基于特征融合的实时语义分割算法

蔡雨*，黄学功，张志安，朱新年，马祥
南京理工大学机械工程学院，江苏 南京 210094

摘要 为满足自动驾驶、人机交互等任务对语义分割算法准确度和实时性的要求，提出一种基于特征融合技术的实时语义分割算法。首先，利用卷积神经网络自动学习图像深层特征的功能，设计一个浅而宽的空间信息网络输出低级别的空间信息，以保持原始空间信息完整性，从而生成高分辨率特征；接着，设计一个语境信息网络来输出深层次、高级别的语境信息，并引入注意力优化机制来代替上采样，优化网络的输出；最后，将两路输出特征图进行多尺度融合，再上采样得到与原始输入尺寸相等的分割图像。两路网络并行计算，提高了算法的实时性。在 Cityscapes、CamVid 数据集上对该网络框架进行一系列实验。其中，在 Cityscapes 数据集上取得了 68.43%的均交并比(MIOU)。对于 640×480 的图像输入，在一块 NVIDIA 1050T 显卡上的速度为 14.14 frame/s。本文算法在准确度上大幅超越现有实时分割算法，基本满足人机交互类任务对实时性的要求。

关键词 图像处理；语义分割；卷积神经网络；特征融合；注意力机制；轻量化模型；并行计算
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.021011

Real-Time Semantic Segmentation Algorithm Based on Feature Fusion Technology

Cai Yu*, Huang Xuegong, Zhang Zhian, Zhu Xinnian, Ma Xiang
College of Mechanical Engineering, Nanjing University of Science & Technology, Nanjing, Jiangsu 210094, *China*

Abstract In this study, we propose a real-time semantic segmentation algorithm based on the feature fusion technology to satisfy the requirements of autopilot, human-computer interaction, and other tasks with respect to accuracy and real-time capability . Here, we use a convolutional neural network to automatically learn deep features of the image. We design a shallow and wide spatial information network to output low-level spatial information for ensuring the integrity of the original spatial information and generating high-resolution features. Furthermore, we design a context information network to output deep-level high-level context information. Then, we introduce an attention optimization mechanism to replace upsampling for optimizing the network output. Finally, we fuse the two output feature maps on multiple scales and perform upsampling to obtain a segmented image with a size equal to the original input size. Subsequently, we perform a simulation using two-way network parallel computing to improve the real-time performance of the proposed algorithm. The network framework achieves 68.43% mean intersection over union (MIOU) on the Cityscapes dataset. In case of an image input of 640 × 480, the speed obtained using an NVIDIA 1050T graphics card is 14.14 frame/s. Furthermore, the accuracy considerably exceeds that of the existing real-time segmentation algorithm, satisfying the real-time requirements of the human-computer interaction tasks.

Key words image processing; semantic segmentation; convolutional neural network; feature fusion; attention mechanism; lightweight model; parallel computing
OCIS codes 100.2000; 100.4996; 200.4260; 200.4960

1 引 言

图像分类、目标检测、语义分割是计算机视觉的三大基本任务。其中，图像分类就是判断图像中的内容所属的分类；目标检测是在识别图像内容的同时，还要给出目标的位置信息；图像语义分割融合了

收稿日期：2019-05-21；**修回日期**：2019-06-14；**录用日期**：2019-07-09
基金项目：国家自然科学基金(11472008,11772160,11202206)
* **E-mail**：1204246973@qq.com

传统的图像分割和目标识别两个任务，其目的是将图像分割成几组具有某种特定语义含义的像素区域，并识别出每个区域的类别，最终获得一幅具有像素语义标注的图像[1]。目前，图像语义分割的研究非常活跃，是遥感测绘、自动驾驶、医疗影像分析、人机交互等应用领域[2-3]的核心问题。例如在时尚领域，通过对人体的语义分割可以定位出人脸、躯干、着装等信息，从而帮助网民在互联网购物过程中实现自动试衣等功能；在自动驾驶领域，通过对车体前方场景的语义分割可以精确定位道路、车体和行人等场景或物体信息，从而提升自动驾驶的安全性。为了保证在这些任务中获得更好的表现，对语义分割算法的准确性、实时性和稳健性提出了较高的要求[4]。

