜能，逐渐成为解决图像语义分割问题的首选。对比前两个时期，全卷积神经网络通过像素级到像素级的训练方式，能够获得更高的精度和更好的运算效率，已经成为图像语义分割的研究热点。然而随着对该领域研究的深入，如何有效提高不同应用场景下图像语义分割的精确度一直是该领域的研究痛点。

目前存在的文献综述[19-23]，虽然对图像语义分割进行了总结，但是普遍缺乏对于应用场景的深刻了解，如文献[19]仅对语义分割进行整体概述介绍；文献[20]将语义分割分为传统方法和深度学习的方法展开分析；文献[21]将语义分割进一步细化为全监督和弱监督学习方法进行阐述；文献[22]从语义分割研究领域入手进行梳理；以及文献[23]侧重于主流语义分割算法的总结。但是这些综述文献都未能根据不同应用领域有针对性的对精度需求和创新方向进行详细的解释，因此对全卷积神经网络图像语义分割方法进行综述必不可少。经过总结和整理了相关研究后得到，如图 1 所示。从语义分割常用神经网络引入。按照图像语义分割模型的应用场景不同，分为经典语义分割方法、实时性语义分割方法和 RGBD 语义分割方法，对每类具有代表性的方法进行叙述总结，并对不同应用场景下的方法进行延展。

第 1 节介绍语义分割常用神经网络，第 2 节对全卷积神经网络图像语义分割方法进行阐述，并对不同应用场景下每类具有代表性的算法展开叙述和延展，第 3 节对图像语义分割的相关实验进行分析和总结，介绍公共数据集和算法性能评价指标，第 4 节对图像语义分割未来的发展方向进行展望。

1 语义分割常用神经网络

1.1　AlexNet

2012 年 Krizhevsky 等人提出的 AlexNet[24]架构以绝对优势在 ImageNet 竞赛中以 84.6%的准确率夺得冠军，掀起 CNN 在各个领域的研究热潮。AlexNet 网络结构共 8 层，包括 5 个卷积层和 3 个全连接层。其网络采用 Relu 激活函数，局部响应归一化（Local Response Normalization,LRN）提高模型的泛化能力，应用重叠池化（Overlapping）和随机丢弃（Dropout）预防过拟合。

1.2　VGGNet

2014 年由牛津大学计算机视觉组合和 Google DeepMind 公司提出的 VGGNet[25]，在 ImageNet 竞赛中以精确度 92.7%获得亚军。它与 AlexNet[24]网络相

比，主要创新是叠加使用 3×3 滤波器将网络深度提升到 16-19 个权重层，使其在感受野不变的条件下，减少参数计算，同时网络深度增加有效的改善网络对语义信息的提取。

1.3 GoogLeNet

2014 年 Szegedy 等人提出的 GoogLeNet[26]以精确度 93.3%取得 ImageNet 竞赛中的冠军。它采用比 VGGNet[25]更深的网络结构，共 22 层，最亮眼的是提出 Inception 模块。Inception 将不同感受野的滤波器对输入图进行卷积和池化，通过 1×1 卷积降维后拼接输出。GoogLeNet 将这些模块堆叠在一起形成一个抽象的网络结构。同时抛弃全连接层。Inception 的引入不仅削减网络复杂性，而且还考虑到内存和计算成本。

表 1 图像语义分割方法分析与总结

Table 1 Analysis and summary of image semantic segmentation methods

方法类别	代表算法和算法特点	方法特点	优缺点总结
经典语义分割方法	**FCN**[18]：将当前分类网络改编为全卷积网络并连接全局信息和局部信息，实现任意尺寸图片输入输出 **U-Net**[28]：使用跳跃连接将编码网络的特征图与相对应的解码网络的特征图直接拼接 **SegNet**[29]：编解码层使用对应编码器层存储的最大池化索引对特征图进行上采样，实现边界特征的精准定位 **DeepLab 系列**[30-33]：引入 ASPP 模块捕获多尺度上下文语义信息，同时解码模块的加入，细化边界分割的准确度	（1）对 FCN 进行优化改进，提取稠密的图像特征信息，增大感受野，有效的捕获上下文语义信息； （2）通过带孔卷积、带孔空间金字塔池化等技术获得多尺度信息，对空间变换具有较高的不变性。	优点：针对 FCN 的不足进行改进，有效稠密反卷积过程、增强感受野、获取图像的多尺度表示，提高特征图的分割精度。 缺点：模型参数量大、分割速度慢，小尺度物体的分割结果不明显且易丢失边界细节信息。
实时性语义分割方法	**ENet**[49]：采用较大编码器和较小解码器结构，减少了模型参数量的同时保持良好的分割精度 **LinkNet**[50]：将编码器和解码器对应特征图直接连接从而提高准确率 **BiSeNet**[51]：提出双向语义分割网络，生成高分辨率特征图的同时获得丰富的感受野，引入特征融合模块充分结合两个分支的特征 **DFANet**[52]：使用改进的轻量级 Xception 网络，分别通过子网络和子阶段级联聚合判别特征	（1）对经典语义分割模型进行剪枝、量化等，去除冗余层，轻量化网络模型，有效提升语义分割的速度； （2）通过设计轻量级网络模型 Xception 等来平衡网络分割精度和处理速度。	优点：合理去除不必要的冗余层，使得模型参数量减少，运行速度提高，适用于移动端设备。 缺点：由于模型结构的简化，图像信息丢失严重，分割精度降低。
RGBD 语义分割方法	**RedNet**[59]：使用残差模块作为基本的积木块应用于编码器和解码器中，并提出一种金字塔监督训练方案 **RDFNet**[60]：明确利用下采样过程中的有用信息，实现远程残差连接提高分割精度，引入链式残差池化结构有效捕获上下文信息	（1）使用深度图对彩色图进行图像语义信息的补充，设计 RGBD 网络模型，有效分割复杂场景下的多尺度信息； （2）通过深度图和彩色图处理方式、融合时期的不同来提高图像语义分割精度。	优点：引入深度图对彩色图进行语义信息的补充，大大提升复杂场景下多尺度信息的分割精度。 缺点：彩色图和深度图不对等融合，造成图像信息丢失，难以达到预估的分割效果。

1.4 ResNet

2015 年由微软研究院提出的 ResNet[27]以精度 96.4%成为 ImageNet 竞赛的冠军。其残差模块，能够成功的训练高达 152 层深的网络结构，残差结构通过引入跳跃连接来解决梯度回传消失的问题，真正解决网络深层架构的问题。

2 全卷积神经网络图像语义分割方法

全卷积神经网络对图像语义分割具有里程碑的意义。按照应用场景不同，从高分割精度的经典语义分割方法，高效率的实时性语义分割方法和复杂场景的 RGBD 语义分割方法三个方面进行阐述。表 1 对这三类方法从方法特点、优缺点等几个方面进行了分析和比较。下面对其进行详细的介绍。

2.1 经典语义分割方法

经典语义分割在应用中具有里程碑的意义。从经典网络模型 FCN[18]、U-Net[28]、SegNet[29]、DeepLab [30-33]和方法延展展开详细的叙述。

2.1.1 FCN

2015 年 Long 等人提出全卷积网络（Fully Convolutional Network,FCN）[18]，首次实现任意图片大小输入的像素级语义分割任务，其结构如图 2 所示。FCN 将 CNN 模型中的全连接层替换为全卷积层以实现像素级的密集预测，使用反卷积对特征图进行上采样，并提出跳层连接充分融合全局语义信息和局部位置信息，实现精确分割。同时 FCN 微调常用经典网络的预训练权重来加快网络收敛速度。

尽管 FCN 实现了分类网络到分割网络的转换，但是 FCN 也有许多不足：1）上采样过程粗糙，导致特征图语义信息丢失严重，严重影响分割精度；2）跳跃连接未能充分利用图片的上下文信息和空间位置信息，导致全局信息和局部信息的利用率低；3）网络整体规模庞大，参数多，导致计算时间过长。正是 FCN 的提出与不足，才为全卷积神经网络的发展奠定了里程碑的基础。

