曾永煌, 张孙杰. 基于多分支结构和注意力机制的实时语义分割网络[J]. 智能计算机与应用,2024,14(5):107-114. DOI:10.20169/j.issn.2095-2163.240514

基于多分支结构和注意力机制的实时语义分割网络

曾永煌, 张孙杰
(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: 在实时语义分割方法研究中,由于目标感受野有限,目前仍然存在大目标分割不准确和细节信息丢失的问题。针对这个问题,提出一种基于多分支结构和注意力机制的实时语义分割算法。首先,本文构建多分支结构的细节路径以保留多尺度细节信息,减少小目标细节丢失;其次,设计空洞金字塔分支扩大感受野,以覆盖视野内大目标,进一步丰富上下文信息;最后,提出双边注意力特征融合模块,以增强特征融合时对关键特征的选择,弥补小目标信息的缺失。在 Cityscapes 测试集、CamVid 测试集所提模型的平均交并比(*mIoU*)为 74.6%与 73.6%,每秒传输帧数(Frames Per Second, *FPS*)为 94 与 74;较于 BiSeNet,本文算法的 *mIoU* 分别提高了 6.2、8.0 个百分点。实验结果表明,本文算法在实时性和准确性方面获得了很好的平衡。

关键词: 实时语义分割; 多分支结构; 注意力机制; 特征融合
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)05-0107-08

Real-time semantic segmentation network based on multi-branch structure and attention mechanism

ZENG Yonghuang, ZHANG Sunjie
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: To address the problem of limited receptive fields in current real-time semantic segmentation methods leading to inaccurate segmentation of large objects and loss of detail information, this paper proposes a real-time semantic segmentation algorithm based on multi-branch structure and attention mechanism. First of all, design detail path of multiple branch structures to preserve multi-scale detail information and reduces the loss of small target details; Secondly, design the atrous pyramid branch to expand the receptive field and cover large targets within the field-of-view, thereby enriching context information; Finally, design a bilateral attention feature fusion module to enhance the selection of key channels during feature fusion and compensate for the missing of small target information. Experimental results on Cityscapes test set and CamVid test set show that the mean Intersection over Union (*mIoU*) of the proposed model is 74.6% and 73.6%, Frames Per Second (*FPS*) is 94 and 74. In comparison with BiSeNet, *mIoU* of the proposed model is increased by 6.2 and 8.0 percentage points respectively. Experimental results show that the algorithm proposed in this paper has achieved a good balance between real-time performance and accuracy.

Key words: real-time semantic segmentation; multi-branch structure; attention mechanism; feature fusion

0 引 言

语义分割(semantic segmentation)是计算机视觉领域的一项重要技术,其主要思想是将图像分割成几种不同语义类别的像素区域并识别出每个区域的类别,最后做图像的像素级分类获得语义分割预测图。图像语义分割在自动驾驶、医学影像分析、智能交通、机器人、视频监控等领域均有广泛的应用[1]。

随着人工智能研究领域的全面发展,深度学习理论在图像语义分割任务中取得了可观成就。全卷积网络(Fully Convolutional Network, FCN)[2]首次将深度学习理论引入到语义分割任务中,使用反卷积层替换传统卷积网络中的全连接层,做像素级的

基金项目: 国家自然科学基金(61673276,61603255)。
作者简介: 曾永煌(1999-),男,硕士研究生,主要研究方向:计算机视觉,图像分割。
通讯作者: 张孙杰(1988-),男,博士,副教授,主要研究方向:非线性控制,图像处理。Email:zhang_sunjie@126.com
收稿日期: 2023-07-06

密集预测。U-Net[3]网络采用 U 型对称结构,并将编码器提取的特征图与解码器提取的特征图进行多层级特征融合。虽然 U-Net 分割精度高,但其对称结构计算量较大,而且空间信息缺失较多。为提取更多空间信息,DeepLab[4]网络引入膨胀卷积(Dilated Convolution)在不增加计算量的同时获得更大的感受野,以提取更丰富的上下文信息,但对小目标分割效果欠佳。在此基础上,DeepLabV3[5]网络引入多层级空洞卷积来提取多尺度特征图,并采用金字塔池化对多尺度特征进行融合。但由于空洞卷积的特殊性,空间信息缺失问题一直存在。混合空洞卷积网络[6]解决了空洞卷积出现的空间信息缺失的问题。然而,上述网络的参数量与计算量较大,难以满足实际应用中嵌入式设备对实时性的要求。

