基于深度学习的语义分割问题研究综述

张祥甫，刘健*，石章松，吴中红，王智
海军工程大学兵器工程学院，湖北 武汉 430032

摘要 语义分割是计算机视觉领域的核心技术，通过对图像中的每个像素点进行分类，将图像分割成若干个具有特定语义类别的区域。近年来，卷积神经网络(CNN)不断取得突破性进展，利用深度学习方法处理语义分割问题展示出具大的潜力。首先从语义分割的定义出发，探讨了目前语义分割领域存在的挑战。在介绍 CNN 相关原理的基础上，详细对比了几种用于语义分割算法评测的数据集，并重点对近年来语义分割领域基于解码器、信息融合和循环神经网络的深度学习方法进行综述。最后进行总结和展望，阐述了未来语义分割领域在进一步丰富数据库场景、提高算法实时性和开展三维点云语义分割三方面的发展趋势。

关键词 图像处理；语义分割；深度学习；卷积神经网络；特征融合
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.150003

Review of Deep Learning-Based Semantic Segmentation

Zhang Xiangfu, Liu Jian*, Shi Zhangsong, Wu Zhonghong, Wang Zhi
College of Weapons Engineering, Naval University of Engineering, Wuhan, Hubei 430032, *China*

Abstract Semantic segmentation, which classifies all pixels in an image and divides the image into several regions with specific semantic categories, is a key technology in the field of computer vision. In recent years, convolutional neural networks (CNNs) have been making breakthroughs and have demonstrated great potential in using deep learning to perform semantic segmentation. Herein, beginning with the definition of semantic segmentation, existing challenges in the field of semantic segmentation are discussed. Based on CNN principles, several datasets used for semantic segmentation algorithm evaluation are compared in detail, and recent deep learning methods based on decoders, information fusion, and recurrent neural networks in semantic segmentation are summarized. Finally, future development trends (e.g. enriching database scenes, improving real-time performance of algorithms, and researching the semantic segmentation) of three-dimensional point cloud data in semantic segmentation are summarized.

Key words image processing; semantic segmentation; deep learning; convolutional neural network; feature fusion
OCIS codes 100.3008; 100.5010; 100.2000

1 引 言

语义分割是计算机视觉和模式识别领域的关键技术，与物体检测、图像分类共同组成环境感知的三大核心任务。语义分割融合了传统的图像分割和目标识别两大技术，其目的是将图像分割成几组具有特定语义类别的区域，属于像素级别的密集分类问题。语义分割是自动驾驶[1-3]、医学影像分割[4]、行人检测[5]等领域的技术基础，具有广泛的应用价值。例如：在穿衣搭配领域[6]，语义分割可以区分出头部、上身、下身等人体部位，为用户推荐合理的穿搭风格；在医学影像分割领域[4]，通过语义分割可以完成人体器官部位的提取与测量，从而辅助医生诊断与治疗。由此可见，语义分割技术具有广阔的应用前景，对语义分割开展相关研究具有重要意义。

然而，图像语义分割是计算机视觉领域非常具有挑战性的问题，分割效果直接影响后续的图像处理任务。语义分割既要克服同类物体因光线、角度

收稿日期：2019-01-25；**修回日期**：2019-02-24；**录用日期**：2019-03-05
基金项目：国家自然科学基金(61773395)
* **E-mail**：liujian_nue@163.com

和状态等不同而产生的差异性，还要解决不同类物体之间的高度相似性。此外，语义分割的实际场景往往是复杂多样的，物体间常伴随有遮挡、割裂现象，这进一步增加了语义分割的难度。

作为计算机视觉领域最成功的一种深度学习模型，卷积神经网络（CNN）近年来取得了突破性进展[7-8]。2017 年，Krizhevsky 等[9] 使用著名的 AlexNet 模型在 ImageNet 图像分类挑战赛上以领先第二名传统方法 10%的准确率夺得冠军，使得深度学习再次受到广泛关注。此后，以 CNN 为基础架构的深度学习算法[10-13] 不断在图像分类与识别领域的大规模竞赛中取得突破，并广泛应用于图像分类、语音识别、机器翻译等领域，它的识别精度在部分领域甚至超过了人工识别精度。因此，设计深度学习模型处理语义分割问题具有很大的潜力，且已取得重要进展。本文对基于 CNN 的语义分割方法进行综述，首先介绍了 CNN 的相关知识，然后对现有的数据库进行详细介绍，重点对基于深度学习的语义分割方法进行全面综述，最后对其未来发展趋势进行展望。

2 CNN

在深度学习发展进程中，CNN 的地位不言而喻，对计算机视觉发展做出了巨大贡献。现有的 CNN 语义分割方法主要是依据语义分割任务的特点，将图像分类基础网络进行合理改进，从而实现较好的语义分割效果。为此，首先对 CNN 的基础内容进行介绍，然后以大规模视觉图像识别竞赛（ILSVRC）为主要脉络，对近年来表现优异的 CNN 基础网络进行梳理和总结。

2.1 网络基础

在 CNN 出现之前，传统的图像算法一般都采用人工设计的多种特征相结合的方式[14-17] 处理计算机视觉任务，不仅效率特别低，而且精度也很差。对于不同的任务，传统的图像算法还需要不断地组合已有特征，以实现最优的效果，未能实现端到端的学习效果。传统的神经网络由于其自身结构的局限性，难以构造深层次网络，分类效果一般，与人工特征相比优势并不明显。CNN 凭借其强大的拟合能力，通过卷积学习大量数据，能提取到最大限度区分物体的抽象特征，精度远超传统算法。

在本质上，CNN 其实是一种特殊的前馈神经网络或多层感知器。标准的 CNN 基础层结构主要包括卷积层（Conv）、池化层和全连接层，通过连接这些层结构，组成一个完整的 CNN 结构，如图 1 所示。

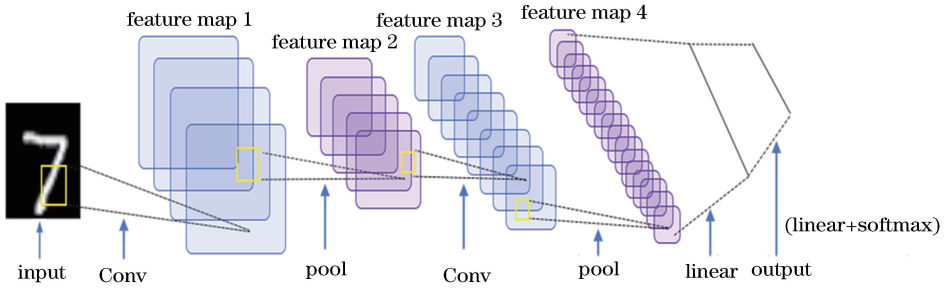

图 1 标准 CNN 基本组成示意图

Fig. 1 Diagram of basic composition of standard convolutional neural network

2.1.1 卷积层

卷积层是 CNN 的核心部分，网络中的大部分计算都在该层进行。每个卷积层都包含一系列固定尺寸的卷积核（也叫滤波器），在训练初始化阶段主要是对这些滤波器的权重和偏置等参数进行初始化。卷积操作的核心思想是将滤波器按照一定的步长，沿着输入数据的宽度和高度滑动，经过卷积和激活函数作用后生成激活图，激活图中的每个位置都记录了输入数据对于该卷积核的反应，将这些二维（2D）激活图在深度方向层叠起来就产生了卷积层的输出。