传统基于机器学习的语义分割算法，是将输入图像分割成一系列具有独特性质的区域块，从而提取每个区域块的图像特征（如纹理、颜色、形状等），然后根据一定的规则建立图像特征与高层语义之间的概率图模型，最后通过学习得到模型参数。这类方法需要依靠先验知识进行人工选择和设计，耗时耗力，虽然算法精度较高，但稳健性较差，可用于一些图像特征变化不大的特定场合，如医学影像的分割[5]。卷积神经网络（CNN）是一种多层的监督学习网络，应用在计算机视觉领域能表现出很好的稳健性。近些年 CNN 的快速发展，为语义分割提供了新的解决方案。

2017 年，Shelhamer 等[6]提出一种全连接卷积神经网络（FCNN），该网络将普通的分类网络的全连接层替换为对应尺寸的卷积层，再通过上采样将特征图恢复成原始输入图像尺寸。FCNN 在 PASCAL VOC 数据集上的均交并比（MIOU）得分，相比 2012 年提高了 20%，成为当时最出色的语义分割算法，成为后来评价算法准确性的标杆，语义分割算法正式进入深度学习时代。之后，其他基于神经网络方法进行特征提取的语义分割算法相继被提出。Badrinarayanan 等[7]提出 SegNet，该网络是一种“编码器-解码器”的结构，在编码过程中，通过卷积提取特征，利用池化过程增大感受野。在解码过程中，在相应编码器的最大池化过程中，使用计算的池化索引来执行非线性上采样，能减少对上采样过程的学习。因此，SegNet 在保证性能的同时，能减少对内存的占用，提高了计算速度，为后来的实时性语义分割算法研究提供了一个思路。很多实时性算法，如 ENet，就是基于 SegNet 的设计改进而来的。Chen 等[8]提出的 DeepLab 网络，使用空洞卷积进行上采样，在不增加参数量的基础上增大感受野，能够更好地提取语境信息，并结合了马尔科夫随机场的概率模型来提高物体的边界信息还原精度。DeepLab 网络首次提出语境信息这一概念，在准确性上取得了很好的成绩，但该算法利用概率模型来还原边界信息，实时性很差。Peng 等[9]提出全局卷积网络（GCN），该网络引入基于残差的边界细化模块，进一步提高目标边界附近的定位性能，并且首次验证了在语义分割任务上，卷积结构内核尺寸的增大有利于平衡特征提取和细节还原性能之间的矛盾。

上述研究的主要工作都是为了恢复网络不断下采样时造成的信息损失，评定算法优劣的性能指标也更多聚焦在像素精度上。然而，近些年来自动驾驶、人机交互等技术的快速发展，设计出兼顾准确度和实时性的网络成为了学者研究的重点。

语义分割算法的实时性与模型中基础网络的复杂度、模型的输入尺寸、卷积核的大小、网络的通道数目都有关系。因此目前，提高算法实时性的方法主要有三大类[10-11]：

1）通过剪裁或调整尺寸来减少输入图像的数据量；

2）减少网络的通道数，尤其是在基本网络的早期阶段；

3）放弃模型的上采样阶段。

这些方法虽然简单有效，但空间细节的丢失会破坏预测，特别是在边界周围。换句话说，上述方法只是简单粗暴地减少待处理的数据量，并没有平衡网络实时性和准确度之间的矛盾。基于此，本文设计一种基于多尺度融合的实时语义分割算法，能兼顾算法的实时性和准确性。

2 实时分割网络框架设计

本文设计了一个由两路网络融合而成的神经网络框架。两路网络分别解决空间信息损失和提升感受野的问题，并设计特征融合模块来整合这两路网络的特征。两路网络同时对原始图像进行处理，大大提升了效率。网络整体框架如图 1 所示。