为了满足语义分割算法在现实场景下的部署需求,已经出现许多轻量级的网络权衡了语义分割的准确性和实时性。ENet[7]网络遵循 UNet 网络编码器-解码器的设计理念,但 ENet 利用不对称的编码器-解码器结构,减少解码器的内存成本,并利用早期下采样提高计算速度,ENet 最先对实时语义进行分割;双边分割网络(Bilateral Segmentation Network,BiSeNet)[8]为实时语义分割提供了新的思路,其主要思想是利用浅层网络的细节路径,以获取带空间位置信息的低维特征图;再用轻量化网络快速下采样提取深层次的带语义信息的高维特征图,最后将两者融合得到高级的高分辨率特征图。此后的研究表明,双分支结构方法是实时语义分割领域的高效方法。

但是双分支结构的网络采用轻量化的模型作为主干网络(如 Xception[9]、MobileNet[10]等),这些模型的卷积层数不够深,导致模型整体的感受野有限,难以覆盖比较大的特征对象,如建筑物、公交车和卡车等,会影响最终的分割精度。此外保留空间信息的空间路径设计过于简单,空间位置信息不够丰富,细节信息不足,并且双分支结构的空间路径和语义路径的特征维度相差较大,空间路径提取的是边界、位置等低层次信息,语义路径提取的是比较具体的带上下文的深层次信息,低层次的细节信息会产生较大的噪声,缺失语义信息,融合时低层次产生的噪声会影响深层次提取到的语义信息,导致最终的分割效果欠佳。

针对上述问题,本文提出一种基于多分支结构和注意力机制的实时分割网络,主要工作如下:

(1)重新设计细节路径,利用深度可分离卷积和注意力机制的结合设计一个多分支结构的细节路径;

(2)构建一个轻量化的空洞金字塔结构,使语义路径获得丰富的上下文信息的同时收获大的感受野;

(3)引入双边注意力特征融合模块,在特征融合中加入不同的注意力强化模型对关键特征的识别,提高模型的分割效果。在公开数据集 CamVid 和 Cityscapes 上的实验表明,本文模型兼顾了分割精度和推理速度。

1 网络模型与结构

1.1 网络整体结构

本文提出了基于多分支结构和注意力机制的实时语义分割网络,如图 1 所示,由一个提取高级语义特征的语义路径(Semantic Path,SP)和一个提取空间细节信息的细节路径(Detail Path,DP)构成的双路径实时语义分割网络。

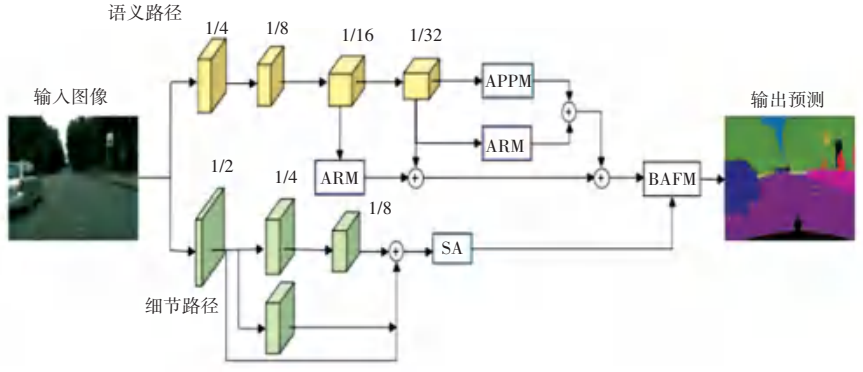


图 1 网络整体结构图

Fig. 1 Overall network architecture diagram

细节路径采用的是一个仅采样 3 次的多分支结构，保留更多不同尺度的空间细节信息，并引入空间注意力机制加强细节特征，有利于对杆、交通信号灯和交通指示牌等小目标的精准分割；语义路径采用预训练的轻量化模型（STDC）[11]快速下采样提取上下文语义信息，输入到空洞金字塔模块（Atrous pyramid pooling module，APPM）中，再使用注意力细化模块[8]（attention refinement module，ARM）细化最后 2 个阶段的输出特征；空洞金字塔从语义分支中提取不同尺度的深层特征，并用扩张卷积增大感受野，有利于覆盖视野内的建筑、公交车等大目标。细节路径保留多尺度空间细节信息，语义路径则提取丰富的语义信息，而空洞金字塔提供更大的感受野和更深层次的特征。最后，通过双边注意力特征融合模块（Bilateral attention feature fusion module，BAFM）来指导模型融合这些特征，提高模型融合时的特征选择能力，以进行最后的预测。该方法既具有实时性，又能保证高分割精度。