2.1.2 池化层

卷积层之后通常会紧跟着一个池化层，该层也叫降采样层，通过对输入特征图中邻近区域的聚合统计，能够降低数据体的空间尺寸，减少网络中参数的数量，减少计算资源耗费，同时也能够有效地控制过拟合现象，改善分类器的性能。池化层一般采用尺寸为 2×2 的窗口对图像进行下采样，将其中大部分的激活信息都丢弃，保留特征最大值，具体处理过程如图 2 所示。常用的池化操作还有平均池化、L2 范数池化。实践表明，在卷积层中采用最大值池化的效果最好，平均池化通常放在卷积网络的最后一层。

2.1.3 全连接层

在卷积网络的最后，通常会有一两个全连接层，全连接层的每个神经元同上一层所有的神经元相

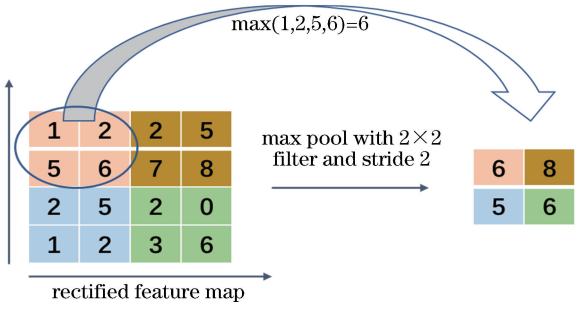

图 2 最大值池化处理过程示意图

Fig. 2 Diagram of max pooling process

连。全连接层与传统神经网络中的结构是一样的，把经过系列卷积和池化层输出的 2D 特征图转化为一维向量，再经过几个隐藏层后得到最终的输出结果。研究发现，在这个过程中引入丢失输出(dropout)和在全连接层前使用全局平均池化能够有效抑制过拟合现象。

2.2 常用网络

常见的图像分类网络主要有 AlexNet、VGGNet、GoogLeNet 和 ResNet 等，基本结构通常包括输入层、交替的卷积和池化层、全连接层和输出层。表 1 给出了这些网络的信息综述，直观地展示了各网络的提出时间、模型结构、主要操作及识别结果等内容。

表 1 常见图像分类网络信息汇总

Table 1 Information summary of common image classification networks

Item	LeNet5	AlexNet	VGGNet	GoogLeNet	ResNet
Year	1994	2012	2014	2014	2015
Layer	7	8	19	22	152
Conv	2	5	16	21	151
Kernel size	5	11,5,3	3	7,1,3,5	7,1,3,5
Linear	3	3	3	1	1
Linear size	120,84,10	4096,4096,1000	4096,4096,1000	1000	1000
Activation function	Sigmoid	ReLU	ReLU	ReLU	ReLU
Classifier	Multi-layer perception	Softmax	Softmax	Softmax	Softmax
Data augment	×	√	√	√	√
Bath normalization	×	×	×	×	√
Local response normalization	×	√	×	√	×
Graphics processing unit	×	√	√	√	√
Inception	×	×	×	√	×
Dropout	×	√	√	√	√
TOP-5(error)	N/A	16.4%	7.32%	6.67%	3.57%

AlexNet[9]是深度学习发展进程中的突破性成果，使神经网络再次处于人工智能领域的风口浪尖。它使用两个图形处理器(GPU)和校正线性函数(ReLU)，大大提高了 CNN 的学习训练速度，并在 2012 年的 ILSVRC 中以领先第二名 10%的准确率夺得冠军。AlexNet 主要由输入层、5 个卷积层和 3 个全连接层构成，其中 3 个卷积层还进行了最大池化。此外，AlexNet 还使用了数据增强和 dropout 的训练技巧，有效地抑制了过拟合现象。

VGGNet[10]是 CNN 的突破模型。AlexNet 的出色表现说明，通过增加网络的深度可以提升网络的性能。牛津大学的 Visual Geometry Group 沿着逐步加深卷积网络结构这一思路建立起 VGGNet 系列模型 VGGNet-16 和 VGGNet-19，并凭此获得了 ImageNet 2014 年的亚军，其模型架构如图 3 所示。与 AlexNet 相比，VGGNet 使用了很多小的滤波器，在达到与 1 个大的滤波器相同感受野效果的同时，还能减少训练参数，增大网络的深度，使决策函数区分度变强，模型也更容易训练。由此可见，VGGNet 主要是对网络层进行不断地层叠，并没有太大的创新，但也表明增加网络深度确实可以在一定程度上改善模型效果。

GoogLeNet 是由牛津大学的 Szegedy 等[11]提出的，在 ILSVRC-2014 竞赛上以 top-5 上 93.3%的准确率取得冠军，如今已经发展到 V4 版本。与 VGGNet 相比，GoogLeNet 采取了更深的网络结构，共有 22 层。同时，GoogLeNet 通过引入新颖的网络结构——Inception 模块提高了计算效率。

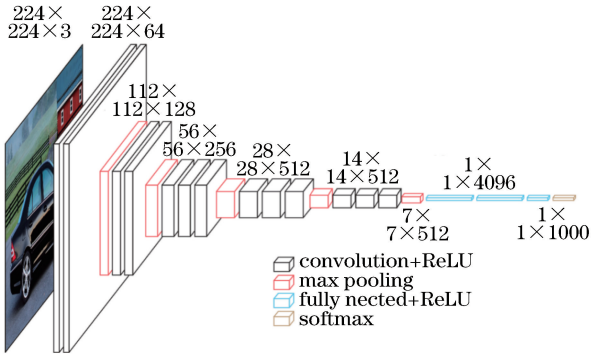

图 3 VGGNet 模型的结构示意图

Fig. 3 Structural diagram of VGGNet model

Inception 模块可以看作是一个局部的网络拓扑结构，如图 4 所示，它的核心处理层包括 4 个通道：1×1 卷积通道、3×3 卷积通道、5×5 卷积通道和 3×3 最大池化通道。该结果能够使用多个具有不同感受野的滤波器对输入进行卷积和池化，充分学习不同尺度的特征信息，从而提高网络性能。

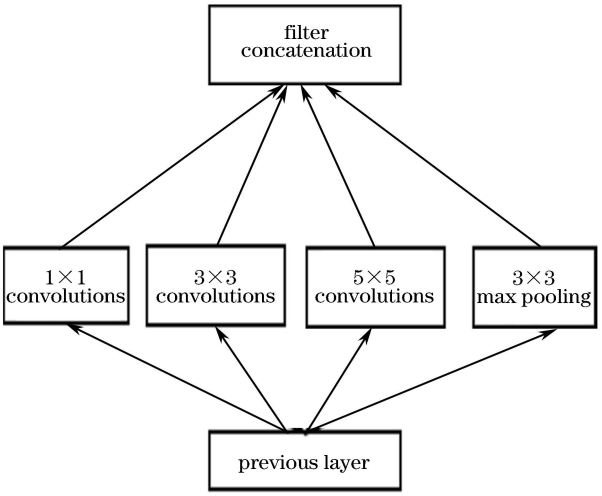

图 4 GoogLeNet 网络中的 Inception 模块

Fig. 4 Inception module in GoogLeNet network

微软研究院提出的 ResNet[12] 在 ILSVRC-2015 竞赛中以 96.4% 的准确率获得冠军。与 GoogLeNet 相比，ResNet 将网络的深度扩展到 152 层。另外，该网络构造了残差模块(图 5)，通过在输入与输出间加入捷径连接，保证了下一层能够从输入中学到与已学习信息不一样的内容，从而解决了训练更深层网络时普遍存在的退化问题。图 5 中，X 是该层残差块的输入，$F(X)$为激活后的输出。