2.1 空间信息网络

空间信息网络主要提取图像的空间信息，尤其是对于边缘丰富的图像，要保留足够的空间分辨率，才能保证较好的分割效果。CNN 用连续的下采样操作编码高级语义信息，并已经被证明能够有效地

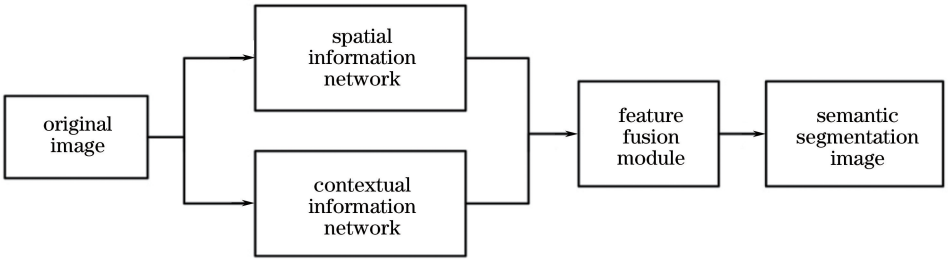

图 1 语义分割网络框架图

Fig. 1 Framework of semantic segmentation network

提取特征。但是 CNN 的层数越多，步长越大，空间位置信息的损失就越严重，而且全连接或池化操作也会使空间细节丢失。

根据上述描述，本文设计了一个“浅而宽”(层数少、步长短)的 CNN。同时，不在卷积层后添加全连接或者全局池化等操作，不对图像进行剪裁或调整尺寸的处理，而是使用原始图像作为网络的输入。这样的设计能保证丰富的空间信息不丢失，由于网络结构被简化，也提高了实时性。网络框架及细节如表 1 所示，其中 W 为卷积层的宽度，H 为卷积层的高度。

表 1 空间信息网络结构

Table 1 Spatial information network structure

Structure	Input	Convolution size	Mapping	Output
Conv1	$W\times H\times 3$	$3\times 7\times 7\times 64$	ReLU	$\frac{1}{2}W\times\frac{1}{2}H\times 64$
Conv2	$\frac{1}{2}W\times\frac{1}{2}H\times 64$	$64\times 3\times 3\times 64$	ReLU	$\frac{1}{4}W\times\frac{1}{4}H\times 64$
Conv3	$\frac{1}{4}W\times\frac{1}{4}H\times 64$	$64\times 3\times 3\times 64$	ReLU	$\frac{1}{8}W\times\frac{1}{8}H\times 64$
Conv4	$\frac{1}{8}W\times\frac{1}{8}H\times 64$	$64\times 1\times 1\times 128$	ReLU	$\frac{1}{8}W\times\frac{1}{8}H\times 128$

该网络为 4 层(Conv1、Conv2、Conv3、Conv4)，前 3 层每一层包含一个步长为 2 的卷积层，每一个卷积层都有对应的批量归一化层(BN)和激活层(ReLU)[12-13]。最后一层为一个“1×1”的卷积层，在保持特征图尺寸不变的情况下，大幅增加网络的非线性特性，可增加网络的维度。表中输入图像的原始尺寸为 $W\times H$，最后输出的图像为原始图像的 1/8。

2.2 语境信息网络

在满足丰富的细节信息的同时，语义分割需要语境信息来提高分割的效果。现有的关注语境信息的提取算法有金字塔池化(PSPNet)[14]、大内核等，但这些方法会消耗内存，降低速度。

考虑到实时性要求，本文算法选择轻量级模型 Xception[15] 作为基础网络，可以快速对特征进行下采样，并且拥有更大的感受野。然后在模型的尾部添加一个全局平均池，它可以提供具有全局上下文信息的最大接收域，能稳定最大感受野。同时，设计了注意力优化模块，优化感受野网络中每一个阶段的特征输出，便于整合全局语境信息，大大降低计算成本。语境信息网络框架如图 2 所示。