1.2　语义路径

空间路径编码丰富的空间信息，而语义路径旨在提供足够的感受野。在语义分割任务中，感受野对性能具有重要意义。语义路径利用轻量级模型和空洞金字塔来提供多尺度信息和较大的感受野。在语义路径中使用短期密集级联模块（Short－Term Dense Concatenate module，STDC module）作为主干网络，该网络有以下 2 个优点：

（1）通道数逐级减少，极大地减少计算复杂性；

（2）STDC 模块的最后输出是连接了所有块，保留了可变感受野和多尺度信息。

STDC 各层输出见表 1。本文在轻量级模型的尾部添加一个空洞金字塔模块，可以提供多尺度信息和大的感受野，并使用 U 型结构来融合最后 2 个阶段的特征，是一个非完整的 U 型结构。

表 1　STDC 网络各阶段输出

Table 1　Outputs of STDC each layer

stage	输出通道数	输出分辨率
输入	3	512×512
stage 1	64	256×256
stage 2	256	128×128
stage 3	512	64×64
stage 4	1 024	32×32

在深层次网络中，感受野的大小体现了模型获得上下文信息的能力。轻量化网络 STDC 的网络层数相比其他卷积网络的卷积层数不够深，提取的语义信息有限，所以本文使用空洞金字塔模块提取不同尺度的深层特征进行融合，如图 2 所示。由图 2 可看到，输入特征图，分成 4 个分支，前 3 个分支先使用 3×3 的深度可分离卷积进行再次下采样，获取更深层的信息，接着又经过空洞率分别为 4、8、12 的空洞卷积，提供比普通卷积更大的感受野，能更好地识别如建筑物、公共汽车等大目标，最后一个分支通过全局池化获取全局最大的感受野。在前述处理后通过 Concat 操作进行特征拼接，通过注意力细化模块（ARM）细化特征，最终经过 1×1 的卷积降维得到输出。

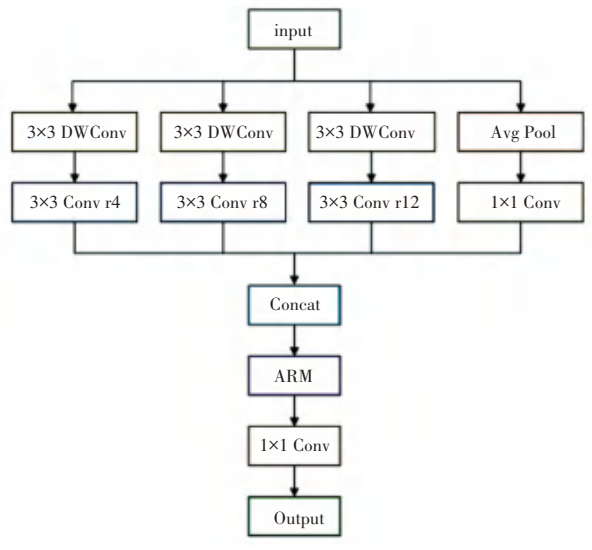


图 2　空洞金字塔模块

Fig. 2　Atrous pyramid pooling module

1.3　细节路径

在语义分割任务中，图像的空间信息（细节信息）对于预测输出至关重要。现有许多方法来编码更丰富的细节信息，比如使用膨胀卷积以保持特征图的空间大小并获得大的感受野，但由于膨胀卷积的卷积核是有间隔的，会导致关键细节信息的遗失；或者利用大的卷积核编码丰富的细节信息与边界信息，但是大的卷积核会带来巨大的计算量，所以为保留丰富的细节信息、且不增加过多的计算量，本文提出了细节路径来提取丰富的细节信息。

细节路径如图 3 所示。由图 3 可看到，本文提出了一条可以编码空间细节信息的特征提取路径，利用浅层网络，提取了较高分辨率的特征图，保留了更多的空间细节，有利于对像素信息少的小目标与边界精确分割。首先使用 7×7 的普通卷积进行 1 次下采样，然后分成 3 个分支。第 1 个分支通过 2 个3×3的深度可分离卷积代替标准卷积来减少计算

量和参数量,最后使用 1×1 卷积对上一步输出的特征图做通道的升降维;第 2 个分支用深度可分离卷积和最大池化进行下采样,最大池化可以保留纹理(边界)信息;第 3 个分支用最大池化进行 2 次下采样,在此基础上进行特征融合,并将融合后的特征图输入到空间注意力模块(Spatial Attention,SA)中。