2.1.2 U-Net

2015 年 Ronneberger 等人提出的用于医学图像分割的 U-Net[28]，是一个对称编解码网络结构，如图 3 所示。U-Net 的独特之处是使用镜像折叠外推缺失的上下文信息，补充输入图片的语义信息，通过跳跃连接将编解码器中的特征图直接拼接，有效的融合了深层细节信息和浅层语义信息。

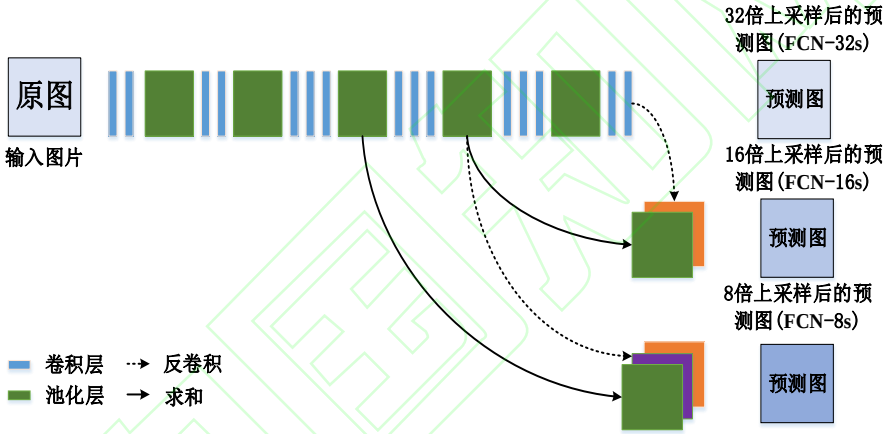

图 2 FCN 结构图[18]

Fig.2 Structure diagram of FCN[18]

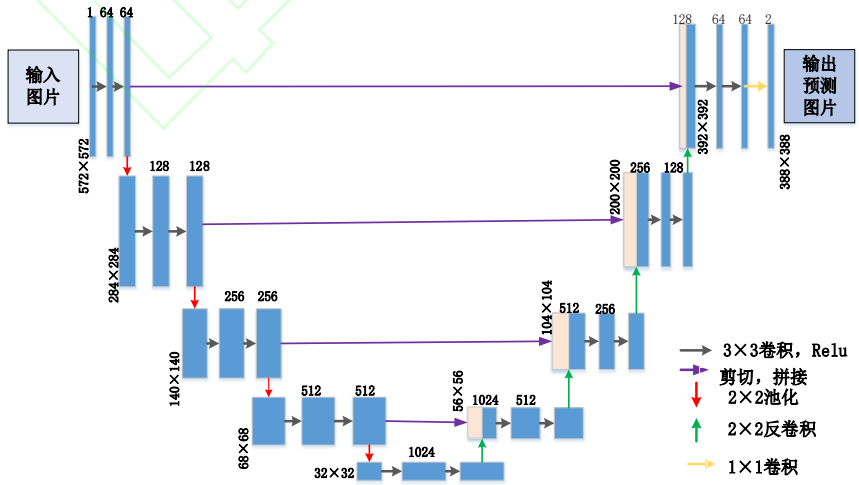

图 3 U-Net 网络架构[28]

Fig.3 Network architecture of U-Net[28]

同时 U-Net 提出一个加权交叉熵损失函数，如公式（1）所示。

$$E=\sum_{x\in\Omega}\omega(x)\lg(p_{l(x)}(x)) \quad (1)$$

其中，ω 是一个权重图谱，通过形态学操作的计算方式计算获得，如公式（2）所示：

$$\omega(x)=\omega_c(x)+\omega_0\cdot\exp(-\frac{(d_1(x)+d_2(x))^2}{2\delta^2}) \quad (2)$$

损失函数的目的是紧密细胞间的分割。此外，U-Net 网络采用了自适应权重初始化方法：标准方差为 $\sqrt{2/N}$（N 为神经元输入结点的数量）的高斯分布初始化权重。

显然 U-Net 网络编解码器同层间直接进行跳跃连接，特征图之间语义差别大，不可避免的增加了网络学习的难度。因此基于 U-Net 出现了一系列改进的网络结构，如 UNet++[34]、UNet3+[35]、Attention UNet[36] 等，目的是充分利用深浅层语义信息，稠密特征图融合，提高语义分割精度。

2.1.3 SegNet

SegNet[29]将对称编解码结构推向高潮，其结构如图 4 所示。SegNet 没有跳层结构，使用批标准化（Batch Normal,BN）加快收敛抑制过拟合，其最大的创新是上采样使用最大池化（Max-pooling）方法[22]，即编码阶段的下采样过程中保留最大池化值和对应索引值，在解码阶段利用最大池化索引对输入的特征图进行上采样，最后经过卷积层得到稠密的特征图。SegNet 使用极少数据量保存索引值却将低分辨率特征映射到输入分辨率中，实现对边界特征的精确定位。

SegNet 充分考虑内存占用问题，在空间复杂度上具有优势，然而除非存储量十分有限，SegNet 就其网络本身，优势并不明显。

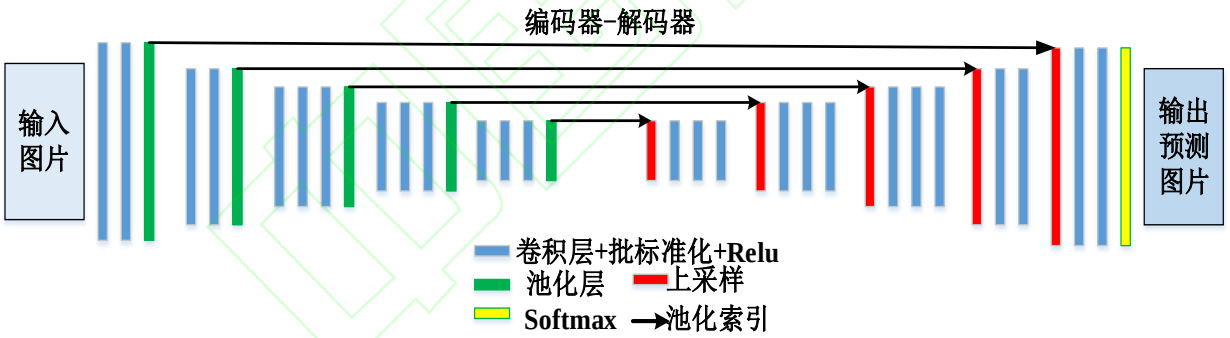

图 4 SegNet 网络架构[29]

Fig.4 Network architecture of SegNet[29]

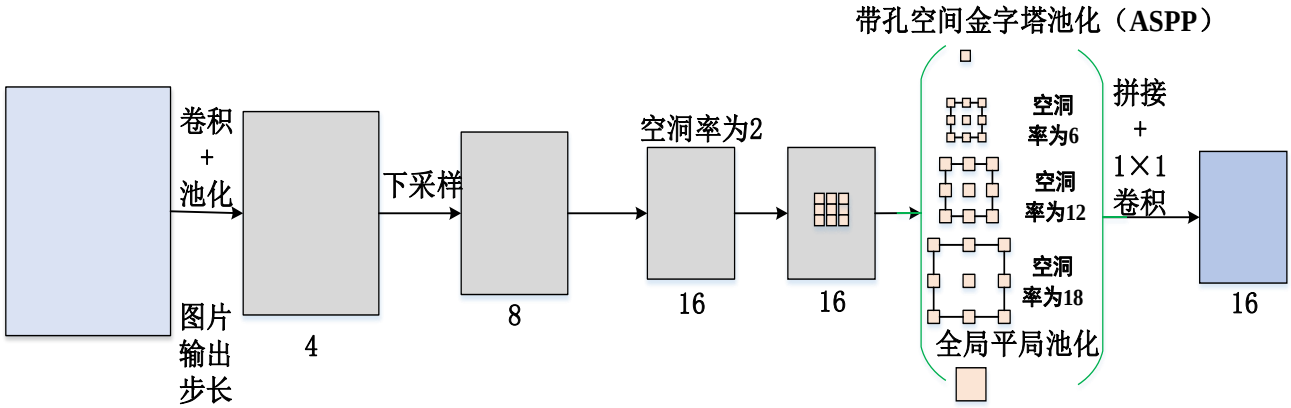

图 5 DeepLab v3 模型结构[32]