2.2.1 轻量化模型

Xception 是一种基于 Inception 系列网络的借鉴深度可分卷积思想的网络[16]。图 3 为设计的轻量化深度学习基础模型。Xception 网络首先对图像进行“1×1”的卷积处理；卷积后生成的所有特征图按通道全部分离(通常为 3 个)，每一个通道对应进行“3×3”卷积操作；最后将各个通道简单相加。这样构建的网络，参数减少，计算复杂程度降低，模型的计算速度提高。实验验证 Xception 网络在训练过程中的收敛速度比 Inception 更快，在 Image-Net 上也取得了较好的成绩。

2.2.2 注意力模块

注意力模块可以使用高级信息来指导前馈网络[17]。本文算法在感受野网络中引入注意力机制，对基础模型的输出进行处理，得到与尺度特征对应的可解释的权重图，决定为不同位置像素赋予多少注意力，使得模型自适应地为同一场景中的不同对象选择合适的分割尺度，可以有效地提高算法的准确度和稳健性。

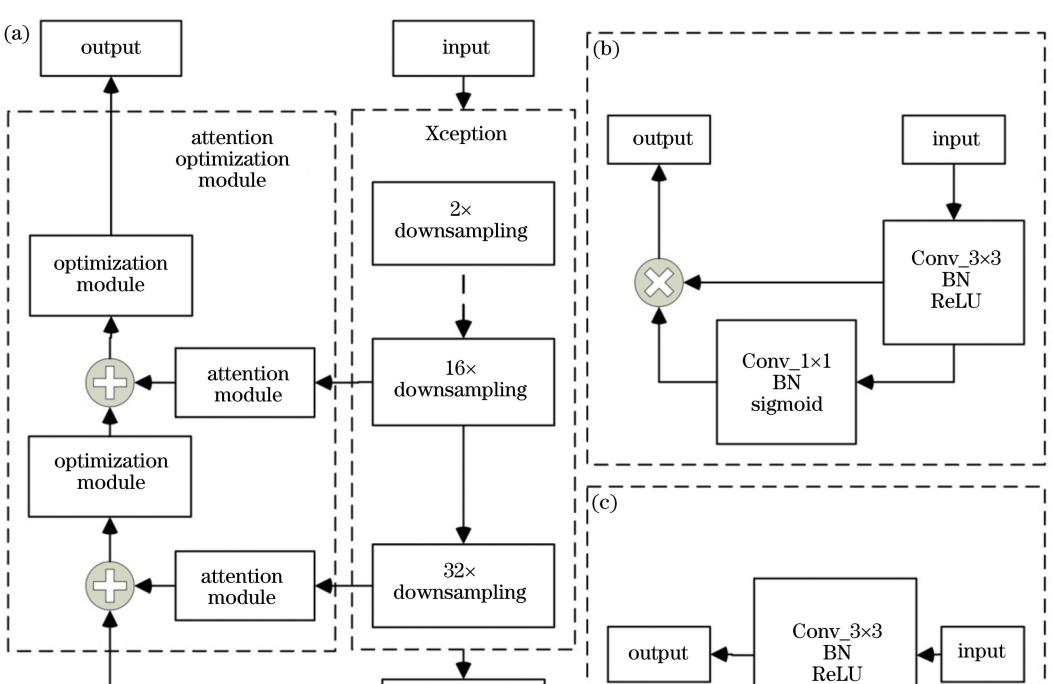

图 2 语境信息网络框架图。(a)整体框架;(b)注意力模块;(c)优化模块

Fig. 2 Framework of context information network. (a) General framework; (b) attention module; (c) optimization module