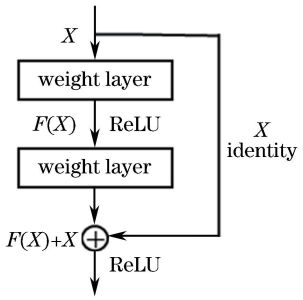

图 5 ResNet 网络中的残差模块

Fig. 5 Residual module in ResNet network

3 常见数据库

对于开展基于深度学习的语义分割算法研究而言，一个庞大且有效的数据集是极为关键的。对于相同的神经网络，数据量的增加往往能够提升语义分割的效果。但是，创建一个数据量足够大且能准确反映相关场景的数据集，不仅会耗费大量的时间与精力，而且需要一定的专业知识和设备来获取数据。为促进语义分割及场景分类问题的研究，许多科研机构提供了能够代表语义分割场景的开放数据集，这些数据集不仅能用来测试相关算法，还给不同算法之间的性能对比提供了可靠的平台。

本节主要从数据集类别、各类型图像数量、用途及年份等角度出发，详细介绍语义分割领域主流的数据集，并分析其在语义分割领域表现出的优缺点，具体信息如表 2 所示。图 6 为部分来自这些常见数据集的图像及其对应的语义分割示意图，其中不同颜色的像素区域表示不同的语义信息。

1) PASCAL VOC 2012。这是 PASCAL VOC[18] 为图像分类和语义分割等任务发布的一套数据集，提供了 1464 张具有标签的训练图像和 1449 张验证图像。图像主要是常见的生活物体，总共划分为 21 类，包括人、动物、植物和交通工具等。如果某像素不属于任何类，则将其视为背景。数据集属于多标签类型，每幅图像都包含了一个或多个类别物体。2011 年，文献[19]在此基础上重新标注了约 10000 幅图像，将训练图像的数量提升到 10582，被称为 PASCAL VOC 2012+。由于图像背景复杂、物品尺度变化大，且不同物体间存在遮挡，该数据集对语义分割算法提出了较大挑战。

2) PASCAL-CONTEXT。文献[20]在 PASCAL VOC 2010 物体识别竞赛数据集的基础上，对全部训练集图像进行像素级别标注。包括 540 个语义类别的语义分割图像，除之前 20 类物体外，还标注了草地、天空和道路等场景信息。该数据集种类繁多，在使用该数据集测试语义分割算法时，通常只采用出现频率最高的 59 个类，其他类则重新标定为背景。与 PASCAL VOC 2012 数据

表 2 常见语义分割数据集信息汇总

Table 2 Information summary of common semantic segmentation datasets

Dataset	Classes	Sample(training)	Sample(validation)	Sample(test)	Purpose	Year
PASCAL VOC 2012[18]	21	1464	1449	1452	Generic	2012
PASCAL VOC 2012+[19]	21	10582	1449	1452	Generic	2014
PASCAL-CONTEXT[20]	540	4998	5105	—	Generic	2014
PASCAL-PERSON-PART[20]	6	1716	—	1817	Person	2014
PASCAL-COW-PART[21]	4	294	—	227	Cow	2015
SBD[22]	21	8498	2857	—	Generic	2011
MICROSOFT COCO[23]	80+	82783	40504	81434	Generic	2014
CITYSCAPES(fine)[24]	19	2975	500	1525	Urban	2015
CITYSCAPES(coarse)[24]	19	22973	500	—	Urban	2015
CAMVID[25-26]	32	361	100	233	Driving	2009
KITTI-Ros[27]	11	170	—	46	Driving	2015
KITTI-Zhang[28]	10	140	—	112	Driving	2015

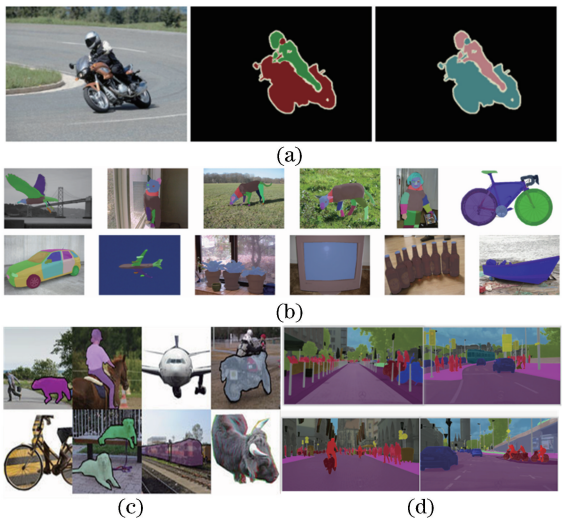

图 6 部分数据集图像及其对应的语义分割效果图。(a) PASCAL VOC 2012; (b) PASCAL-CONTEXT; (c) MICROSOFT COCO; (d) CITYSCAPES

Fig. 6 Partial dataset images and corresponding semantic segmentation effect diagrams. (a) PASCAL VOC 2012; (b) PASCAL-CONTEXT; (c) MICROSOFT COCO; (d) CITYSCAPES

集相比,该数据集包含了更多的场景信息,语义分割的难度进一步加大。PASCAL-PERSON-PART 在该数据集的基础上,为人体各个部位提供了精细的语义标注,通常包括 6 类,即头部、躯干、上臂、下臂、大腿和小腿。文献[21]提出的 PASCAL-COW-PART 标注了 531 张包含牛和马的图像,主要分为 4 个类别:头部、身体、四肢和尾巴。这两个数据集在语义分割时面临的难点主要来自物体的尺度和姿态。

3) SEMANTIC BOUNDARIES DATASET (SBD)。作为 PASCAL VOC 数据集的扩展[22],该数据集对其未标注图像进行了语义标注,提供 PASCAL VOC 2011 数据集中 11355 张图像的标签,但训练集与验证集的划分形式与以往不同,训练集图像有 8498 张,验证集图像有 2857 张。除各物体的边界信息外,该数据集还包含相应的类别及实例信息。近年来,该数据集的训练数据不断增多,实践中逐渐采用该数据集取代 PASCAL VOC 数据集。

4) MICROSOFT COCO。它是 2014 年建立的一个大规模图像识别与处理数据集[23],包含 80 多个类别的物体,提供 82783 张训练图像、40504 张验证图像和 80000 多张测试图像。其中,测试集图像分为 4 个子集,每个子集包含 20000 张图像:test-standard 是默认测试数据子集,用于性能对比;test-challenge 是竞赛专用子集;test-reserve 是防作弊检验子集;test-dev 通常用于额外验证及调试。该数据集规模庞大,内容丰富,一般用于 CNN 预训练,有助于提高深度学习算法的性能。

5) CITYSCAPES。一个专注于城市街道场景的大规模数据集[24],通常包含 19 种语义类别,包括 3475 张精确标注图像和 20000 多张模糊标注图像。数据从 50 个城市场景中采集而来,以视频格式存储,包含了不同的天气和时间段,具有动态信息丰富、场景布局多样和街道背景复杂等特点。因此,该数据集也对语义分割算法提出了挑战:一方面,该数据集的复杂背景和尺度差异对算法精度提出了挑战;另一方面,较大的图像尺寸(2048 pixel × 1024 pixel)对自动驾驶领域追求的实时性也提出了挑战。

6) CAMVID。它是由 Brostow 在 2009 年发布的道路场景数据集[25],采用 960 pixel×720 pixel 的车载摄相机拍摄的 5 个视频流,包含 32 类物体共 701 幅图像。Sturgess 等[26]将该数据集进一步划分为 367 张训练图像、100 张验证集图像和 233 张测试集图像,但只使用了 11 类标签。