Fig.5 DeepLab v3 model structure[32]

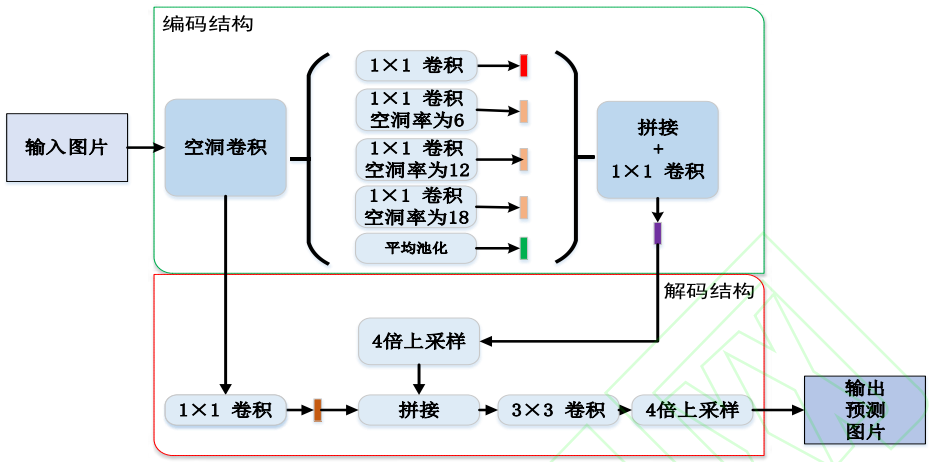

图 6 DeepLab v3+模型结构[33]

Fig.6 DeepLab v3 model structure[33]

2.1.4 DeepLab 系列

2016 年 Chen 等人提出的 DeepLab v1[30]，抛弃 VGG16[25]的全连接层，将最后两次池化步长改为 1，深度卷积网络（Deep Convolutional Neural Networ, DCNN）的部分卷积层替换为空洞卷积（atrous convolution），通过增大感受野来获得更多的语义信息。同时提出全连接条件随机场（Conndtional Random Field,CRF）的后处理方法对分割结果图进行细节增强，但是易丢失图片中详尽的细节信息。

2017 年，Chen 等人对 DeepLab v1 进行扩展提出了 DeepLab v2[31]，使用网络为 ResNet[27]并提出带孔空间金字塔池化（Atrous Spatial Pyramid Pooling, ASPP）模块，实现多尺度目标的处理。多尺度特征提取的采样率（rate）分别为：6,12,18,24。同时 DeepLab v2 仍然需要 CRF 做后处理。

同年 12 月，Chen 等人在 DeepLab v1、v2 的基础上提出 DeepLab v3[32]，如图 5 所示。使用 ResNet[27]，在级联 ASPP 模块中增加了全局平均池化和 1×1 的卷积层，有效处理多尺度分割目标的任务，同时引入批标准化 BatchNormal（BN）。DeepLab v3 在丢弃 CRF 后处理的情况下，取得比 DeepLab v1 和 DeepLab v2 更高的精确值。

2018 年 Chen 等人提出 DeepLab v3+[33]，结合编解码结构设计了一种新的编解码模型，如图 6 所示。以 DeepLab v3 为编码器结构提取丰富的上下文信息，简单有效的解码器用于恢复语义对象边界信息，同时在 ASPP 模块和解码网络中添加深度可分离深度卷积（Depth wise Separable Convolution），提高了网络的运行速率和鲁棒性，大幅度提升了分割准确度。

DeepLab 系列尽管成果斐然，但就其网络而言，存在细节分割丢失严重、计算量大、上下层语义信息关联性差等问题。因此基于 DeepLab 网络结构以及针对网络某个问题提出很多新的网络结构，如文献[37,38]等，有针对性的完善网络结构，解决多尺度目标的分割任务。

2.1.5 方法延展

Lin 等人提出了多路径细化网络（RefineNet）[39]。RefineNet 用于解决空间信息丢失问题，首先输入来自 ResNet[27]网络中 4 个不同尺度、不同分辨率的特征图，然后把 4 个特征图分别送入由残差卷积单元构成的 4 个精细化模块（RefineNet block）中求和，充分利用下采样过程中的所有可用信息，有效的实现高分辨率的预测任务。

Zhao 等人提出金字塔场景解析网络（PSPNet）[40]，提出一个金字塔池化模块。该模块级联多个具有不同步长的全局池化操作来聚合更多的上下文信息实现高质量的像素级场景解析，同时提出深度监督优化策略，降低模型优化的难度。

Peng 等人提出 GCN[41]。GCN 提出对于输入图片进行分类和定位操作时有效的感受野至关重要，提出 GCN 模块采用大的卷积核替代通常小卷积核堆叠的方法来提高感受野，使用边界细化模块细化边界信息。论文作者提出当卷积核大小为 11 时效果最好。

Yu 等人提出 DFN[42]网络。DFN 从宏观角度出发针对类内不一致和类间不一致的问题，提出平滑网络（Smooth Network，SN）和边界网络(Border Network，

BN)。前者通过引入注意力机制和全局平均池化选择更具区分性的类别特征信息，后者通过深度语义边界监督来区分不同类别的特征。同时还有改编于 U-Net[28]的网络 Fusionnet[43]用于自动分割连接组学数据中的神经元结构，它在网络中引入基于求和的跳跃连接，用更深的网络结构来实现更精确的分割。DeconvNet[44]的解码器部分将反卷积和反池化组成上采样组件，逐像素分类完成分割任务。还有针对视频的语义分割的文献[45-47]。文献[45]提出将静态图像语义分割的神经网络模型转换为视频数据的神经网络技术，主要原则是使用相邻帧的光流来跨时间扭曲内部网络表示，提高性能的端到端训练。文献[46]提出基于时空变压器门控递归单元 STGRU (Spatio-Temporal Transformer Gated Recurrent Unit) 的 GRFP 模型，结合多帧未标注信息来提高分割性能。以及文献[47]采用类似生成对抗网络（Generative Adversarial Networks ,GAN）[48]的网络结构。通过预测未来帧学习判别特征，与单帧的简单分割相比，语义分割效果显著。由此可知，经典模型发展相对饱和，横向领域研究将会为其精度提升注入新的血液。

2.2 实时性语义分割方法

实时执行像素级语义分割的能力在延时满足的应用中至关重要，针对这一应用场景，实时性语义分割应运而生。通过具有代表性的实时性网络架构 ENet[49]与 LinkNet[50]、BiseNet[51]、DFANet[52]展开阐述，并对模型优化方向提出方法延展。

2.2.1 ENet 与 LinkNet

2016 年 AdamPaszke 等人提出 ENet[49]，次年 Chaurasia 等人提出 LinkNet[50]。其中 ENet 针对低延迟操作的任务提出适合的网络模型结构，采用较大的编码结构和较小的解码结构，大大削减参数数量。同时采用 PReLUs 激活函数确保分割精度。LinkNet 则是直接将编码器和解码器对应部分连接起来提高准确率，在不增加额外操作同时保留编码层丢失的信息，减少计算量。然而编解码网络的简化，不可避免丢失空间分辨率，减弱分割精度。如何平衡语义分割精度和分割效率，成为实时性分割模型的重要突破口。