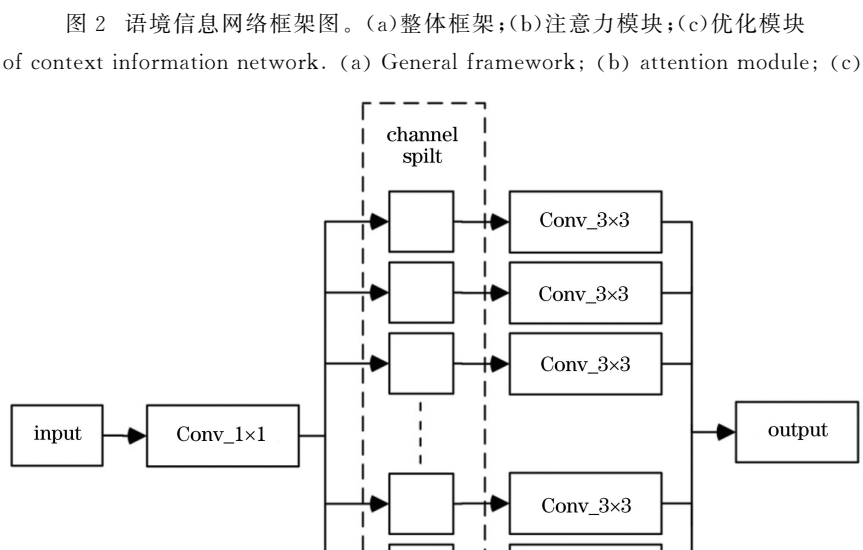

图 3 可分卷积网络框架图

Fig. 3 Framework of separable convolution network

本文将注意力模块引入上采样操作,用来整合全局语境信息。实验证明注意力模块对算法的准确性有明显的提升作用。本文通过引入额外的 CNN 来构建注意力模块,网络细节及设计原因如下。

1) 机制结构细节如图 2(a)所示;

2) 模型输入:对基础模型 Xception 的最后两层的输出进行处理,因为这两层能学习到具有更高语义的尺度信息;

3) 将特征图的不同通道导入注意力模块中进行处理,注意力模块中的“3×3”滤波器(Conv_3×3)可以考虑周围像素对该位置像素的影响,“1×1”的滤波器(Conv_1×1)可以实现跨通道的交互和信

息整合；

4）采用线性插值进行上采样：算法速度快，满足实时性要求。

2.3 特征融合模块

两条路径的特征在特征表示的层次上是不同的。空间信息网络捕获的空间信息编码了大量的低层细节信息，语境信息网络输出的主要是深层的语境信息。由于两路网络的特征并不同，因此不能简单地加权这两种特征。因此，本文设计了一个卷积网络以训练学习如何叠加这两个网络来融合这些功能，并将这个网络命名为特征融合模块。网络框架细节如图 4 所示。

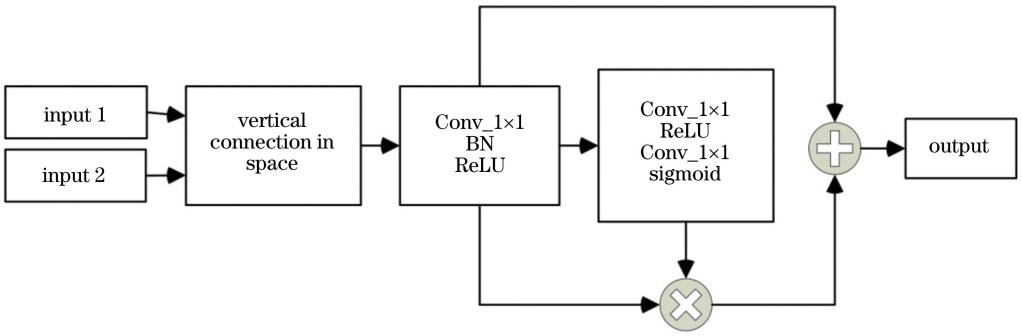

图 4 特征融合网络框架图

Fig. 4 Framework of feature fusion network

本文算法先将两路网络的输出在尺度空间上纵向连接起来；然后利用一个“1×1”的卷积网络来实现跨通道的信息整合，并接上对应的 BN 层和 ReLU 层来平衡特征的尺度；接着将特征权重与特征图相加，得到融合后的特征图；最后将其双线性插值放大 8 倍，即得到与原始图像大小相等的分割图像。

