DOI:10.16136/j.joel.2022.10.0035

基于高度有效驱动注意力与多层级特征融合的城市街景语义分割

赵 迪[1],孙 鹏[1],陈奕博[1],熊 炜[1,2,3*],刘 粤[1],李利荣[1,2]

(1.湖北工业大学 电气与电子工程学院,湖北 武汉 430068;2.襄阳湖北工业大学 产业研究院,湖北 襄阳 441100;3.美国南卡罗来纳大学 计算机科学与工程系,南卡罗来纳 哥伦比亚 29201)

摘要:针对DeepLabv3+网络在进行城市街景图像分割任务时,没有充分利用到网络中多层级特征信息,导致分割结果存在大目标有孔洞、边缘目标分割不够精细等不足;并且考虑到城市街景数据具有天然的空间位置特殊性,本文提出在DeepLabv3+网络的基础上引入高度有效驱动注意力机制(height-driven efficient attention model,HEAM)与多层级特征融合模块(multi-stage feature fusion model,MFFM),将HEAM嵌入特征提取网络与空洞空间金字塔池化(atrous spatial pyramid pooling,ASPP)结构中,使其对目标关注更多垂直方向上的空间位置信息;MFFM通过融合多层特征图,在网络中形成多条融合支路依次连接到网络解码端,采用逐次上采样提高解码时像素上的连续性。将改进的网络通过CamVid城市街景数据集验证测试,实验结果表明,该网络能有效改善DeepLabv3+的不足,并且合理运用了数据集的位置先验性,增强了分割效果,在CamVid测试集上平均交并比(mean intersection over union,*MIoU*)达到了68.2%。

关键词:DeepLabv3+;城市街景;注意力机制;语义分割;特征融合

中图分类号:TP391 **文献标识码**:A **文章编号**:1005-0086(2022)10-1038-09

Urban street view semantic segmentation based on height-driven effective attention and multi-stage feature fusion

ZHAO Di[1], SUN Peng[1], CHEN Yibo[1], XIONG Wei[1,2,3*], LIU Yue[1], LI Lirong[1,2]

(1. School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, Hubei 430068, China; 2. Xiangyang Industrial Research Institute, Hubei University of Technology, Xiangyang, Hubei 441003, China; 3. Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201, USA)

Abstract: Deeplabv3+ network does not make full use of multi-stage feature information in urban street view image segmentation, which leads to the shortcomings of large targets with holes, imprecise segmentation of edge target and so on. Considering the natural spatial position particularity of urban street view data, this paper proposes to introduce a height-driven effective attention model (HEAM) and a multi-stage feature fusion model (MFFM) on the basis of Deeplabv3+ network, and it embeds HEAM into the feature extraction network and atrous spatial pyramid pooling (ASPP) structure, which makes it pay attention to more spatial position information in the vertical direction. MFFM integrates multi-layer feature images to form multiple branches in the network and connect them to the network decoding end in turn. Successive up-sampling is used to improve the continuity of pixels during decoding. The improved network is verified and tested by CamVid urban street view data set. The results show that the network can effectively improve the deficiency of DeepLabv3+, and the location priori of the data set is properly used to enhance the segmentation effect. Mean intersection over union (*MIoU*) on CamVid test set reaches 68.2%.

* **E-mail**: xw@mail.hbut.edu.cn

收稿日期:2022-01-15 **修订日期**:2022-03-03

基金项目:国家自然科学基金(61571182,61601177)、国家留学基金(201808420418)、湖北省自然科学基金(2019CFB530)、湖北省科技厅重大专项(2019ZYYD020)和襄阳湖北工业大学产业研究院科研项目(XYYJ2022C05)和资助项目

Key words: DeepLabv3+; urban street view; attention mechanism; semantic segmentation; feature fusion