2.2.2 BiSeNet

2018 年 Yu 等人提出 BiSeNet[51]，分为空间分支路径（Spatial Path,SP）和上下文分支路径（Context Path,CP）,如图 7 所示。SP 共三层，每层包括一个步长为 2 的 3×3 的卷积，BN 层和 Relu 层，有效的保留原始图片的空间尺寸并编码丰富的空间信息。CP 采用轻量级网络 Xception 和平均池化来兼顾感受野和实时性。同时模型加入注意力机制模块（ARM）来引导特征学习，最后使用特征融合模块（FFM）将全局特征和局部特征进行有效融合。

BiSeNet 证实了实时分割中双路径网络的有效性能，但是不可避免造成算法耗时增加。STDC[53]重新思考 BiSeNet，进一步缩短了实时推理时间，削减网络冗余，也为网络瘦身提供新的研究思路。

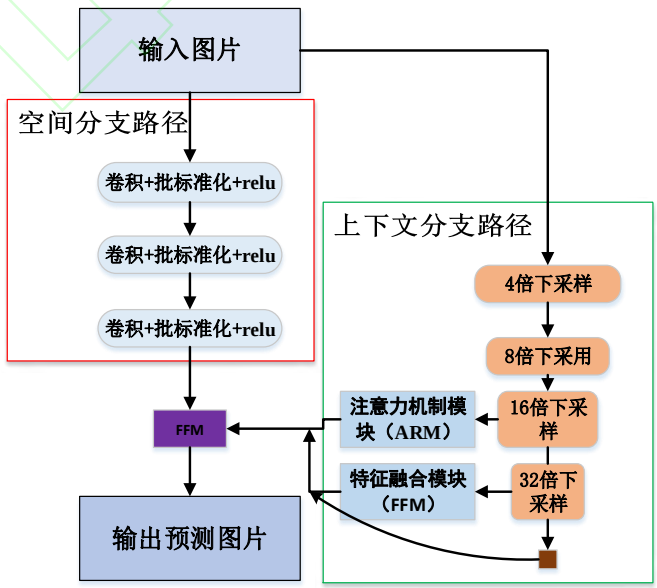

图 7 BiSeNet 模型结构[51]

Fig.7 BiSeNet model structure[51]

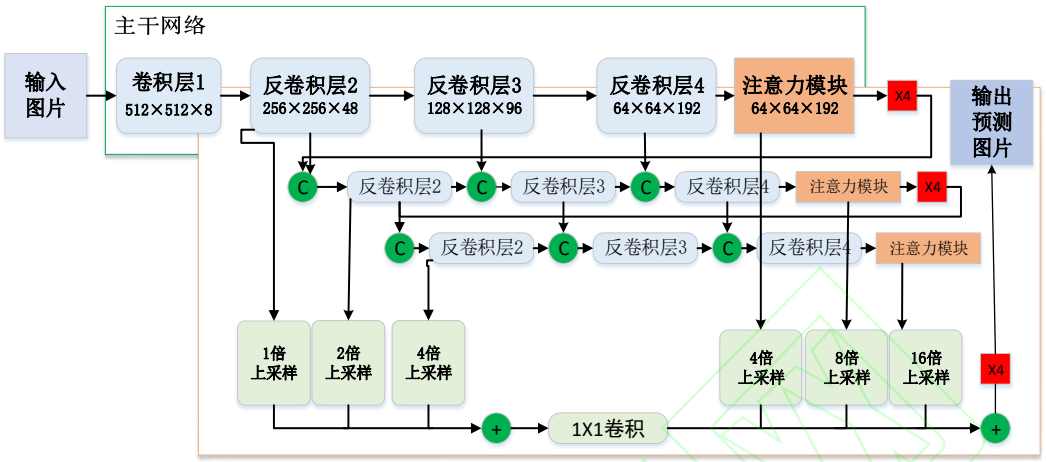

图 8 DFANet 模型结构[52]

Fig.8 DFA model structure[52]

2.2.3 DFANet

2019 年 Li 等人提出 DFANet[52]，如图 8 所示，DFANet 开起了在主流移动端处理器上做高清视频级应用的可能性。其中编码器是 3 个改进的轻量级 Xception 网络，由网络级特征聚合和阶段级特征聚合连接在一起。作者保留全连接层增加感受野，并和 1×1 卷积组成注意力模块。解码器是将编码器 3 个阶段的特征图采用双线性差值的方式上采样后融合细化语义信息。

DFANet 改进轻量级网络的思想，刷新了实时语义分割的计算量的记录。但是优化计算成本、内存占用，会损失分割精度，因此如 EsNet[54]、DFPNet[55]等网络的提出很好的平衡了实时性网络中速度和精度的追求。

2.2.4 方法延展

Light-Weight RefineNet[56]在 RefineNet[39]基础上，将网络改编为更加紧凑的架构，使其适用于在高分辨率输入图片上实现更快速率的分割任务。类似于将网络模型轻量化的模型压缩方法有模型裁剪、模型量化、知识蒸馏[57]、神经结构搜索(Neural Architecture Search，NAS) [58]等，其中模型裁剪按照裁剪规则和敏感度分析对参数进行重要性分析，剪掉不重要的网络连接。模型量化是将浮点数映量化到最低位数，使得参数计算量和模型体积减少，从而加快模型的推理速度。知识蒸馏将复杂网络的知识迁移到小网络，通常的实现过程是就用复杂网络监督小网络的训练，从而提高小网络的精度。以及 NAS 是通过模型大小和推理速度力约束来设计更高效的网络结构。因此，有效的模型瘦身和轻量化网络结构会促进实时性语义分割性能，实现对高分辨率图像的精准快速分割。

2.3 RGBD 语义分割方法

随着室内复杂场景分割问题的显露，提出 RGDB 语义分割。主要思想是使用深度图（Deep Image）对 RGB 图进行语义信息的补充。其中深度图也叫距离影像，指将从图像采集器到场景中各点的距离（深度）作为像素值的图像。首先从 RedNet[59]、RDFNet[60]来介绍 RGBD 语义分割。然后针对其算法融合阶段进行方法延展。

2.3.1 RedNet

2018 年，Jiang 等人提出的 RedNet[59]网络，如图 9 所示。RedNet 使用残差模块作为基本块应用于编码解码结构中，深度图和彩色图使用相同下采样方式。网络先短跳进行深度图和彩色图融合，再将融合结果通过远跳和同尺寸的解码器模块融合，并提出一种金字塔监督的监督训练方法来提高复杂场景的分割精度。

然而，彩色图和深度图本身差异明显，如何让深度图有效的给彩色图以语义补充，提高模型分割精度，是复杂场景下 RGBD 语义分割追求的目标。目前有文献[61,62]对深度图进行有效处理。

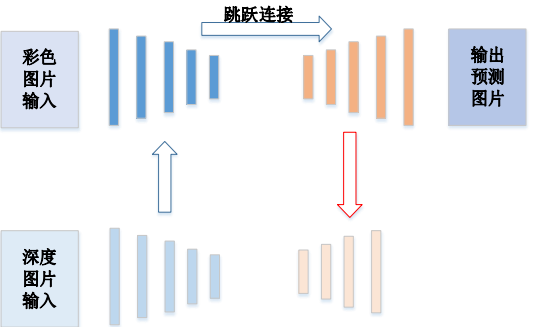

图 9 RedNet 模型结构[59]

Fig.9 RedNet model structure[59]

2.3.2 RDFNet

2017 年 Park 等人提出的 RDFNet[60]，编码部分使用多模态特征融合模块（Multi-Modal Feature Fusion, MMF），如图 10 所示。该模块充分利用彩色图和深度图之间的互补特征提取语义信息。解码器特征优化模块与 RefineNet[39]一样，采用多个级别学习融合特征的组合，以实现高分辨的预测。