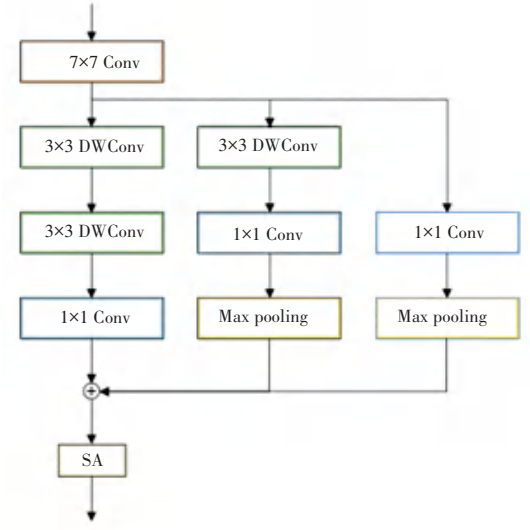


图 3 细节路径

Fig. 3 Detail path

空间注意力(Spatial Attention, SA)让模型更加关注空间位置信息,更加突出关键特征的位置,增强特征表达能力,在分割任务中起着重要作用。空间注意力模块如图 4 所示。由图 4 可以看到,对经过细节路径得到的特征图 F_d 在通道维度上分别通过最大池化和平均池化得到特征图 F_m 和 F_a, 接着利用 Concat 操作将这 2 个特征图进行通道的拼接,再使用 7×7 卷积、*Sigmoid* 激活函数获取到空间注意力特征图 F_{cs}, 最后让 F_{cs} 与输入特征 F_d 相乘得到具有空间位置关注的特征图 F_s。

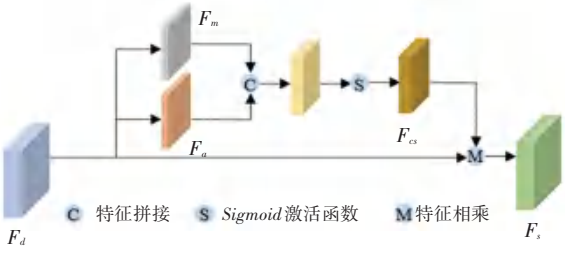


图 4 空间注意力模块

Fig. 4 Spatial attention module

1.4 双边注意力特征融合模块

本文模型是采用双路径提取特征,2 条路径的特征在特征层次上的差异是比较大的,不能进行简单的特征融合。空间路径提取浅层丰富的细节信息,语义路径提取深层次的语义信息。也就是说,细节路径的输出特征是低维的,而语义路径的输出特征是高维的,简单地融合低级和高级特征带来的收益并不高,虽然低层特征含有丰富的空间位置信息,有利于模型识别小目标的细节信息,但却具有大量的噪声,容易干扰深层特征中的语义信息。因此给出的联合注意力特征融合模块如图 5 所示。由图 5 可看到,本文提出了一个特定的双边注意力特征融合模块(Bilateral Attention Feature Fusion Module, BAFM)来指导模型融合这些特征,提高模型融合时的特征选择能力,使模型能够自主地学习各通道特征的重要性,并用于指导深层特征与浅层特征的融合。