3 实 验

3.1 模型训练

为验证本文算法的有效性，采用 Cityscapes 数据集中的训练集进行训练，训练出的模型分别用 Cityscapes 的测试集和 CamVid 的测试集进行测试。Cityscapes 数据集总共有 5000 张高分辨率图片，共有 19 种语义类别。CamVid 数据集是第一个具有对象类语义标签的视频集合，其中包含元数据，共有 32 种语义类别。实验中 FCNN、ENet[18]、PSPNet、DeepLab 算法使用原作者的开源代码，本文算法使用 Pytorch 框架编写。实验平台的硬件环境为 IntelI7(3.5 GHz)及一块 NVIDIA GTX1050T 显卡，软件环境为 Ubuntu16.04 操作系统，联合使用 cuda 7.5并行计算架构和 cudnnv2 计算加速方案进行 GPU 加速。

本文以语义分割领域公认的指标 MIOU 作为评价标准。MIOU 计算公式为

$$M_{\mathrm{IOU}}=\left(\frac{1}{n}\right)\sum_{i}n_{ii}/\left(t_i+\sum_{j}n_{ji}-n_{ii}\right),\quad(1)$$

式中：n 表示数据集中的语义类别数；n_{ji} 表示属于 i 类而被分为 j 类的像素个数；n_{ii} 表示正确分类的像素数目；t_i 表示类别为 i 的像素总数目。

在训练过程中，通过一个损失函数监督整个分割网络的输出。损失函数都是 Softmax 损失。本文使用随机梯度下降(SGD)来优化收敛过程，应用“多边形”学习率衰减策略[19]，批量大小为 4，冲量为 0.9，权值衰减率为 1×10^{-4}，初始学习率为 2.5×10^{-2}。

3.2 实验及结果

本文利用训练出的模型，对 Cityscapes、CamVid 数据集中的图像进行处理，输出结果使用 Visdom 进行可视化。算法分割结果如图 5 所示。对比图 5 中的三张图，可以看出本文算法能够有效分割图片中的各类物体，分割效果接近真实分割图。

对本文算法进行实验，验证算法结构设计的合理性，并与其他算法进行对比实验，从准确性和实时性两方面分析算法效果。

3.2.1 算法模块性能实验

对算法内部各模块分别进行实验，以 MIOU 为评价标准，验证算法各模块是否能够提升整体算法的性能。实验结果如表 2 所示。

直接对语境信息网络进行双线性插值放大 8 倍，基于 Cityscapes 进行验证，MIOU 达到了 60.56%。添加空间信息网络并使用特征融合模块对两路网络进行特征融合，再线性放大，得到 63.47%的 MIOU，性能提升了 4.8%。最后加上注意力机制模块，得到了68.43%的MIOU，相比前一

图 5 基于 Cityscapes 数据集本文分割算法的示例结果图。(a)原始图片;(b)真实分割结果;(c)本文算法分割结果

Fig. 5 Example results of proposed segmentation algorithm based on Cityscapes dataset. (a) Original image; (b) true segmentation result; (c) result of proposed algorithm

表 2 各模块性能

Table 2 Module performance

Experiment	Method	MIOU /%
Experiment 1	Spatial information network	60.56
Experiment 2	Experiment 1+context information network	63.47
Experiment 3	Experiment 2+attention model	68.43

算法,该算法性能提升了 7.8%。

上述实验证明,算法中各模块对基础模型的性能都有提升作用,算法结构设计合理。尤其是注意力机制,配合前文设计的损失函数,训练出的模型性能得到大大提升。

3.2.2 对比实验和适用性实验

首先,基于 Cityscapes 测试集,用 FCNN、ENet、PSPNet、DeepLab 算法与本文算法进行对比,分析算法的性能表现。之后,将基于 Cityscapes 训练出的模型放到 CamVid 数据集的测试集中进行性能测试,验证本文算法的适用性。算法部署在 NVIDIA 1050T 显卡上运行,输入的图像尺寸统一调整为 640×480。