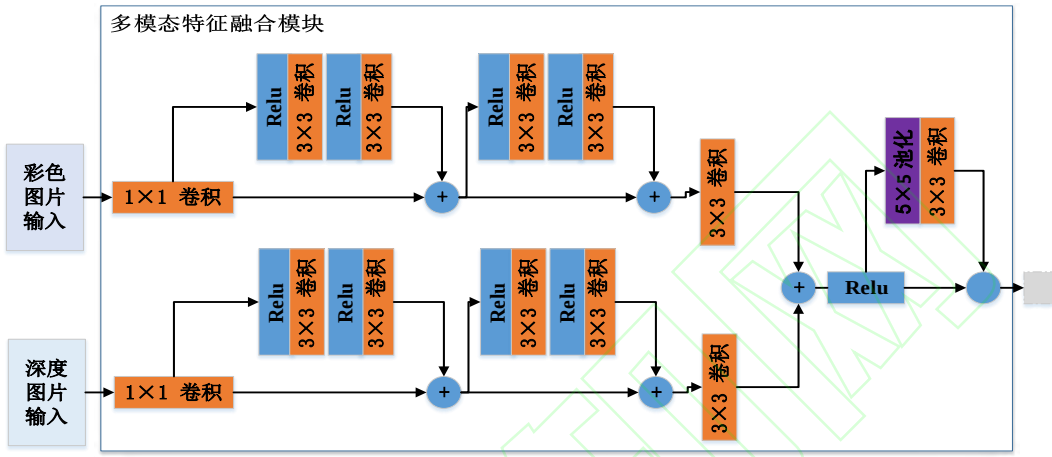

图 10 RDFNet 模型结构[60]

Fig.10 RDFNet model structure[60]

RDFNet 网络提出 MMF 模块对彩色图和深度图进行处理，通过考虑深度信息来实现更好的分割性能。目前，有效使用深度信息仍然是值得深入思考的问题，如 ACNET[63]、MCA-Net[64]等网络的提出，为 RGBD 语义分割在复杂场景下的应用提供了新的创新思路。

2.3.3 方法延展

通过 RedNet[59]和 RDFNet[60]可知，RGBD 模型的关键是彩色图和灰度图的有效融合。Fenwick 等人[65]提到 RGBD 模型可分为早期、中期和晚期融合，根据中期融合又细分为浅层中期融合和深层中期融合。然而早期融合与浅层中期融合的网络虽然在融合过程中很好的保留了空间线索，但是 RGDB 图像中的视觉信息和深度图中的几何信息在底层没有得到矫正，特征信息较少。而后期融合与深层中期融合，他们融合了表示语义信息的高级特征，在不同模式下得到的结果更加的兼容，但是两者互补的空间线索被大大削弱。因此，如文献[66,67]对深度图进行详细介绍，文献[68-71]通过对 RGBD 模型融合方式进行创新来提高分割精度。合理应用深度图对 RGB 图进行补充，一定会有效提高 RGBD 语义分割的分割精度。

3 图像语义分割实验分析与对比

3.1 数据集

根据全卷积语义分割方法应用场景的不同，整理了语义分割的常用公共数据集，分为 2D 数据集和 2.5D 数据集，如表 2 所示。

3.1.1 2D 数据集

PASCAL Visual Object Classes[72](简称 PASCAL VOC)：数据集由一个国际计算机挑战赛提供，从 2005 年一直发展到 2012 年，由于每年发布带标签的图像数据库并开展算法竞赛而产生一系列高质量的数据。目前数据集 PASCAL VOC 2012 最为常用。数据集包含 20 种类别（人、动物、交通工具、室内物品等），图片大小不固定，背景复杂多变。

PASCAL Context[73]：数据集由 PASCAL VOC 数据集扩展得到，总共有 540 个类，包含 10103 张语义标注的图像。该数据集类别繁多且许多类比较稀疏，因此在评估语义分割算法性能时，通常使用前 59 个类作为分割评判标准。

Semantic Boundaries Dataset[74](简称 SBD)：数据集由斯坦福大学建立，继承了 PASCAL VOC 中的 11355 张语义标注图像，其中训练集 8498 张图像，验证集 2857 张图像，图片大多数为户外场景类型，实际应用中已逐渐替代 PASCAL VOC 数据集。

Microsoft Common Objects in Context[75](简称 COCO)：数据集由微软公司开源和推广，包含 80 个图像实例，82782 张训练图片，40504 张验证图片和 81434 张测试图片，其中测试图片分为四类用于不同的测试。数据集中图像类别丰富，大多数取自复杂的日常场景，图中的物体具有精确的位置标注。

Cityscapes[76]：数据集由奔驰公司于 2015 年推行发布，专注于对城市街景的语义理解。提供了 50 个不同城市街景记录的立体视频序列，包含 20000 张弱注释图片和 5000 张的高质量的强注释的图片，涵盖了各种时间及天气变化下的街道动态物体，同时提供了 30 个类别标注，像素为 2048×1024 的高分辨率图像，

图像中街道背景信息复杂且待分割目标尺度较小。此数据集可用于实时语义分割研究。

CamVid[77]:数据集由剑桥大学的研究人员与2009年发布，CamVid 由车载摄像头拍摄得到的 5 个视频序列组成，提供了不同时段 701 张分辨率为 960×720 的图片和 32 个类别的像素级标签，包括汽车、行人、道路等。数据集中道路、天空、建筑物等尺度大，汽车、自行车、行人等尺度小，待分割物体尺度丰富。

KITTI[78]：目前国际上最大的用于自动驾驶场景的算法评测数据集，可进行 3D 物体检测、3D 跟踪、语义分割等多方面研究。数据集包含乡村、城市和高速公路采集的真实数据图像，原始数据集没有提供真实的语义标注，后来 Alvarez 等人[79,80]、Zhang 等人[81]和 Ros 等人[82]为其中部分图添加了语义标注。

表 2　常用分割数据集

Table 2　Popular segmentation datasets

分类	数据集名称	时间	应用场景	类别数量	数量总数	训练集数量	验证集数量	测试集数量	分辨率
2D 数据集	PASCAL-VOC 2012[72]	2012	多种场景	21	9993	1464	1449	1452	非固定
	PASCAL Context[73]	2014	多种场景	540	---	10103	10103	9637	非固定
	SBD[74]	2011	多种场景	21	---	8498	2857	---	非固定
	Microsoft COCO[75]	2014	多种场景	81	328000	82783	40504	81434	非固定
	Cityspaces[76]	2012	城市街道	30	5000	2975	500	1525	2048× 1024
	CamVid[77]	2009	城市街道	32	>700	367	100	233	960× 720
	KITTI-Layout[79][80]	2012	城市街道	3	---	328	---	---	非固定
	KITTI-Ros[81]	2015	城市街道	11	---	170	---	46	非固定
	KITTI-Zhang[82]	2015	城市街道	10	---	140	---	112	1226× 370
	SiftFlow[83]	2009	户外场景	33	---	2688	---	---	256× 256
	Standford background[84]	2009	户外场景	8	---	725	---	---	320×240
2.5D 数据集	NYUDv2[85]	2012	室内场景	40	407024	795	654	---	480× 640
	SUN3D[86]	2013	室内场景	---	---	19640	---	---	640× 480
	SUNRGBD[87]	2015	室内场景	37	10335	2666	2619	5050	非固定
	RGB-D Object Dataset[88]	2011	室内场景	51	---	207920	---	---	640× 480

Sift Flow[83]：数据集是 LabelMe 数据集的子集，包含 33 个类别和 2688 张分辨率为 256×256 的训练图像，提供 8 种不同户外场景，包括山脉、海滩、街道、城市等，图片都具有像素级标注。