7) KITTI。它是近些年在智能机器人和无人驾驶领域广受欢迎的数据集之一[29],包含了由多种类型传感器(主要有 RGB 相机、深度摄像机和激光扫描仪)采集到的交通场景信息。该数据集并没有提供完整的语义标注。Alvarez 等[30]和 Ros 等[31]为道路检测比赛中的 323 幅图像贴上了语义标签,主要包括天空、道路和垂直面 3 大类;Zhang 等[28]在跟踪竞赛数据中选择了 252 幅图像进行标注,包括 140 张训练图像和 112 张测试图像,物体分为 10 类;Ros 等[27]在视觉测距数据集中标注了 170 张训练图像和 46 张测试图像,促使其在相应领域发挥实际作用。

4 基于深度学习的语义分割方法

近年来,由深度学习引发的技术革命使计算机视觉领域发生了翻天覆地的变化,包括语义分割在内的许多计算机视觉问题都开始使用深度网络架构来解决。本节对现有的深度学习语义分割方法进行综述,并将其分为 3 类:基于解码器的方法、基于信息融合的方法和基于循环神经网络(RNN)的方法。表 3 对这些方法进行了详细的总结,展示了所提方法产生的时间、采用的基础架构、主要贡献,以及完成相应任务的能力。其中,能力评价分为 3 个等级:A 代表优秀,B 表示良好,C 表示合格。此外,图 7 给出了这些方法的形象化展示。

表 3 常见深度学习语义分割方法的信息汇总

Table 3 Information summary of common deep learning semantic segmentation methods

Model name	Year	Architecture	Accuracy	Efficiency	Training	Contribution
FCN[32]	2015	VGG-16(FCN)	C	C	C	Forerunner
SegNet[33]	2017	VGG-16 + Decoder	A	B	C	Encoder-decoder
DeepLab[34-37]	2017	VGG-16 + ResNet-101	A	C	C	Standalone CRF, Atrous convolutions
CRFasRNN[38]	2015	FCN-8s	C	B	A	CRF reformulated as RNN
ParseNet[39]	2015	VGG-16	A	C	C	Global context feature fusion
SharpMask [40]	2016	DeepMask	A	C	C	Top-down refinement module
PSPNet[41]	2016	ResNet-101	A	B	C	Pyramid pooling module
Multi-scale-CNN-Raj[42]	2015	VGG-16(FCN)	A	C	C	Multi-scale architecture
Multi-scale-CNN-Eigen[43]	2015	Custom	A	C	C	Multi-scale sequential refinement
Multi-scale-CNN-Roy[44]	2016	Multi-scale-CNN-Eigen	A	C	C	Multi-scale coarse-to-fine refinement
Multi-scale-CNN-Bian[45]	2016	FCN	B	C	B	Independently trained Multi-scale FCNs
ReSeg[46]	2016	VGG-16 + ReNet	B	C	C	Extension of ReNet to semantic segmentation
LSTM-CF[47]	2016	Fast R-CNN +DeepMask	A	C	C	Fusion of contextualinformation from multiple sources
RCNN[48]	2014	MDRNN	A	B	C	Different input sizes, image context
2D-LSTM[49]	2015	MDRNN	B	B	C	Image context modelling
DAG-RNN [50]	2015	Elman network	A	C	C	Graph image structure for context modelling
MINC-CNN[51]	2015	GoogLeNet(FCN)	C	C	C	Patchwise CNN, Standalone CRF
DeepMask[52]	2015	VGG-A	A	C	C	Proposals generation for segmentation

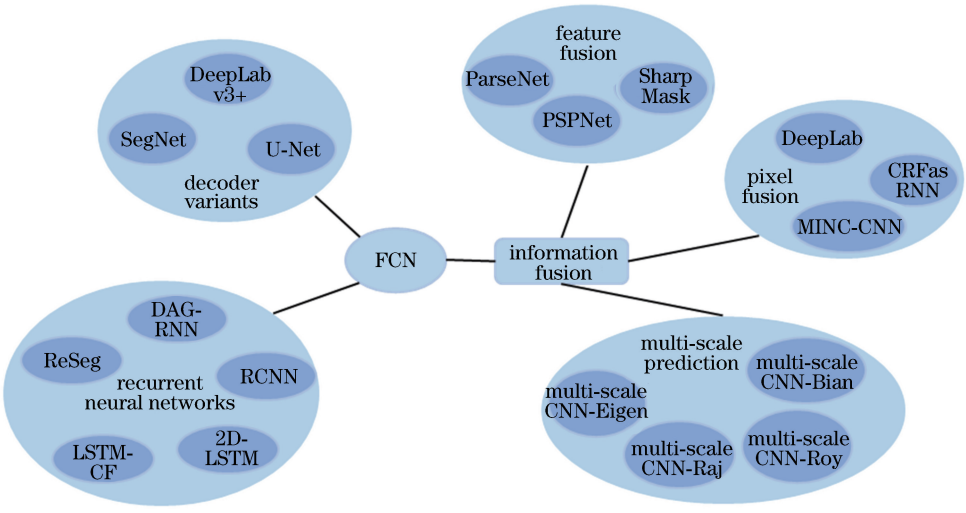

图 7 常见的深度学习语义分割方法分类

Fig. 7 Classification of common deep learning semantic segmentation methods

4.1 基于解码器的方法

2017 年，Shelhamer 等[32]提出了基于全卷积神经网络(FCN)的语义分割方法，该工作使得语义分割领域取得了跨越式的进步，它不仅回答了 CNN 如何在语义分割问题上实现端对端训练的问题，还有效解决了如何对任意尺寸的输入产生像素级别输出的语义预测问题。它的思路是将 CNN 中的全连接层替换为卷积层，从而建立全卷积网络，输入任意尺寸的图像后，经过有效的推理和学习产生相应尺寸的输出，从而对每个像素进行分类，这个部分被称为编码器。分类完成后再通过上采样(也称反卷积)将分类结果映射到原图像大小，产生密集的像素级别的标签，从而获得语义分割的结果，此部分被称为解码器。该方法还融合了多分辨率的信息，将不同大小的特征图进行上采样并进行融合，取得了较为精确的分割效果。然而，该方法也存在着一定的局限性，在进行上采样时容易丢失像素的位置信息，从而影响分割精度。

FCN 是语义分割深度学习算法的开山之作，其处理过程如图 8 所示，它明确了一种语义分割经典模型，选用去掉全连接层的常用分类网络作为基础架构，结合其他的解码变体，将原用于分类的网络(如 VGG-16 等)转化为用于语义分割的网络。如何巧妙地设计解码器是这些方法的区别所在。2017 年，由 Badrinarayanan 等[33]提出的 SegNet 算法设计了一个用于道路场景语义分割的编码器-解码器网络，如图 9 所示。该网络在进行池化时保留了池化层索引，记录下池化层的值在特征图中的空间位置，在恢复图像尺寸时调用池

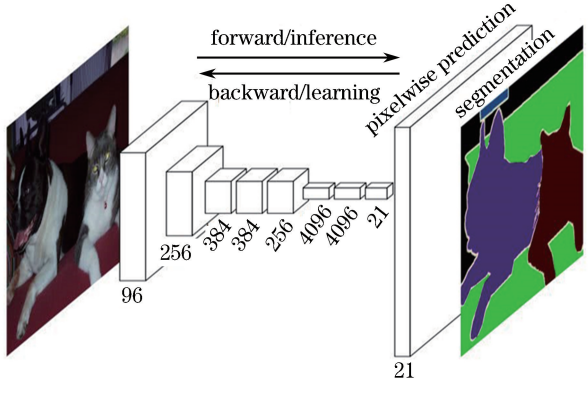

图 8 FCN 网络处理过程图

Fig. 8 FCN network processing diagram

化层索引，从而将该值准确地反映射到其初始位置。该改进不需要进行额外的学习，减少了训练参数，同时能够更加准确地恢复图像边界信息，从而改善了图像分割的效果。但是其在物体边界处的分割精度仍有待进一步提升。