1 引　言

城市场景分割作为一种基础且又具挑战性的语义分割任务,其目的是将场景目标细分并深度解析为与语义类别信息相关的不同区域,近年来城市街道场景分割算法已经广泛应用到计算机视觉自动驾驶任务中,并且取得了显著成果。全卷积神经网络(fully convolutional networks, FCN)[1]作为分割任务的开山之作,它将卷积神经网络(convolutional neural networks, CNN)的最后一层全连接层用卷积层替代,使各种尺寸大小的图像输入网络后能输出同样大小的分割图像,进而实现网络端到端的图像分割方法。但 FCN 仅使用深层特征对像素进行分类,而相对具有丰富空间信息的浅层特征并没有被充分利用,导致最终的分割结果较为粗糙。随后,如 U-Net 和 SegNet 等[2]更注重浅层特征的分割网络模型大量涌现,此类模型均采用编码-解码结构,编码端通过卷积网络对特征进行深度提取,而解码端用于将提取到的深层特征逐步恢复至图像原始尺寸大小,同时为弥补细节信息丢失,会将部分浅层特征通过跳跃连接与解码端合并。由于空洞卷积(atrous convolution)能汇集上下文信息,DeepLab 系列分割网络均采用了空洞卷积的方法,其中 DeepLabv1 通过增大特征图感受野的方法替代下采样操作,获得具有深层语义信息的稠密特征图,在图像的后处理过程中,采用全连接条件随机场(conditional random field, CRF),在全局信息上做类似平滑处理,以提高分割精度。DeepLabv2[3]网络改进 SSP-Net 网络,新提出空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)代替原先做预处理 resize 方法。DeepLabv3[4]网络由于在 ASPP 模块中融合了相关位置信息,故而在图像后处理阶段舍弃了全连接 CRF 处理。最终的 DeepLabv3+[5]网络采用编码-解码结构,将 V3 网络解码端中最后预测分类的 8 倍上采样细化为两次 4 倍上采样,并有效融合特征提取网络的浅层特征信息,进而增强了网络分割能力。但 DeepLabv3+网络依然存在分割大目标有孔洞、边缘目标分割精度低、个别目标分割不够精细等问题。

由于城市街景图像的独特拍摄角度,街景图像具有特定的位置相关性,将城市街景数据集 CamVid 按类别分布排序,同时将图像按垂直方向做类分布排序,会发现占比较大的几个类别在垂直空间位置有很强的依赖性,如顶部区域主要存在类别 0(sky)和类别 1(building);中部区域主要存在类别 1(building)、类别 8(car)和类别 3(road)和一些小目标类别;底部区域主要存在类别 3(road)和类别 8(car)。

鉴于 DeepLabv3+网络存在的一些不足,且为了更好地利用城市数据的空间位置先验性和高度上下文信息,本文基于 DeepLabv3+网络采用 HANet[6]进行适当改进,提出了高度有效驱动注意力机制(height-driven efficient attention model, HEAM)与多层级特征融合模块(multi-stage feature fusion model, MFFM)两个新模块,并在 CamVid 数据集验证其分割效果。