Standford background[84]：数据集由斯坦福大学 2009 年发布，主要来自 LabelMe、MSRC、PASCAL VOC 等公共数据集。包含 715 张图片，分辨率为 320×240。包括道路、树木、草、水、建筑物、山脉、天空和前景物体共 8 个类别。

3.1.2　2.5D 数据集

NVUDv2[85]:数据集大都来自微软 Kinect 数据库，提供了 1449 个 RGBD 图像，捕获了 464 种不同的室内场景，并附有详细的标注，能够验证 3D 场景的提示和推断，实现更好的对象分割内场景，并附有详细的标注，能够验证 3D 场景的提示和推断，实现更好的对象分割。

SUN3D[86]：数据集由美国普林斯顿大学研究小组 2013 年发布，包含使用 Asus Xtion 传感器捕获的

415 个 RGBD 序列，是一个具有摄像机姿态和物体标签的大型 RGBD 视频数据库。每一帧均包含场景中物体的语义分割信息以及摄像机位态信息。

SUNRGBD[87]：数据集由 4 个 RGBD 传感器获取而得和 NYU depthv2、SUN3Dd 等数据集组成。包含了 10335 张室内场景、146617 个二维多边形标注、58657 个三维边界框标注以及大量的空间布局信息和种类信息。

RGB-D Object Dataset[88]：数据集由美国华盛顿大学的研究小组 2011 年发布，包括 11427 幅人工手动分割的 RGBD 图像组成，包含 300 个对象，分为 51 个类别。另外，还提供了 22 个带注释的自然场景视频序列，用于验证过程以评估性能。

表 3 不同语义分割方法在不同数据集上的性能

Table 3 Performance of different semantic segmentation methods on different datasets

分类	分割方法	时间	基础网络	MIoU(%)				
				VOC 2012	Cityscapes	CamVid	SUNRGBD	NYUDv2
经典语义分割方法	FCN[18]	2015	VGG16	62.2	--	--	--	34.0
	U-Net[28]	2015	VGG16	--	--	--	--	--
	SegNet[29]	2016	VGG16	--	--	60.1	22.57	--
	DeepLab v1[30]	2016	ResNet	71.6	--	--	--	--
	DeepLab v2[31]	2017	ResNet	79.7	70.4	--	--	--
	DeepLab v3[32]	2017	ResNet	86.9	81.3	--	--	--
	DeepLab v3+[33]	2018	ResNet	**89.0**	**82.1**	--	--	--
	RefineNet[39]	2017	ResNet	83.4	73.6	--	45.9	43.1
	PSPNet[40]	2017	ResNet	85.4	80.2	--	--	--
	GCN[41]	2017	ResNet	82.2	76.9	--	--	--
	DFN[42]	2018	ResNet	86.2	80.3	--	--	--
	FusionNet[43]	2016	VGG16	--	--	--	--	--
	DeconvNet[44]	2015	VGG16	72.5	--	--	--	--
实时性语义分割方法	ENet[49]	2016	--	--	58.3	51.3	26.3	--
	LinkNet[50]	2017	--	--	58.6	55.8	--	--
	BiseNet[51]	2018	Xception	--	68.4	**68.7**	--	--
	DFANet[52]	2019	Xception	--	**71.3**	64.7	--	--
	Light-weight RefineNet[53]	2018	ResNet	**81.1**	--	--	--	41.7
RGBD语义分割方法	RedNet[59]	2018	ResNet	--	--	--	**47.8**	--
	RDFNet[60]	2017	ResNet	--	--	--	47.7	**50.1**

3.2 性能评价指标

为了衡量分割算法的性能，需要使用客观评价指标来确保算法评价的公正性，运行时间、内存占用和精确度是常用的算法评价指标[89]。

3.2.1 运行时间

运行时间包括网络模型的训练时间和测试时间。由于运行时间依赖硬件设备及后台的实现，某种情况下，提供确切的运行时间比较困难。但是提供算法运行硬件的信息及运行时间有利于评估方法的有效性，以及保证相同环境下测试最快的执行方法。

3.2.2 内存占用

内存占用是分割方法的另一个重要的因素。图像处理单元（graphics processing unit,GPU）具有高效并行特征以及高内存带宽，但是相比于传统的中央处理器（cencer processing unit,CPU），时钟速度更慢以及处理分支运算的能力较弱。在某些情况下，对于操作系统及机器人平台，其显存资源相比高性能服务器并不具优势，即使是加速深度网路的 GPU，显存资源也相对有限。因此，在运行时间相同的情况下，记录算法运行状态下内存占用的极值和均值都是有意义的。

3.2.3 精确度

精确度包括像素精度（Pixel Accuracy,PA）、均像素精度（Mean Pixle Accuracy，MPA）、均交并比（MeanIntersection over Union,MIOU）、频率加权交并比(Frequency Weighted Intersection over Union，FWIoU)，常使用 MIoU 来衡量语义分割模型的性能。

像素准确度（PA）是语义分割中最简单的像素级评价指标，仅需计算机图像中正确分类的像素占图像中总像素比值，如公式（3）所示。

$$PA=\frac{\sum_{i=0}^{n}p_{ii}}{\sum_{i=0}^{n}\sum_{j=0}^{n}p_{ij}} \quad (3)$$

其中 p_{ii} 表示正确分类的像素个数，p_{ij} 表示本应属于第 i 类却被分为第 j 类的像素个数，n 是类别数。

平均准确度（MPA）表示图像中所有物体类别像素准确率的平均值，如公式（4）所示。

$$MPA=\frac{1}{n+1}\sum_{i=0}^{n}\frac{p_{ii}}{\sum_{j=0}^{n}p_{ij}} \quad (4)$$

平均交并比（MIoU）是分割结果真值的交集与其并集的比值，按类计算后取平均值，如公式（5）所示。

$$MIoU=\frac{1}{n+1}\sum_{i=0}^{n}\frac{p_{ii}}{\sum_{j=0}^{n}p_{ij}+\sum_{j=0}^{n}p_{ji}-p_{ii}} \quad (5)$$

频率加权交并比（FWIoU）是对 MIoU 改进后的新的评价标准，旨在对每个像素的类别按照其出现的频率进行加权，如公式（6）所示。

$$FWIoU=\frac{1}{\sum_{i=0}^{n}\sum_{j=0}^{n}p_{ij}}\sum_{i=0}^{n}\frac{\sum_{j=0}^{n}p_{ij}p_{ii}}{\sum_{j=0}^{n}p_{ij}+\sum_{j=0}^{n}p_{ji}-p_{ii}} \quad (6)$$

3.3 实验结果分析与对比

不同应用场景下语义分割方法在不同数据集上的实验结果对比如表 3 所示。选用分割领域标准数据集 VOC 2012、Cityscapes、CamVid、SUNRGBD 和 NYUDv2 对经典语义分割方法，实时性语义分割方法和 RGBD 语义分割方法进行实验结果分析和对比。

针对高精度追求的应用场景经典语义分割方法，多用于室外场景数据集，从表 3 可知，在 VOC 2012 数据集上 DeepLab v3+的精度高达 89.0%，在数据集 Cityscapes 是可达到 82.1%的精度；针对延时满足要求高这一应用场景，实时性语义分割网络 DFANet 和 Light-weight RefineNet 在数据集 Cityscapes 和 VOC 2012 分别达到 71.3%和 81.1%的准确率，并且后者每秒传输帧数需要 2055fps；而针对复杂场景下 RGBD 语义分割方法，在对室内复杂场景分割效果要优于经典语义分割和实时性语义分割的模型。

4 结束语

随着全卷积神经网络在图像语义分割领域的应用，如何提高分割精度成为目前研究的难点和痛点。本文从不同应用场景，针对不同场景下的经典网络结构展开分析总结，发现该领域仍然存在许多未知的问题值得深入探究。

（1）实时性语义分割

现阶段语义分割在实时性网络分割任务上，依旧不够完善，如何平衡语义分割精度和效率依旧是一个必不可少的研究方向。

（2）RGBD 语义分割

RGBD 网络模型目前的难点依旧是如何充分利用深度信息，有效的融合两者互补的模态，目前依旧是一个悬而未解的问题。

（3）三维场景的语义分割技术

深度图的引入让研究开始关注三维场景。尽管 3 维数据集难以获取，且标注工作很难，但是 3 维数据

集比 2 维数据集包含更多的图像语义信息，使得 3 维场景语义分割有较高的研究价值和广阔的应用前景。

（4）应用于视频数据的语义分割

可用的视频序列数据集较少，导致针对视频语义分割的研究进展缓慢。更多高质量的视频数据的获取和视频中时空序列特征的分析，将是语义分割领域的重要研究方向。

（5）弱监督和无监督语义分割技术

随着基于目标边框、基于图像类别便签、基于草图等弱监督方法的出现，降低了标注成本。但是分割效果并不理想，所以弱监督和无监督的语义分割需要进一步的研究。

参考文献：

[1] 汪海洋,潘德炉,夏德深.二维 Otsu 自适应阈值选取算法的快速实现[J].自动化学报,2007,33(9):968-971.