4.2 基于信息融合的方法

常见的 CNN 的池化层具有空间不变性，但也丢失了全局上下文信息。为进一步优化语义分割效果，充分利用目标空间信息，通常需要对不同层次的信息进行融合，主要分为以下 3 种：像素级融合、特征图融合和多尺度融合。

4.2.1 像素级融合

条件随机场(CRF)作为一种常用的后处理模块，通常被引入到卷积网络模型框架中，以实现底层图像信息像素间的融合。与 CNN 相比，CRF 能够更好地学习像素之间的关联性。Chen 等[34]提出的 DeepLabv1 使用 CRF 模型作为其网络中独立

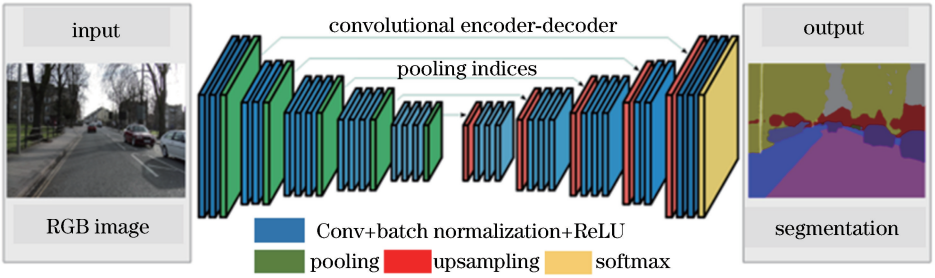

图 9 SegNet 模型的结构示意图

Fig. 9 Structural diagram of SegNet model

的优化环节，将图像中每个像素与模型中的某个节点一一对应，衡量任意像素之间的联系，对分割结果实现了细节增强。图 10 是 CRF 后处理过程发挥模型优化作用的形象化展示。

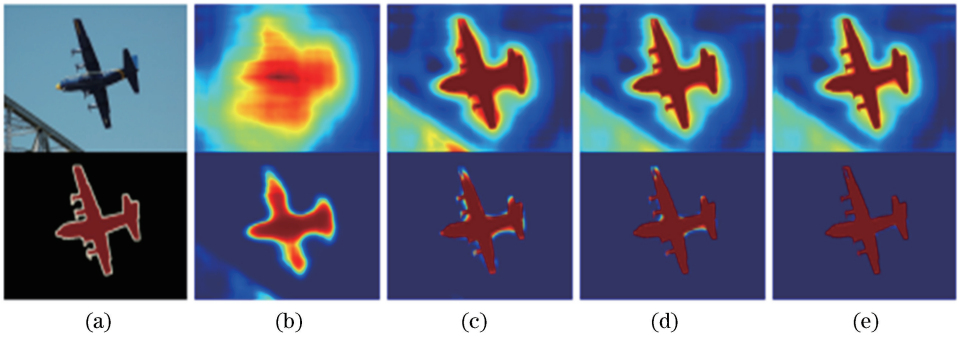

图 10 DeepLab 中利用 CRF 调优迭代产生的影响。(a) GT；(b) CNNout；(c) CRFit1；(d) CRFit2；(e) CRFit10

Fig. 10 Effects of using CRF tuning iterations in DeepLab. (a) GT; (b) CNNout; (c) CRFit1; (d) CRFit2; (e) CRFit10

于 2016 年提出的 DeepLabv2 在 DeepLabv1 的基础上引入了金字塔型空洞池化(ASPP)模块[35]，选择不同采样率的带孔卷积处理特征图，提高了分割精度。DeepLabv3[36]继续优化 ASPP 结构，并引入 Resnet block 模块，通过级联多个空洞卷积结构，有效地提取了表现力强的特征。2018 年，Chen 等[37]提出的 DeepLabv3＋把 DeepLabv3 作为编码器，使用 Xception 网络结构作为基准模型，并设计了一个新型的解码器结构，其测试结果目前在 PASCAL VOC 2012 竞赛中排名第一。考虑到介绍 DeepLab 系列方法的连贯性，本文并没有将此方法划分到解码器一类中进行介绍。

考虑到像素间的局部关联，即邻近像素属于同类别的可能性应该更大，Zheng 等[38]提出了 CRFasRNN 模型，这是通过 CRF 来调优 FCN 网络语义分割结果的另一项重要工作。该工作首先引入平均场的概念，并将其近似为 RNN 结构，成功地将 CRF 与 RNN 整合为一个完整的端对端网络，然后利用随机梯度下降法来求解参数。该工作详细介绍了将 CRF 改造为 RNN 模型进而构造深度网络的过程，得到的效果比 FCN-8s 和 DeepLab 好，如图 11 所示。

4.2.2 特征图融合

在语义分割问题中，另一种对 FCN 进行信息融合的常见做法是进行特征图融合。特征图融合是指将网络中前面层提取到的全局特征图与后面层提取到的局部特征图进行结合。早期 FCN 网络提出的跳跃结构便是基于这一思想对特征映射进行延迟融合。

ParseNet[39]的上下文模块采用了提前融合的方法，其核心思想是利用反池化操作将全局特征图转化为与局部特征图相同的尺寸，合并后输入下一层或用于学习分类器。该网络更加有效地利用了前面层所提供的上下文信息，取得了比 FCN 跳跃结构更好的分割效果。SharpMask[40]引入了一种先进的 Refinement 模块，该网络不仅具有传统自下而上的 CNN，还设计了自上而下的通道。将前向 CNN 网络产生的 mask encoding 和自下而上传递过来的 feature map 融合生成一个新的 mask encoding，持续进行类似操作直到输出具有相同分辨率的精细分割结果。但是该方法主要面向实例分割，这里不再赘述。

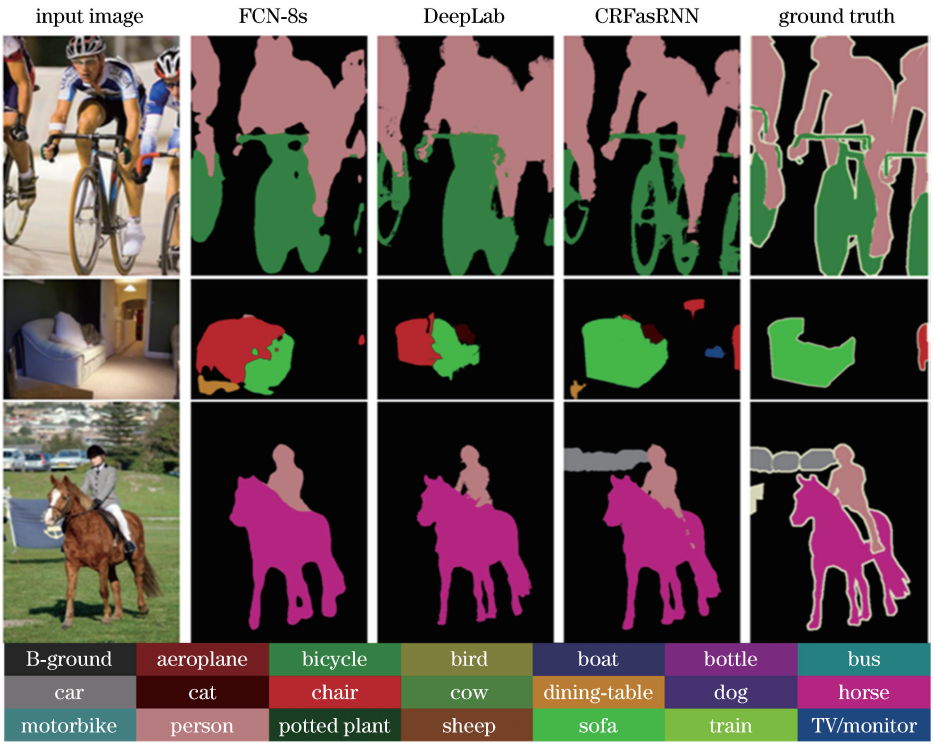

图 11 CRFasRNN、FCN-8s 和 DeepLab 模型效果对比图

Fig. 11 Comparison of three models of CRFasRNN, FCN-8s, and DeepLab

以往，不同形式的特征融合一般都是融合前层与后层的 feature map，未能充分获取全局信息，因此效果并不理想。为充分利用全局特征层次的先验知识，PSPNet[41]采用如图 12 所示的金字塔池化模块聚合不同子区域的不同尺度信息，从而完成多层次的语义特征融合，提高其获取全局信息的能力，在 Cityscapes、PASCAL VOC 2012、ImageNet 2016 三个数据集上都获得了较好的效果。