2 基本原理

2.1 DeepLabv3+网络介绍

DeepLabv3+网络是标准的编码-解码结构。如图 1所示,在编码端,输入图像进入骨干网络完成特

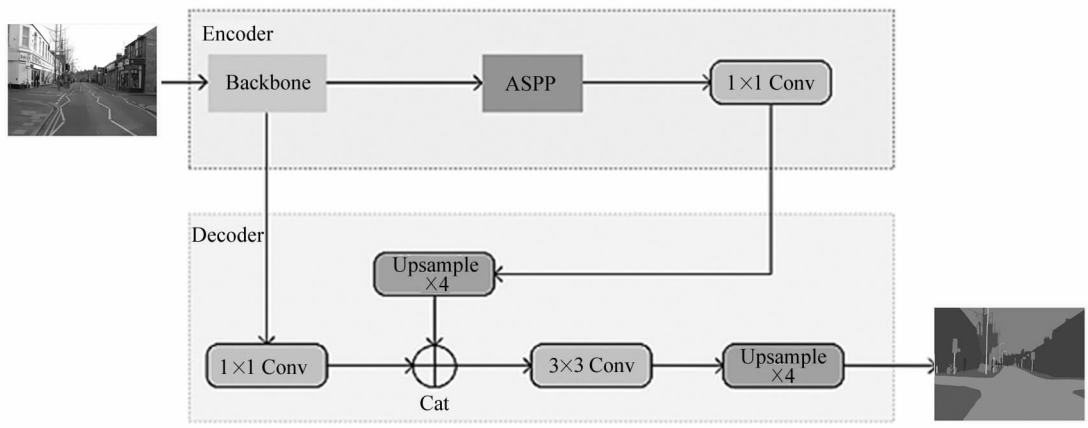

图 1　DeepLabv3+网络模型

Fig. 1　DeepLabv3+ network

征提取后，获得原始图像1/16大小的特征图，然后经过 ASPP 模块后会获得 5 个不同尺度的特征图，将 5 个特征图在通道维度上拼接成的新的特征图，此特征图将获得更大的感受野。最后再使用 1×1 卷积将特征图通道压缩为 256 并送入解码端。解码端一般通过上采样将编码后的特征图逐次恢复到原始图像尺寸大小，实现端到端的语义分割方法。在 DeepLabv3＋网络中，为了弥补下采样导致的目标边界轮廓信息的丢失，在解码端中融合骨干网络第一层的浅层特征信息，并且采用双线性插值进一步恢复图像。

2.2　改进的 DeepLabv3＋网络

注意力机制在深度学习各个领域应用都表现的十分出色，在图像处理方面，利用注意力机制使网络更加关注目标的重要特征，而忽略掉不相关的信息，进而建立两个像素之间的深度依赖关系。结合城市街景图像的位置先验性，在改进的网络中加入 HEAM，从而加强网络在不同高度位置上对目标的特征提取能力。本文将 HEAM 嵌入特征提取网络与 ASPP 结构中，使网络在做深层卷积和通过 ASPP 做多层空洞卷积时，提升对深层特征和多尺度特征的提取能力。

对于多目标语义分割任务，不同目标物体具有不同尺寸大小，如果仅使用同一层特征进行分割，难以完成对多目标的精准分割。浅层特征往往意味着分辨率更高、拥有丰富的细节和空间信息，而深层特征分辨率更低、对细节感应较弱，但拥有更强的整体语义信息。因此，经常将两者特征进行深度融合，促使特征具有细节和语义信息，以提升特征的完整信息表达。如图 2 所示，在改进的 DeepLabv3＋网络中，加入 MFFM，从骨干网络 Resnet50[7] 中分别抽取各层特征作为特征融合支路。支路一，通过通道拼接第三层与第四层特征，然后与第二层特征融合，由于特征图尺寸大小不同，需要对其进行 FPN（feature pyramid networks）[8] 多尺度特征融合，最后与解码块中第一个 2 倍上采样后的特征做跳跃拼接。支路二，支路一完成拼接经过一个 2 倍上采样操作后，从骨干网络中抽取第一层和第二层特征做改进的 FPN 融合操作，进而引出另一条支路跳跃连接到第二个 2 倍上采样特征。改进后的网络将原网络中的 4 倍上采样细化为两个 2 倍上采样操作，并在每个上采样操作后跳跃连接骨干网络中多阶段的深层特征，有效地改善了解码块图像恢复过程信息的丢失。