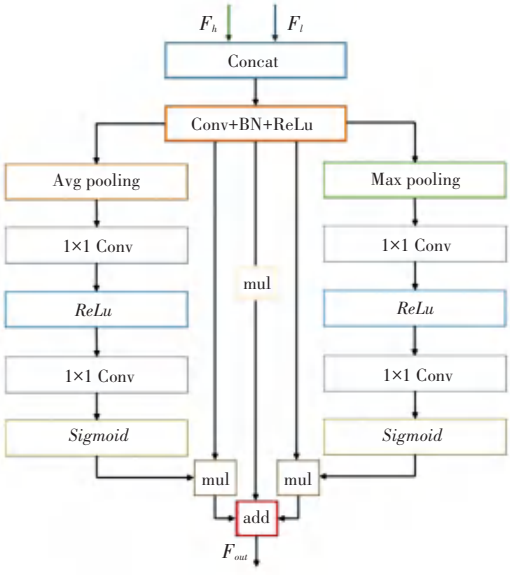


图 5 联合注意力特征融合模块

Fig. 5 Bilateral attention feature fusion module

BAFM 由 2 个轻量化的通道注意力模块组成,通过获取不同的全局的语义信息,选择性地融合有关键空间细节信息的低级特征和深层语义信息的高级特征。首先,输入为语义路径的高维特征 F_h 与细节路径的低维特征 F_l,然后将 2 种不同的特征直接拼接得到 F_{cat},接着通过卷积、批量归一化和激活函数得到 F_C; 再按通道维度分别对输入特征使用全局平均池化和全局最大池化进行特征压缩,使特征通道变为一个特征张量;然后,利用 1×1 卷积代替原 SENet 模块中的全连接层为各特征通道生成权重,可以减少计算量和参数量,通过 *Sigmoid* 激活函数将值限定在 0 与 1 之间,得到权重 θ_1 和 θ_2。 其次,把各特征通道的输出权重作为每个特征通道的重要程度,再用乘法逐通道分别加权到特征 F_C 上,分别得到带注意力的特征 F_A 和 F_M,最后将各特征叠加

得到最终的输出 F_{out}。相应的计算过程分别如下：

$$F_C = Conv(Concat(F_l, F_h)) \tag{1}$$

$$\theta_1 = Sigmoid(Conv(Avg(F_C))) \tag{2}$$

$$\theta_2 = Sigmoid(Conv(Max(F_C))) \tag{3}$$

$$F_{\text{out}} = F_C + \theta_1 F_C + \theta_2 F_C \tag{4}$$

2　实验结果与分析

2.1　实验设置

本文实验显卡型号为 NVIDIA RTX 3080Ti，采用随机梯度下降（SGD）算法优化模型，其中批处理的大小为 16，初始动量为 0.9，权重衰减为 0.000 5。采用 poly 的学习率调整策略在网络训练过程中动态调整学习率，该策略的运算公式为：

$$lr_{\text{cur}} = lr_{\text{init}} \times \left(1 - \frac{epoch}{max\_epoch}\right)^{\text{power}} \tag{5}$$

其中，lr_{cur} 表示当前学习率；lr_{init} 表示初始学习率；*epoch* 表示当前迭代次数；*max_epoch* 表示最大迭代次数；*power* 设置为 0.9；初始学习率为 0.005。

在 Cityscapes 数据集中，为提升模型的泛化能力，实验中采用随机水平翻转、随机缩放等数据增强策略，且随机缩放因子设置为{0.75，1.0，1.25，1.75，2.0}，将图像转换为不同的尺度，而后随机剪裁图片尺寸至 1 024 ´ 1 024 用于训练。在 CamVid 数据集上训练网络时的实验设置与在 Cityscapes 数据集上训练网络模型的实验设置相同，批处理大小变为 12。

2.2　评价指标

本文实验使用平均交并比（mean Intersection-over- Union，*mIoU*）和每秒处理帧数（Frames Per Second，*FPS*）作为评估算法语义分割精度和推理速度的评价指标。对此，研究给出阐释分述如下。

（1）*mIoU*。是真实值和预测值集合的交集与并集之比，可用于评价算法的分割能力，*IoU*（Intersection over Union）是每一个类别的交集与并集之比，而 *mIoU* 是所有类别的平均 *IoU*，其值可由式（6）计算求出：

$$f_{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{P_{ij}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \tag{6}$$

其中，k 表示前景对象的个数，P_{ij} 表示原本属于第 i 类、却分类到第 j 类的像素的数量。

（2）*FPS*。用于评价算法速度。其值可由式（7）计算求出：

$$f_{FPS} = \frac{N}{\sum_{j=1}^{N} T_j} \tag{7}$$

其中，N 表示图像数量，T_j 表示算法处理第 j 幅图像的时间。

2.3　CamVid 数据集实验结果

本文在 CamVid 公共数据集上进行对比验证。CamVid 公共数据集是通过自动驾驶汽车的视角来截取街道场景的数据集，数据集大小适中，标注精度高。该数据集是从 5 个视频片段中抽帧截取出来的城市街道场景分割数据集，比较符合对实时语义分割算法在实时场景的有效性的验证，并且在自动驾驶场景中增加了观察对象的类别。该数据集共包含 701 幅图片，其中 367 幅图像用于训练、101 幅图像用于验证、233 幅图像用于测试，图像分辨率均为 720×960，共包含 11 个语义类别。