4.2.3 多尺度融合

另一种实现信息融合的思路便是多尺度融合方法，通过选用多个不同尺度的网络，并结合其预测结果，从而产生一个综合性的输出。Raj 等[42]提出了基于 VGGNet-16 的一种多尺度融合思路，网络包含两个通道：第一个通道使用浅层卷积网络在原始分辨率上对输入图像进行处理，第二个通道使用全卷积 VGGNet-16 和额外的卷积层在 2 倍分辨率上进行处理。将第二个通道的输出进行上采样后与第一个通道的输出相结合，再经过一系列卷积操作，最终的结果对尺度变换的稳健性更强。Roy 等[43]发展了 Eigen 等[44]设计的 4 个多尺度 CNN 的网络架构，利用一个从粗糙到精细的尺度序列来逐步提取特征，对输出进行优化，具体流程如图 13 所示，取得了令人满意的效果。

另一项利用多尺度融合的重要工作是 Bian

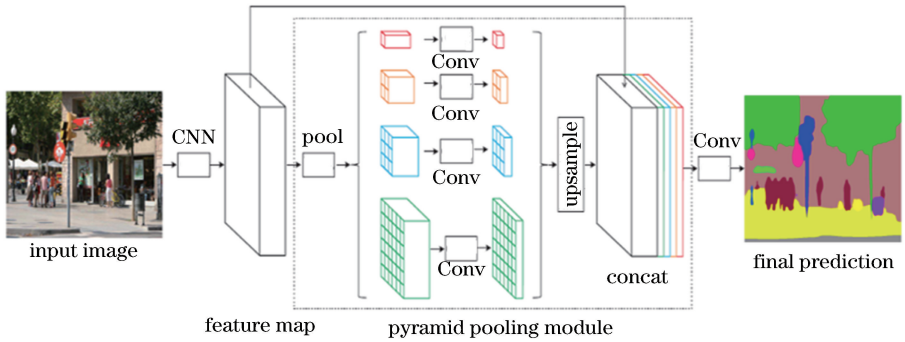

图 12 PSPNet 中的金字塔池化模块示意图

Fig. 12 Diagram of pyramid pooling module in PSPNet

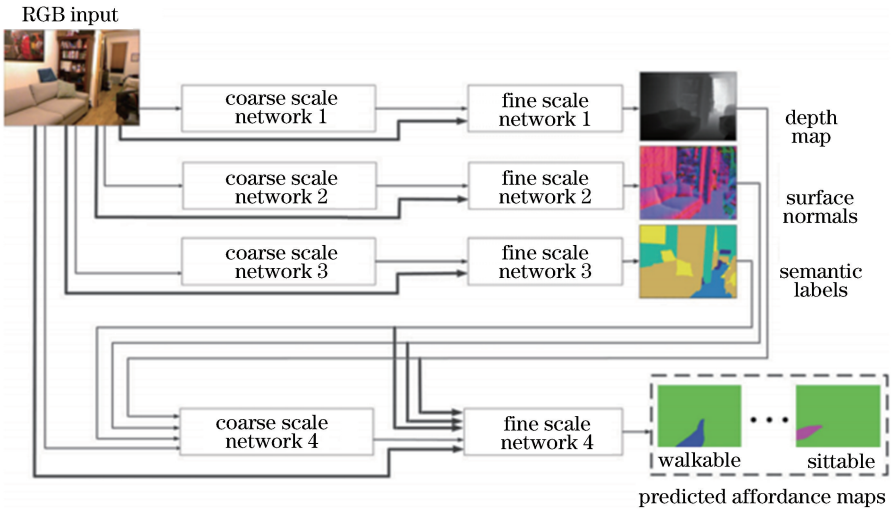

图 13 Roy 等[43]提出的多尺度 CNN 网络架构示意图

Fig. 13 Diagram of multi-scale CNN network architecture proposed by Roy

等[45]提出的多尺度全卷积网络模型，该工作的主要创新是独立训练多个不同尺度的网络，经过必要的上采样后将网络提取的特征进行融合，经卷积处理后得到最终分割结果。该模型可以方便地训练新网络并将其组合，具有很强的灵活性。

4.3 基于 RNN 的方法

CNN 可以看作是一种执行人类视觉功能的模型，但是它缺少记忆功能，只能处理特定的视觉任务，不能依据之前的记忆处理新任务。RNN 是一种基于记忆的网络模型，能够记住网络前面出现的特征，并依据这些特征对输出进行推断，整体网络结构能够不断循环，故而得名循环神经网络。得益于其独特的拓扑结构，RNN 将像素级与局部信息结合起来，成功地应用到全局信息建模和改善语义分割结果中。

Visin 等[53]基于 ReNet 分类网络，提出了面向语义分割的 ReSeg 模型[46]，如图 14 所示。在该方法中，输入图像首先经 VGG-16 的卷积层处理，得到特征映射结果，然后连续通过 5 个 ReNet 层进行微调，经上采样后得到与输入图像尺寸相同的输出结果。在这里，HHA 图像指的是将深度图转换为包含水平差异、对地高度和表面法向量角度 3 种通道的图像。但是，由于 RNN 具有遗忘性，一般的 RNN 在实际应用中都会面临长时依赖问题，不能取得满意的效果。由此出现了两种 RNN 网络结构的改进形式：长短期记忆网络(LSTM)[54]和门控循环单元(GRU)[55]。这两种变式都较好地解决了长时依赖的问题，逐渐成为两种主流的 RNN 结构。其

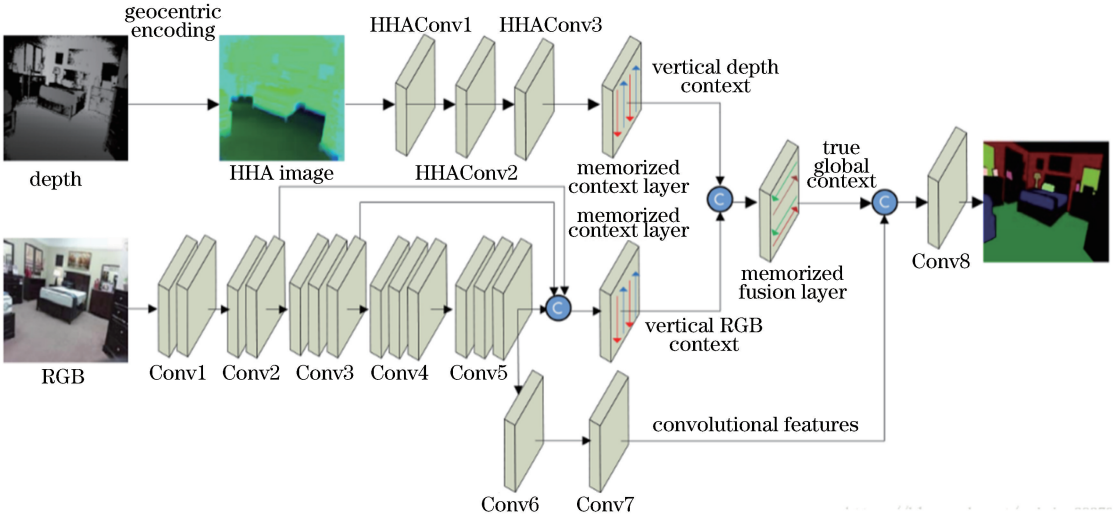

图 14 ReSeg 模型的结构示意图

Fig. 14 Structural diagram of ReSeg model

中，GRU 与 LATM 的最大区别在于 GRU 把输入门和遗忘门合并成一个“更新门”，而且网络没有额外产生记忆状态，只是把输出结果 h_t 当作记忆状态向后不断循环传递，这样网络输入和输出都变得更加简单。图 15 给出了 GRU 内部的网络结构，其中 X_t 表示网络的输入值，z_t 和 r_t 分别表示更新门和重置门，t 表示时间。