其中,FCNN 是语义分割中一种简单有效的算法,在语义分割领域具有代表地位。FCNN 将传统 CNN 分类网络中的全连接层用卷积层代替,这样的设计使卷积层保留了像素之间的位置信息和关联信息。同时,首次提出采用反卷积的方法来进行上采样,输出与原始图像大小一致的分割图,但是直接通过反卷积输出的结果是很粗糙的,会损失很多细节。FCNN 使用一种跳跃结构,结合上采样和上卷积赤化后的数据修复还原的图像。近些年,FCNN 算法不断发展,在各测试集上都获得了不错的表现,已经成为语义分割算法在准确度上的标杆。这里,我们选择 FCNN 中效果最好的 FCNN-8s 算法。

DeepLab 算法提出使用空洞卷积来代替上采样,并首次提出感受野、语境信息等概念,这对语义分割任务的发展具有重大意义。DeepLab 算法不断发展,已经迭代出多个版本。这里我们使用最新的 DeepLab-v3 算法,DeepLab-v3 通过引入多孔空间金字塔(ASPP)模块,可以在多尺度上捕获信息,增大感受野,提升边界分割效果,同时使用全连接条件随机场,利用低层的细节信息对分类的局部特征进行优化。

在目前主流的语义分割算法中,PSPNet 是精度最高的算法之一,这主要归功于金字塔池化模块。PSPNet 提出的金字塔池化模块能够聚合不同区域的上下文信息,达到语境和空间细节的融合,从而提高获取全局信息的能力。

ENet 是基于实时性的语义分割算法,是目前速度最快的语义分割算法之一,同时在准确性上取得了较好的成绩。ENet 算法框架采用经典“编码器-解码器”结构,将编码阶段进行下采样的索引信息保留到解码阶段进行上采样时使用。这样的设计大大减少了内存的需求。同时,ENet 算法使用空洞卷积扩大来更好地提取语境信息。

基于 Cityscapes 数据集的 5 种算法的对比实验结果如表 3 所示。

从表 3 中可以看出,在 Cityscapes 数据集上,本文算法获得了 68.43%的 MIOU,相比 FCNN-8s 提高了 10.76%,达到了目前语义分割领域的优秀水准。但相比于 DeepLab-v3 的 70.44%和 PSPNet 的 75.84%,还有一定差距。对比实时分割算法 ENet,在实时性上,本文算法取得了 14.14 frame/s 的成绩,略优于 ENet 的 13.62 frame/s,实时性能提升了 4%,基本满足交互性任务对实时性的要求;在准确度上,相比 ENet 算法,MIOU 提升了 30.74%。分析最后两组实验发现,本文算法基于 Cityscapes 数据集训练出的模型,在其 CamVid 测试集上同样取得了不错的成绩。综上分析,本文算法是比 ENet 性能更好的实时分割算法,精确度上达到优秀水准,

表 3 算法性能对比

Table 3 Comparison of algorithm performance

Algorithm	Base model	Data set	Time /ms	Speed /(frame·s^{-1})	MIOU /%
FCNN-8s	Paper source	Cityscapes	330.03	3.03	61.78
DeepLab-v3+CRF	Paper source	Cityscapes	1162.79	0.86	70.44
PSPNet	Paper source	Cityscapes	2380.95	0.42	75.84
ENet	Paper source	Cityscapes	73.42	13.62	52.34
Proposed algorithm	Xception	Cityscapes	70.72	14.14	68.43
Proposed algorithm	Xception	CamVid	70.72	14.14	67.28

并具有一定的适用性。

4 结 论

提出的语义分割网络在设计上兼顾了准确性和实时性。轻量化模型、减少网络层数、引入注意力机制代替上采样、双线性插值、两路网络并行计算等设计，有效地提高了算法的实时性。两路网络分别对低级的空间信息和高级的语境信息进行提取，并用卷积网络进行特征融合的设计，保证了算法的准确性。而引入注意力机制后，算法获得了更大的感受野，在准确性上又有了不小的提升。本文算法在 Cityscapes 数据集上得到了 68.43% 的 MIOU，部署在 NVIDIA 1050T 显卡上能够达到 14.14 FPS，表现优于 ENet 实时分割算法。

参 考 文 献

[1] Domke J. Learning graphical model parameters with approximate marginal inference [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2454-2467.