首先，对本文语义路径的骨干网络进行实验，以验证替换骨干网络后语义路径的性能。图像分类与目标检测的高效骨干网络可以帮助图像分割快速地构建相应的分割骨干网络，但用于分类的主干网络用在图像分割任务上效果有限，所以本文使用了 STDC 网络作为本文语义路径的主干网络。不同主干网络的性能对比实验结果见表 2。由表 2 分析可知，采用参数量、*mIoU* 和 *FPS* 作为评价标准，本文模型的骨干网络 STDC1 相比较 Xception39 来说，虽然其参数量增加了将近 3 倍，但是分割精度提升了 6.48%，得到 72.08% 的精度；主干网络 STDC2 相比较 ResNet18 而言，其参数量减少了将近 1 倍，分割精度提升得也比较明显，得到 73.32% 的精度。在对比实时语义分割模型总体性能时，精度比参数量和计算量占有更大权重。分割精度与速度的平衡保证本文模型能够在有限的计算资源的场景中高效、实时地分割图像。

表 2　不同主干网络的性能比较

Table 2　Performance comparison of different backbone networks

Backbone	参数量/MB	*MIoU*/ %	速度/FPS
Xceptin39	5.8	65.60	175
ResNet18	49.0	68.70	116
STDC1	13.8	72.08	125
STDC2	21.8	72.32	114

为了验证本文模块的性能，对各模块进行了对比实验，采用参数量（MB）、*mIoU*（%）和速度（*FPS*）为评价标准，对提出的各模块同步进行消融实验，验证各模块的性能表现，实验结果见表 3。实

验中包括了细节路径(detail path, DP)、空洞金字塔模块(Atrous pyramid pooling module,APPM)和双边注意力特征融合模块(Bilateral attention feature fusion module,BAFM)。

表 3 各组成模块消融实验结果

Table 3 Ablation experimental results of different modules

Backbone	DP	APPM	BAFM	参数/MB	*MIoU*/ %	速度/FPS
Xceptin39				5.8	65.60	175
STDC1				13.8	72.08	125
STDC1	√			14.2	72.23	114
STDC1	√	√		21.5	73.04	95
STDC1	√	√	√	22.1	73.28	86
STDC2	√	√	√	28.6	73.64	74

首先本文验证的基准模型是采用骨干网络为 ResNet18 的 BiSeNet 网络,验证时不改变基准模型中的各模块的结构,采用 STDC1 作为本文模型细节路径的骨干网络做对比试验。研究采用 BiSeNet 模型的基本结构,只替换语义路径中的骨干网络得到 72.08%的精度和 125 帧/s 的速度,然后在此基础上替换细节路径,使用深度可分离卷积和多分支结构在减少参数量的同时提取多尺度信息,并利用空间注意力模块增强细节路径的空间细节信息,得到 72.23%的精度和 114 帧/s,分割精度有一定的提升。随后在语义路径的最后阶段引入轻量的空洞金字塔模块,以获得大的感受野和多尺度信息, 提升模型对全局信息的识别,得到 73.04%的精度和 95 帧/s 的速度。接着,再引入双边注意力特征融合模块,提升模型对关键信息特征的融合,最后得到 73.28%的精度和 86 帧/s 的速度。将语义路径的主干网络替换成 STDC2 得到 73.64%的精度和 74 帧/s 的速度。上述实验表明,本文提出的所有模块对模型整体分割精度都有一定的提升。

为验证模型的优越性,在 CamVid 公共数据集上给出该算法的分割精度和推理速度。数据集输入图像为 720×960,在测试集上测试得到分割精度和推理速度。所提方法和现有方法的比较结果见表 4。对比方法中,有非实时的方法、包括 SegNet[12], DeepLab,PSPNet[13],还有实时语义分割方法、包括 ENet, ICNet[14], BiSeNet, BiSeNetv2[15], STDC, S^2-FPN。所提方法在 CamVid 测试集上得到 73.6%的精度和 74 帧/s 的推理速度,在分割精度和速度上优于非实时方法,并且在实时语义分割方法上也取得了较优的分割效果,比基准网络 BiSeNet 提升了 8%的平均交并比,比 BiSeNetV2 提升了 1.2%的平均交并比。因此,对比其他轻量级实时语义分割算法,本文模型在保证高精度的前提下仍达到了实时效果。实现分割精度和速度之间最优平衡。

表 4 不同算法在 CamVid 测试集的对比分析

Table 4 Comparative analysis of different algorithms in CamVid test set

Model	Backbone	*Speed*/ FPS	*mIoU*/ %
SegNet	VGG16	16.7	55.6
DeepLab	ResNet50	4.9	61.6
PSPNet	ResNet50	5.4	69.1
ENet	No	61.2	51.3
ICNet	PSPNet50	27.8	67.1
BiSeNet1	Xception39	175.0	65.6
BiSeNet2	ResNet18	116.0	68.7
S^2-FPN18	ResNet18	124.2	69.5
S^2-FPN34	ResNet34	107.2	71.0
BiSeNetv2	No	124.5	72.4
STDC1-Seg	STDC1	197.0	73.0
Ours	STDC1	86.0	73.2
Ours	STDC2	74.0	73.6