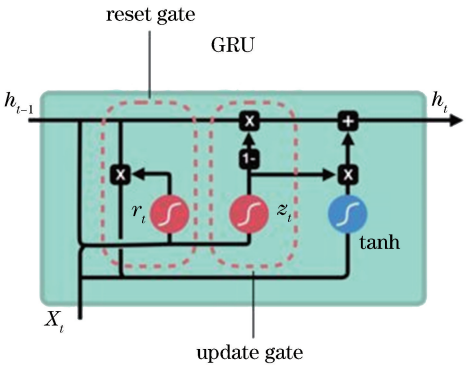

图 15 GRU 计算过程示意图

Fig. 15 Diagram of GRU calculation process

受 ReNet 模型的启发，文献[47]为 RGB-D 语义分割问题提出了一种新型的长期短期记忆网络(LSTM-CF)模型，该模型从颜色通道(RGB)和深度通道(D)中垂直提取和融合特征信息，构建水平方向的记忆网络，结合到深度 CNN 中进行端到端训练，最后生成像素级别的标签图。以往的 RGB-D 语义分割方法通常需要对颜色通道和深度通道分别建立两个 CNN，并将其简单融合后由 FCN 输出结果，忽略了两个通道的内在联系，图像的语义信息损失严重。而该方法可以很好地捕捉到这些信息，语义标记的准确率明显提升。

捕获全局信息的方法依赖于使用更大的输入窗口，虽然能够考虑到更大范围的上下文信息，但也带来了图像分辨率低和窗口重叠等问题。Pinheiro 等[48]使用不同的窗口大小循环地训练循环卷积神经网络(RCNN)，这样就利用了之前层中的预测信息，使得网络语义分割的效果更加平滑。Byeon 等[49]提出了自然场景图像的 2D-LSTM 网络，将输入图像分割为互不重叠的窗口，送入 4 个独立的 LSTM 记忆单元中进行处理。该方法模型简单，与以往方法相比计算复杂度明显降低，实现了较好的分割性能。文献[50]的工作说明 RNN 还能够与有向图(DAG)相结合，通过设计 DAG-RNN 网络处理 DAG 结构化图像，对图像单元之间的远程语义关联进行建模，在 CAMVID 等测试集中取得了理想的测试结果。

综上，应用深度学习来解决图像语义分割问题的发展势头迅猛[56-59]。除上述方法外，近年来仍不断有新的思路和方法出现。为克服不同物体尺寸给语义分割效果带来的负面影响，文献[56]提出了一个新的 FoveaNet 网络来学习图像中的全局透视信息，以解决城市场景图像的理解问题。同时，该网络还引入了一种将透视几何作为先验势场的稠密 CRF 模型，有效地解决了大尺寸物体的识别问题，并在数据集 CITYSCAPES 和 CAMVID 上达到了 state-of-the-art 的性能。针对现有语义分割方法普遍存在的类内一致性和类间区分性的挑战，文献[57]提出一种判别特征网络(DFN)，其包含两个子网络——平滑网络和边界网络，DFN 可以取得理想的分割效果。文献[58]首次将生成对抗网络(GAN)用于语义分割，提出了一个基于 GAN 的半监督网络框架结构，将传统的分割网络作为生成网络，并在其后添加一个判别网络结构，通过 GAN 产生高质量的生成图像来改进像素分类，该方法在 PASCAL、CAMVID 等基准视觉数据集上展现出了很强的竞争力。

5 结束语

从语义分割的基本定义出发，对语义分割中存在的困难和挑战进行了分析和探讨，在介绍 CNN 背景知识的基础上，详细介绍了用于评测语义分割算法的典型数据集，重点对近年来语义分割领域基于解码器、信息融合和 RNN 的深度学习方法进行深入分析和总结。总的来看，利用深度学习方法处理语义分割问题已经取得了较大进步，但在以下方面还有待进一步研究。

1) 训练数据库及应用场景需要进一步丰富。神经网络的训练学习需要海量数据作为支撑，语义分割通常要求对训练集进行像素级别的标注，这需要耗费大量的精力。如何通过有效的弱监督学习缓解这个问题，是未来的研究热点。而且现有的数据集大多局限于室内生活用品和道路交通场景，建立一整套专注于战场环境感知等领域的标准数据集同样迫在眉睫。

2) 语义分割算法的实时性更加受到关注。现有的语义分割算法在分割准确率上已取得较大进展，但也大大增加了模型的复杂度。未来无人驾驶和环境感知等领域对语义分割算法的实时性提出了更高的要求，如何在保持较高分割准确率的同时，进

一步降低模型的复杂度，实现实时语义分割，是未来重要的研究方向。

3）三维（3D）点云语义分割的前景更加广阔。目前基于2D图像的语义分割算法虽已比较成熟，但始终难以取得较大突破。与仅能提供颜色和纹理信息的2D图像相比，3D点云还包含了稳定的深度信息，为语义分割提供了新思路。近年来，深度采集设备和GPU的迅猛发展，加速了3D点云数据的获取和处理进程。因此，如何开发深度学习模型使其较好地实现3D点云语义分割，具有广阔的发展前景。

参 考 文 献

[1] He Y, Wang H, Zhang B. Color-based road detection in urban traffic scenes [J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(4): 309-318.

[2] An Z, Xu X P, Yang J H, *et al*. Design of augmented reality head-up display system based on image semantic segmentation[J]. Acta Optica Sinica, 2018, 38(7): 0710004.
安喆, 徐熙平, 杨进华, 等. 结合图像语义分割的增强现实型平视显示系统设计与研究[J]. 光学学报, 2018, 38(7): 0710004.

[3] Ros G, Sellart L, Materzynska J, *et al*. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 26-July 1, 2016, Las Vegas, Nevada, USA. New York: IEEE, 2016: 3234-3243.

[4] Yi Z, Criminisi A, Shotton J, *et al*. Discriminative, semantic segmentation of brain tissue in MR images[M]// Yang G Z, Hawkes D, Rueckert D, *et al*. Medical image computing and computer-assisted intervention -MICCAI 2009. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2009, 5762: 558-565.

[5] Liu H, Peng L, Wen J W. Multi-scale aware pedestrian detection algorithm based on improved full convolutional network[J]. Laser & Optoelectronics Progress, 2018, 55(9): 091504.
刘辉, 彭力, 闻继伟. 基于改进全卷积网络的多尺度感知行人检测算法[J]. 激光与光电子学进展, 2018, 55(9): 091504.

[6] Simo-Serra E, Fidler S, Moreno-Noguer F, *et al*. A high performance CRF model for clothes parsing[M]// Cremers D, Reid I, Saito H, *et al*. Computer vision—ACCV 2014. Lecture notes in computer science. Cham: Springer, 2015, 9005: 64-81.