[2] Fang X, Wang G H, Yang H C, et al. High resolution remote sensing image classification combining with mean-shift segmentation and fully convolution neural network [J]. Laser & Optoelectronics Progress, 2018, 55(2): 022802.
方旭, 王光辉, 杨化超, 等. 结合均值漂移分割与全卷积神经网络的高分辨遥感影像分类[J]. 激光与光电子学进展, 2018, 55(2): 022802.

[3] Bao Z Q, Lü C G. Real-time gesture recognition based on Kinect [J]. Laser & Optoelectronics Progress, 2018, 55(3): 031008.
鲍志强, 吕辰刚. 基于 Kinect 的实时手势识别[J]. 激光与光电子学进展, 2018, 55(3): 031008.

[4] Wei Y C, Zhao Y. A review on image semantic segmentation based on DCNN[J]. Journal of Beijing Jiaotong University, 2016, 40(4): 82-91.
魏云超, 赵耀. 基于 DCNN 的图像语义分割综述[J]. 北京交通大学学报, 2016, 40(4): 82-91.

[5] Lu Y F, Jin Q H, Jing J, et al. Detection and segmentation algorithm for bioresorbable vascular scaffolds struts based on machine learning[J]. Acta Optica Sinica, 2018, 38(2): 0215005.
鲁逸峰, 金琴花, 荆晶, 等. 基于机器学习的可降解支架检测与分割算法[J]. 光学学报, 2018, 38(2): 0215005.

[6] 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.

[7] 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.

[8] 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.

[9] Peng C, Zhang X Y, Yu G, et al. Large kernel matters: improve semantic segmentation by global convolutional network[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, Hawaii, USA. New York: IEEE, 2017: 4353-4361.

[10] Brostow G J, Shotton J, Fauqueur J, et al. Segmentation and recognition using structure from motion point clouds [M]//Forsyth D, Torr P, Zisserman A. Computer vision-ECCV 2008. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2008, 5302: 44-57.

[11] Wu Z F, Shen C H, van den Hengel A. Real-time semantic image segmentation via spatial sparsity[J/OL]. (2017-12-01)[2019-04-03]. https://arxiv.org/abs/1712.00213.

[12] Ioffe S, Szegedy C. Batch normalization: accelerating

deep network training by reducing internal covariate shift[J/OL]. (2015-05-02)[2019-04-03]. https://arxiv.org/abs/1502.03167.

[13] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, April 11-13, 2011, Fort Lauderdale, USA. USA: MIT Press, 2011: 315-323.

[14] 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, USA. New York: IEEE, 2017: 6230-6239.

[15] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1800-1807.

[16] Sifre L, Mallat S. Rotation, scaling and deformation invariant scattering for texture discrimination[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 1233-1240.

[17] Li W Y, Wang P, Qiao H. A survey of visual attention based methods for object tracking[J]. Acta Automatica Sinica, 2014, 40(4): 561-576.
黎万义, 王鹏, 乔红. 引入视觉注意机制的目标跟踪方法综述[J]. 自动化学报, 2014, 40(4): 561-576.

[18] Paszke A, Chaurasia A, Kim S, et al. ENet: a deep neural network architecture for real-time semantic segmentation[J/OL]. (2016-07-07)[2019-04-03]. https://arxiv.org/abs/1606.02147.

[19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems 25 (NIPS 2012), December 3-6, 2012, Lake Tahoe, Nevada, United States. Canada: NIPS, 2012: 1097-1105.