2.4 Cityscapes 数据集实验结果

为评估模型实时语义分割的泛化性能、分割不同城市场景的驾驶视角图像,本文在 Cityscapes 城市街景数据集上进行了相关实验。Cityscapes 公共数据集是一个截取了 50 个不同城市的大型城市街道场景图像的数据集, 常用于评估语义分割任务,该数据集包含了5 000幅精细标注的图像和20 000幅粗略标注的图像,分辨率为 1 024×2 048。在本文中只使用有精细标注的图像来进行实验,包含 2 975

幅用于训练的图像、500 幅用于验证的图像和 1 525 幅用于测试的图像。数据集输入图像为 1 024×2 048，首先将输入的分辨率调整为 1 024×1 024 进行训练，然后在验证集和测试集上恢复至原始分辨率评估分割精度。本文算法与其他方法对比结果见表 5，同样对比实时和非实时的语义分割模型。本文模型相比基准网络 BiseNet 在分割精度和速度上有明显的提升，得到 74.6%的分割精度和 94 帧/s 的速度。

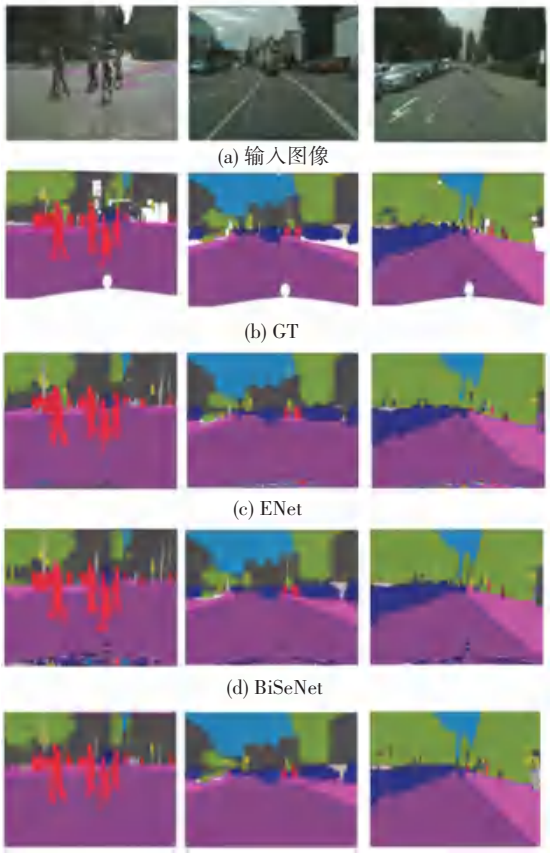
表 5　不同算法在 Cityscapes 测试集的对比分析

Table 5　Comparative analysis of different algorithms in Cityscapes test set

Model	Backbone	*Speed*/ FPS	*mIoU* /%
SegNet	VGG16	17.00	57.0
DeepLab	VGG16	0.25	63.1
PSPNet	ResNet101	0.78	81.2
ENet	No	135.40	58.3
ICNet	PSPNet50	30.30	69.5
DABNet	No	104.00	70.1
LEDNet	No	71.00	70.6
BiSeNet1	Xception	105.80	68.4
BiSeNetv2	No	156.00	72.6
STDC-Seg50	STDC1	250.40	71.9
Ours	STDC1	106.00	73.8
Ours	STDC2	94.00	74.6

图 6 为本文方法对比其他实时语义分割方法的可视化结果。由图 6 的第 1 列图像可以看出，与其他方法相比，本文方法对道路中间行人可以很清晰地分割出来；第 2 列图像中，本文方法对一些远处行人、杆等细小类别和大的目标建筑等有更出色的分割效果；第 3 列图像中，对于车辆分割存在相似类别信息干扰时，本文方法可以更精确地避免相似类别信息干扰。实验结果表明，本文提出的方法具有优秀的语义分割能力和类别判断能力。

表 6 为 SegNet、BiSeNet、DABNet 以及本文在 Cityscapes 测试集上 19 种分类的结果。分析表 6 可知，相较于 BiSeNet，本文在所有分类上都有明显优势，特别是卡车、公交车和火车等大目标提升比较明显，表明本文提出的方法能明显增大感受野；而且对细小的目标如杆、交通信号灯、交通指示牌等分割精度提升也比较明显。