[7] Dollar P, Appel R, Belongie S, *et al*. Fast feature pyramids for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532-1545.

[8] Girshick R, Donahue J, Darrell T, *et al*. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. (2015-04-10)[2019-01-05]. https://arxiv.org/abs/1409.1556.

[11] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 15523970.

[12] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[13] Mohamed A R, Dahl G E, Hinton G. Acoustic modeling using deep belief networks [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.

[14] Cheng G J, Liu L T. Feasibility study of deep learning algorithm applied to rock image processing [J]. Software Guide, 2016, 15(9): 163-166.
程国建, 刘丽婷. 深度学习算法应用于岩石图像处理的可行性研究[J]. 软件导刊, 2016, 15(9): 163-166.

[15] Wang L, Liu Q. A multi-object image segmentation algorithm based on local features [J]. Laser & Optoelectronics Progress, 2018, 55(6): 061002.
王琳, 刘强. 基于局部特征的多目标图像分割算法[J]. 激光与光电子学进展, 2018, 55(6): 061002.

[16] Shi J B, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[17] Wu S Q, Nakao M, Matsuda T. Automatic GrabCut based lung extraction from endoscopic images with an initial boundary[C]// 2016 IEEE 13th International Conference on Signal Processing (ICSP), November 6-10, 2016, Chengdu, China. New York: IEEE, 2017: 1374-1378.

[18] Everingham M, Eslami S M A, van Gool L, *et al*. The Pascal visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[19] Hariharan B, Arbelaez P, Bourdev L, *et al*. Semantic contours from inverse detectors[C]// 2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE, 2012: 991-998.

[20] Mottaghi R, Chen X J, Liu X B, *et al*. The role of context for object detection and semantic segmentation in the wild[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 891-898.

[21] Wang J Y, Yuille A. Semantic part segmentation using compositional model combining shape and appearance[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 1788-1797.

[22] Garcia-Garcia A, Orts-Escolano S, Oprea S, *et al*. A review on deep learning techniques applied to semantic segmentation[J/OL]. (2017-04-22)[2019-01-05]. https://arxiv.org/abs/1704.06857.

[23] Lin T Y, Maire M, Belongie S, *et al*. Microsoft COCO: common objects in context[M]// Fleet D, Pajdla T, Schiele B, *et al*. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[24] Cordts M, Omran M, Ramos S, *et al*. The cityscapes dataset for semantic urban scene understanding[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 3213-3223.

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

[26] Sturgess P, Alahari K, Ladicky L, *et al*. Combining appearance and structure from motion features for road scene understanding[C]// Proceedings of the British Machine Vision Conference 2009, September 7-10, 2009, London. Durham, England, UK: BMVA Press, 2009: 62.

[27] 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, January 5-9, 2015, Waikoloa, HI, USA. New York: IEEE, 2015: 231-238.

[28] Zhang R, Candra S A, Vetter K, *et al*. Sensor fusion for semantic segmentation of urban scenes[C]// 2015 IEEE International Conference on Robotics and Automation (ICRA), May 26-30, 2015, Seattle, WA, USA. New York: IEEE, 2015: 1850-1857.

[29] 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.

[30] Alvarez J M, Gevers T, LeCun Y, *et al*. Road scene segmentation from a single image[M]// Fitzgibbon A, Lazebnik S, Perona P, *et al*. Computer vision-ECCV 2012. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2012, 7578: 376-389.

[31] Ros G, Alvarez J M. Unsupervised image transformation for outdoor semantic labelling[C]// 2015 IEEE Intelligent Vehicles Symposium (IV), June 28-July 1, 2015, Seoul, Korea. New York: IEEE, 2015: 537-542.

[32] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

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

[34] Chen L C, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J/OL]. (2016-06-07)[2019-01-05]. https://arxiv.org/abs/1412.7062.

[35] 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.

[36] Chen L C, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation[J/OL]. (2017-12-05)[2019-01-05]. https://arxiv.org/abs/1706.05587.

[37] Chen L C, Zhu Y K, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]// Ferrari V, Hebert M, Sminchisescu C, *et al*. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.

[38] Zheng S, Jayasumana S, Romera-Paredes B, *et al*. Conditional random fields as recurrent neural networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1529-1537.

[39] Liu W, Rabinovich A, Berg A C. ParseNet: looking wider to see better[J/OL]. (2015-11-19)[2019-01-05]. https://arxiv.org/abs/1506.04579.

[40] Pinheiro P O, Lin T Y, Collobert R, *et al*. Learning to refine object segments[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision—ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 75-91.

[41] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE, 2017: 6230-6239.

[42] Raj A, Maturana D, Scherer S. Multi-scale convolutional architecture for semantic segmentation[R]. Pittsburgh, Pennsylvania: Carnegie Mellon University, 2015: CMU-RITR-15-21.

[43] Roy A, Todorovic S. A multi-scale CNN for affordance segmentation in RGB images[M]// Leibe B, Matas J, Sebe N, *et al*. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9908: 186-201.

[44] 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), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2016: 2650-2658.

[45] Bian X, Lim S N, Zhou N. Multiscale fully convolutional network with application to industrial inspection[C]// 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), March 7-10, 2016, Lake Placid, NY, USA. New York: IEEE, 2016: 16035894.

[46] Visin F, Romero A, Cho K, *et al*. ReSeg: a recurrent neural network-based model for semantic segmentation[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 426-433.

[47] Li Z, Yukang Gan Y K, Liang X D, *et al*. RGB-D scene labeling with long short-term memorized fusion model[J/OL]. (2016-07-26)[2019-01-05]. https://arxiv.org/abs/1604.05000v1.

[48] Pinheiro P H O, Collobert R. Recurrent convolutional neural networks for scene parsing[J/OL]. (2013-06-12)[2019-01-05]. https://arxiv.org/abs/1306.2795.

[49] Byeon W, Breuel T M, Raue F, *et al*. Scene labeling with LSTM recurrent neural networks[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3547-3555.

[50] Shuai B, Zuo Z, Wang B, *et al*. DAG-recurrent neural networks for scene labeling[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 3620-3629.

[51] Bell S, Upchurch P, Snavely N, *et al*. Material recognition in the wild with the Materials in Context Database[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3479-3487.

[52] Pinheiro P O, Collobert R, Dollár P. Learning to segment object candidates[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems, December 7-12, 2015, Montreal, Canada. Cambridge: MIT Press, 2015, 2: 1990-1998.

[53] Visin F, Francesco K, Cho K, *et al*. ReNet: a recurrent neural network based alternative to convolutional networks[J/OL]. (2015-07-23)[2019-01-05]. https://arxiv.org/abs/1505.00393.

[54] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[55] Wu Z Z, King S. Investigating gated recurrent networks for speech synthesis[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 20-25, 2016, Shanghai, China. New York: IEEE, 2016: 5140-5144.

[56] Li X, Jie Z Q, Wang W, *et al*. FoveaNet: perspective-aware urban scene parsing[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 784-792.

[57] Yu C Q, Wang J B, Peng C, *et al*. Learning a discriminative feature network for semantic segmentation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 1857-1866.

[58] Souly N, Spampinato C, Shah M. Semi supervised semantic segmentation using generative adversarial network[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 5689-5697.

[59] Guo C C, Yu F Q, Chen Y. Image semantic segmentation based on convolutional neural network feature and improved superpixel matching[J]. Laser & Optoelectronics Progress, 2018, 55(8): 081005.
郭呈呈, 于凤芹, 陈莹. 基于卷积神经网络特征和改进超像素匹配的图像语义分割[J]. 激光与光电子学进展, 2018, 55(8): 081005.