WANG H Y, PAN D L, XIA D S. A Fast Algorithm for Two-dimensional Otsu Adaptive Threshold Algorithm[J]. Journal of Image, 2007, 33(9):968-971.

[2] PUN T. A new method for gray-level picture thresholding using the entropy of the histogram[J]. Signal Processing, 1985, 2(3):223-237.

[3] OTSU N. A Threshold Selection Method from Gray-Level Histograms[J]. IEEE Transactions on Systems Man &CCybernetics, 2007, 9(1):62-66.

[4] YEN J C, CHANG F J, CHANG S . A new criterion for automatic multilevel thresholding[J]. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 1995, 4(3):370-378.

[5] DERICHE R. Using Canny's criteria to derive a recursively implemented optimal edge detector[J]. International Journal of Computer Vision, 1987, 1(2):167-187.

[6] ROSENFELD A. The Max Roberts Operator is a HueckelTypeEdge Detector[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1981, 3(1):101-3.

[7] YANG L, WU X Y, ZHAO D W, et al . An improved Prewitt algorithm for edge detection based on noisedimage[C]//International Congress on Image and Signal Processing. New York:IEEEPress, 2011:1197-1200.

[8] BOWYER K, KRANENBURG C, DOUGHERTY S. Edge Detector Evaluation Using Empirical ROC Curves[J]. Comput Vision & Image Understand, 2001, 84(1):77-103.

[9] COATES A , NG A Y. Learning Feature Representations with K-means[J]. Lecture Notes in Computer Science, 2012, 7700:561-580.

[10] CHENG Y. Mean shift, mode seeking, and clustering[J]. IEEETransactions on Pattern Analysis and Machine Intelligence, 1995, 17(8):790-799.

[11] FUKUNAGA K, HOSTETLER L. The estimation of the gradient of a density function, with applications in pattern recognition[J]. IEEE Transactions on Information Theory, 2006, 21(1):32-40.

[12] ACHANTA R, SHAJI A, SMITH K, et al. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(11):2274-2282.

[13] HAN S, TAO WD,WANG, et al. Image Segmentation Based on GrabCut Framework Integrating Multiscale Nonlinear Structure Tensor[J]. IEEE Transactions on Image Processing, 2009, 18(10):2289-2302.

[14] TANG M, GORELICK L, VEKSLERr O, et al. GrabCut in One Cut[J]. IEEE, 2013:1769-1776.

[15] BOYKOV Y Y. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images[C]// Proc Eighth IEEE International Conference on Comput Vis. IEEE Computer Society, 2001:105-112.

[16] ROTHER C. GrabCut: Interactive foreground extraction using iterated graph cuts[J]. Proceedings of Siggraph, 2004, 23(3):309-314.

[17] HINTON G E, SALAKHUTDINOV R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 313(5786):504-507.

[18] LONG J, SHELHAMER E, DARRELL T. Fully Convolutional Networks for Semantic Segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[19] GARCIA-GARCIA A, ORTS-ESCOLANO S, OPREA S, et al. A Review on Deep Learning Techniques Applied to Semantic Segmentation[J].arXiv:1704.06857,2017.

[20] 黄鹏,郑淇,梁超.图像分割方法综述[J].武汉大学学报(理学版),2020,66(6):519-531.

HUANG P, ZHENG Q, LIANG C. Overview of ImageSegmentation Methods[J]. Wuhan University Journal of Natural Sciences (college of science), 2020,66(6):519-531.

[21] 田萱，王亮，丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报, 2019, Vol.30Issue(2):440-468.

TIAN X, WANG L, DING Q. Review of image semantic segmentation based on deep learning[J]. Journal of Software, 2019, 30(2): 440-468.

[22] 章琳,袁非牛,张文睿,曾夏玲.全卷积神经网络研究综述[J].计算机工程与应用,2020,56(1):25-37.

ZHANG L, YUAN F N, ZHANG W R, ZENG X L. Review of FullyConvolutional Neural Network[J]. Computer Science and Application,2020, 56(1):25-37.

[23] 徐辉，祝玉华，甄彤，李智慧. 深度神经网络图像语义分割方法综述[J]. 计算机科学与探索，2021，15(1):47-59.

XU H, ZHU Y H, ZHEN T, LI Z H.Survey of Image

Semantic Segmentation Methods Based on Deep Neural Network[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(1): 47-59.

[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2):1097-1105.

[25] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J].arXiv: 1409.1556, 2014.

[26] SZEGEDY C, LIU W, JIA Y, et al.Going Deeper with Convolutions[J]. IEEE Computer Society.arXiv:1409.4842, 2014.

[27] HE K, ZHANG X, REN S, et al.Deep residual learning for image recognition[C]//Proceedings of International Conference on Computer Vision and Pattern Recognition, 2016:770-778.

[28] RONNEBERGER O, FISCHER P, BROX T.U-net: Convolutional networks for biomedical imagesegmentation[C]//Proceedingsof the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015:234-241.

[29] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017:1-1.

[30] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4):357-361.

[31] CHEN L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.

[32] CHEN L C, PAPANDREOU G, SCHROFF F, et al.Rethinking atrous convolution for semantic image segmentation[J]. arXiv:1706.05587, 2017.

[33] CHEN L C, ZHU Y, PPAPNDRROU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision, 2018:801-818.

[34] ZHOU Z , SIDDIQUEE M , TAJBAKHSH N , et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation[C]// 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop.arXiv: 1807.10165, 2018.

[35] HUANG H , LIN L , TONG R , et al. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. arXiv:2004.08790, 2020.

[36] OKTAY O , SCHLEMPER J , FOLGOC L L , et al. Attention U-Net: Learning Where to Look for the Pancreas[J].arXiv: 1804.03999, 2018.

[37] 孟俊熙,张莉,曹洋,张乐天,宋倩.基于 Deeplab v3+的图像语义分割算法优化研究[J/OL].激光与光电子学进展, 2021:1-15..

MENG J X, ZHANG L, CAO Y, et al. Research on optimization of image semantic segmentation algorithms based on Deeplab v3+[J/OL].Laser & Optoelectronics Progress, 2021:1-15.

[38] 赵小强,徐慧萍.分级特征融合的图像语义分割[J].计算机科学与探索,2021,15(5):949-957.

ZHAO X Q, XU H P, Image Semantic Segmentation Method with Hierarchical Feature Fusion[J].Journal of Frontiers of Computer Science and Technology, 2021, 15(5):949-957.

[39] LIN G, MILAN A, SHEN C, et al.Refinenet:Multi-path refinement networks for high- resolution semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1925-1934.

[40] ZHAO H, SHI J, QI X, et al.Pyramid scene parsing network[C]//Proceedings of International Conference on Computer Vision and Pattern Recognition, 2017:6230-6239.