(a) 输入图像　(b) GT　(c) ENet　(d) BiSeNet　(e) 本文

图 6　Cityscapes 数据集上的可视化结果

Fig. 6　Visualization results of the Cityscapes dataset

表 6　Cityscapes 测试集上各个类别的准确率

Table 6　Accuracy for each category on the Cityscapes test dataset

分割类别	SegNet	BiSeNet	DABNet	本文
道路	96.4	96.4	96.8	97.2
人行道	73.2	77.2	78.5	80.3
建筑	84.0	89.8	90.9	91.5
墙	28.4	46.5	45.3	54.1
栅栏	29.0	49.1	50.1	58.8
杆	35.7	43.4	59.1	60.2
信号灯	39.8	58.3	65.2	69.7
指示牌	45.1	66.1	70.7	76.1
植被	87.1	90.6	92.5	92.3
地形	63.8	61.2	68.1	62.4
天空	91.8	92.3	94.6	93.8
人	62.8	75.8	80.5	78.2
骑手	42.8	57.5	58.5	60.7
汽车	89.3	92.1	92.7	92.5
卡车	38.1	52.9	52.7	65.1
公交车	43.1	70.5	67.2	78.9
火车	44.1	61.7	50.9	75.2
摩托车	35.8	54.2	50.4	59.7
自行车	51.9	65.3	65.7	71.3
平均 *mIoU*	57.0	68.4	70.1	74.6

3 结束语

为权衡实时语义分割网络在精度和速度之间的平衡，本文提出了一种基于多分支结构和注意力机制的实时语义分割网络。针对目前的实时语义分割算法对细节特征信息缺失的问题，使用了多分支结构提取多尺度信息，同时结合空间注意力机制提取空间位置信息，提出了高效的细节路径；针对轻量化网络感受野不足的问题，引入了空洞金字塔模块，增大感受野的同时提取多尺度的语义信息，加强了对大目标的识别能力；最后，利用了双边注意力机制的特征融合模块，结合了空间注意力机制提高对小目标细节信息的判别能力。在 Cityscapes 和 CamVid 数据集上进行了实验，实验结果表明：本文研究能够在精度和推理速度之间取得较好的平衡，相较于其他算法而言，表现出了良好的性能。

参考文献

[1] 罗会兰，张云. 基于深度网络的图像语义分割综述[J]. 电子学报，2019，47(10)：2211-2220.

[2] LONG J，SHELHAMER E，DARRELL T. Fully convolutional networks for semantic segmentation [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2015：3431-3440.

[3] RONNEBERGER O，FISCHER P，BROX T. U-Net：Convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham：Springer，2015：234-241.

[4] 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.

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

[6] WANG Panqu，CHEN Pengfei，YUAN Ye，et al. Understanding convolution for semantic segmentation [C]//2018 IEEE Winter Conference on Applications of Computer Vision. Nevada，USA：IEEE，2018：1451-1460.

[7] PASZKE A，CHAURASIA A，KIM S，et al. Enet：A deep neural network architecture for real-time semantic segmentation [J]. arXiv preprint arXiv：1606.02147，2016.

[8] YU Changqian，WANG Jingbo，PENG Chao，et al. BiSeNet：Bilateral segmentation network for real-time semantic segmentation [C]// Proceedings of the 2018 European Conference on Computer Vision. Cham：Springer，2018：325-341.

[9] CHOLLET F. Xception：Deep learning with depthwise separable convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii，USA：IEEE，2017：1800-1807.

[10] HOWARD A G，ZHU Menglong，CHEN Bo，et al. MobileNets：Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv：1704.04861，2017.

[11] FAN Mingyuan，LAI Shenqi，HUANG Junshi，et al. Rethinking BiSeNet for real-time semantic segmentation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，USA：IEEE，2021：9716-9725.

[12] 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.

[13] ZHAO Hengshuang，SHI Jianping，QI Xiaojuan，et al. Pyramid scene parsing network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii，USA：IEEE，2017：2881-2890.

[14] ZHAO Hengshuang，QI Xiaojuan，SHEN Xiaoyong，et al.IcNet for real-time semantic segmentation on high-resolution images[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich，Germany：Springer，2018：405-420.

[15] YU Changqian，GAO Changxin，WANG Jingbo，et al.BiSeNet v2：Bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision，2021，129(11)：3051-3068.