[41] PENG C, ZHANG X, YU G, et al.Large kernel matters-improve semantic segmentation by global convolutional network[C]//Proceedings of International Conference on Computer Vision and Pattern Recognition, 2017:1743-1751.

[42] YU C, WANG J, PENG C, et al. Learning a Discriminative Feature Network for Semantic Segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.arXiv:1804.09337, 2018.

[43] QUAN T M, HILDEBRAND D, JEONG W K. FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics[J]. arXiv:1614.05360, 2016.

[44] NOH H, HONG S, HAN B.Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1520-1528.

[45] GADDE R, JAMPANI V, GEHLER P V. Semantic Video CNNs Through Representation Warping[J]. IEEE, arXiv: 1708.03088, 2017.

[46] NILSSON D, SMINCHISESCU C. Semantic Video Segmentation by Gated Recurrent Flow Propagation[J]. arXiv:1612.08871,2016.

[47] JIN X, LI X, XIAO H, et al. Video Scene Parsing with Predictive Feature Learning[J].arXiv:1612.00119,2016.

[48] RADFORD A, METZ L, CHINTALA S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[C]. arXiv:1511.06434,

2015.

[49] PASZKE A , CHAURASIA A , KIM S , et al. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation[J].arXiv:1606.02147,2016.

[50] CHAURASIA A, CULURCIELLO E. Linknet: Exploiting encoder represen-tations for efficient semantic segmentation[C]//Proceedings of the IEEE Visual Communications and Image Processing.Petersburg: IEEE, 2017:1-4.

[51] YU C, WANG J, PENG C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation[M].Berlin, Germany: Springer, 2018:334-349.

[52] LIH, XIONG P, FAN H, et al. DFANet: deep feature aggregation for real-time semantic segmentation[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Washington D.C. , USA: IEEE Press, 2019:9522-9531.

[53] FAN M, LAI S, HUANG J, et al. Rethinking BiSeNet For Real-time Semantic Segmentation.arXiv:2104. 13188,2021.

[54] LYU H, FU H, HU X, et al. Esnet: Edge-Based Segmentation Network for Real-Time Semantic Segmentation in Traffic Scenes[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019:1855-1859.

[55] FANG Q , QIU J , WU H , et al. DFPNet:Dislocation Double Feature Pyramid Real-time Semantic Segmentation Network[C]//2020 Chinese Automation Congress (CAC). 2020:2587-2592.

[56] VNEKRASOV, SHEN C, REID I. Light-Weight Refine-Net for Real-Time Semantic Segmentation[J]. arXiv:1810. 03272, 2018.

[57] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network[J]. Computer Science, 2015, 14(7):38-39.

[58] ELSKEN T, METZEN J H, F HUTTER. Neural Architecture Search: A Survey[J]. arXiv:1808.05377, 2018.

[59] JIANG J, ZHENG L, LUO F, et al. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation[J]. arXiv:1806.01054,2018.

[60] LEE S, PARK S J, HONG K S. RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation[C]//IEEE International Conference on Computer Vision. IEEE, 2017:4980-4989.

[61] XING Y, WANG J, CHEN X, et al. Coupling Two-Stream RGB-D Semantic Segmentation Network by Idempotent Mappings[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019:1850-1854.

[62] XING Y, WANG J,CHEN X,et al. 2.5D Convolution for RGB-D Semantic Segmentation [C]//2019 IEEE International Conference on Image Processing (ICIP), 2019: 1410-1414.

[63] HU X, YANG K, FEI L, et al. ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation[J]. IEEE, 2019：1440-1444.

[64] SHI W, ZHU D, ZHANG G, et al. Multilevel Cross-Aware RGBD Semantic Segmentation of Indoor Environments[C]//2019 IEEE International Conference on Cyborg and Bionic Systems (CBS). IEEE, 2019:382-390.

[65] LI Y, ZHANG J, CHENG Y, et al. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2018:1262-1266.

[66] EIGEN D, FERGUS R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture[C]//2015 IEEE International Conference on Computer Vision (ICCV). arXiv:1411.4734, IEEE, 2014.

[67] MANCINI M, COSTANTE G, VALIGIi P, et al. Fast Robust Monocular Depth Estimation for Obstacle Detection with Fully Convolutional Networks[J]. IEEE, arXiv: 1607.06349,2016.

[68] HU X, YANG K, FEI L, et al.ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation[J]. IEEE, arXiv:1905.10089, 2019.

[69] CHEN S,ZHU X, LIU W, et al. Global-Local Propagation Network for RGB-D Semantic Segmentation[J]. arXiv: 2101.10801, 2021.

[70] CHEN X, LIN K Y, WANG J, et al. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation[J]. arXiv:2007.09183, 2020.

[71] QI X, LIAO R, JIA J, et al. 3D Graph Neural Networks for RGBD Semantic Segmentation[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017:2380-7504.

[72] EVERINGHAM M, ESLAMI S, GOOL L V, et al. The Pascal Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[73] MOTTAGHI R，CHEN X，LIU X，et al.The role of context for object detection and semantic segmentation in the wild[C]//Proceedings of IEEE Conference on Compu- ter Vision and Pattern Recognition, 2014:891-898.

[74] CHEN X, MOTTAGHIR, LIU X, et al. Detect what you can: detecting and representing objects using holistic models and body parts[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recogni- tion. Washington D.C.,USA: IEEE Press,2014:1971-1978.

[75] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[A].Proceedings of the European Conference on Computer Vision[C].USA:IEEE, 2014:740-755.

[76] CORDTS M, OMRAN M, RAMOUS S, et al.The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition, 2016:3213-3223.
[77] BROSTOW G J, SHOTTON J, FAUQUEUR J, et al. Segmentation and recognition using structure from motion point clouds[C]//Proceedings of the European Conference on Computer Vision, 2008:44-57.
[78] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research,2013,32(11):1231-1237.
[79] ALVAREZ J M, GEVERS T, LECUN Y, et al. Road Scene Segmentation from a Single Image[C]//European Conference on Computer Vision. Springer-Verlag, 2012: 376-389.
[80] G ROS†, JM ALVAREZ‡. Unsupervised image transformation for outdoor semantic labelling[C]//2015 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2015:537-542.
[81] ZHANG R, CANDRA S A, KAI V, et al. Sensor fusion for semantic segmentation of urban scenes[C]//IEEE International Conference on Robotics & Automation. IEEE, 2015: 1850-1857.
[82] ROS G, RAMOS S, GRANADOS M, et al. Vision-Based Offline-Online Perception Paradigm for Autonomous Driving[C]//2015 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2015:231-238.
[83] LIU C, YUEN J, TORRALBA A. Nonparametric scene parsing: Label transfer via dense scene alignment[J]. Proceedings/CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009:1972-1979.
[84] GOULD S, FULTON R, KOLLER D. Decomposing a scene into geometric and semantically consistent regions[C]//IEEE International Conference on Computer Vision. IEEE, 2009:1-8.
[85] SILBERMAN N, HOIEM D, KOHLI P, et al.Indoor segmentation and support inference from RGBD images[C]//Proceedings of European Conference on Computer Vision, 2012:746-760.
[86] XIAO J, OWENS A H, TORRALBA A. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels[C]//2013 IEEE International Conference on Computer Vision (ICCV). IEEE, 2013:1625-1632.
[87] SONG S, LICHTENBERG S P, XIAO J. SUN RGB-D: A RGB-D scene understanding benchmark suite[C]//IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2015:567-576.
[88] LAI K, BO L, REN X, et al. A large-scale hierarchical multi-view RGB-D object dataset[C]//IEEE International Conference on Robotics & Automation. IEEE, 2011:1817-1824.
[89] GARCIA-GARCIA A, ORTS-ESCOLANO S, OPREA S, et al. A review on deep learning techniques applied to semantic segmentation[J].arXiv: 1704.06857, 2017.