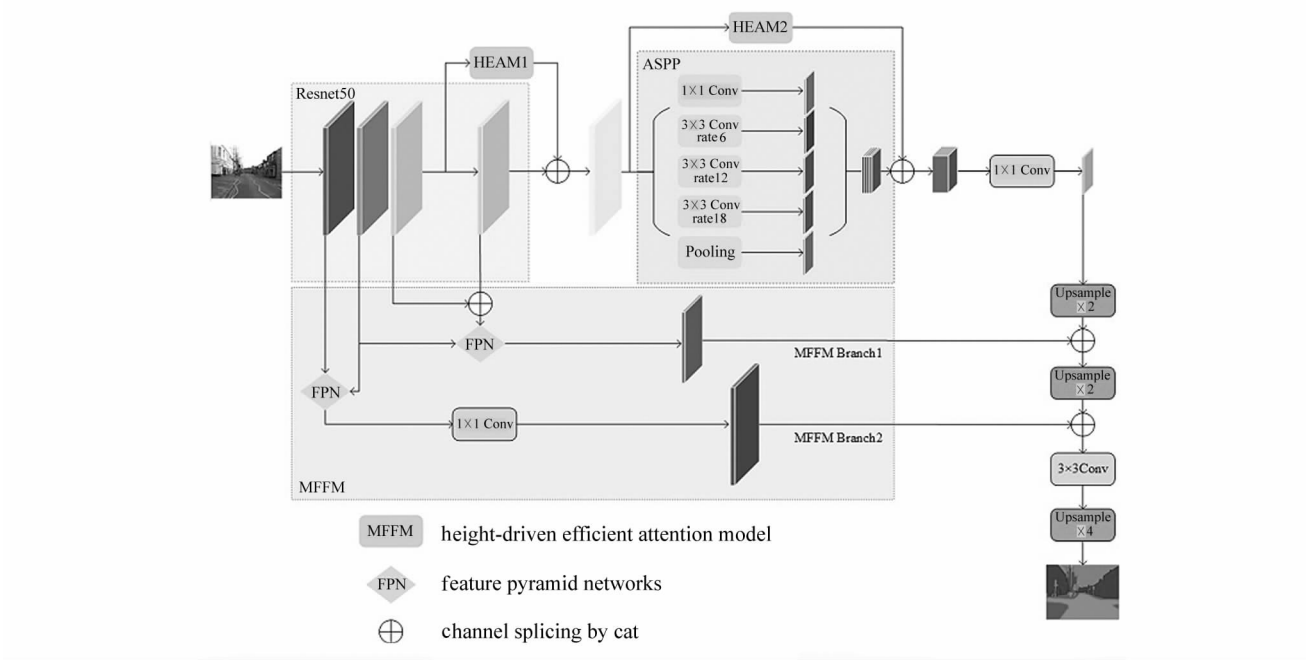

图 2　改进的基于 HEAM 与 MFFM 的 DeepLabv3＋网络

Fig. 2　Improved DeepLabv3＋ network based on height-driven effective attention model and multi-stage feature fusion

2.2.1　HEAM

HEAM 通过压缩宽度维度可以从其上下文信息中生成逐通道的缩放因子，以获得逐通道高度上的权重大小。如图 3 所示，$\boldsymbol{X}_l \in \mathscr{R}^{C_l \times H_l \times W_l}$ 和 $\widetilde{\boldsymbol{X}}_h \in \mathscr{R}^{C_h \times H_h \times W_h}$ 分别代表主网络中的浅层特征图和深层特征图，其中 C 是通道数，H 和 W 分别是特征图的高

度和宽度大小。通过 HANet 产生一个由逐通道高度缩放因子组成的通道注意力图 $\boldsymbol{A}\in\mathscr{R}^{C_h\times H_h}$，$\dot{\boldsymbol{X}}_h$ 可以将深层特征图 $\boldsymbol{X}_l$ 和注意力图 $\boldsymbol{A}$ 逐元素乘法获得，$\ddot{\boldsymbol{X}}_h$ 由深层特征图通过 ECA-Net[9] 生成，最后将 $\dot{\boldsymbol{X}}_h$ 和 $\ddot{\boldsymbol{X}}_h$ 融合生成 $\widetilde{\boldsymbol{X}}_h$，如式(1)、(2)、(3)所示：

$$\dot{\boldsymbol{X}}_h = F_{\mathrm{HANet}}(\boldsymbol{X}_l)\odot\boldsymbol{X}_h = \boldsymbol{A}\odot\boldsymbol{X}_h, \tag{1}$$

$$\ddot{\boldsymbol{X}}_h = F_{ECAM}(\boldsymbol{X}_h), \tag{2}$$

$$\widetilde{\boldsymbol{X}}_h = \dot{\boldsymbol{X}}_h \oplus \ddot{\boldsymbol{X}}_h, \tag{3}$$

式中，$\odot$ 为逐元素相乘，$\oplus$ 为逐元素相加，F_{HANet} 具体步骤如下：(a) 宽度池化：首先将特征图 $\boldsymbol{X}_l\in\mathscr{R}^{C_l\times H_l\times W_l}$ 压缩宽度维度生成特征图 $\boldsymbol{Z}\in\mathscr{R}^{C_l\times H_l\times 1}$，以获得每行的高度上下文信息，如式(4)所示：

$$\boldsymbol{Z} = G_{\mathrm{pool}}(\boldsymbol{X}_l), \tag{4}$$

(b)和(d)插值处理：由于城市街景图像在高度方向上分布有很大的差异性，不需要考虑矩阵 $\boldsymbol{Z}$ 的所有行信息，因此通过下采样对 $\boldsymbol{Z}$ 插值产生特征图 $\hat{\boldsymbol{Z}}\in\mathscr{R}^{C_l\times\hat{H}\times 1}$，本文中令超参数 $\hat{H}=16$，之后(d)步骤再上采样恢复到维度为 $C_l\times H_l\times 1$。(c)高驱动注意力图计算：以特征图 $\hat{\boldsymbol{Z}}$ 作为输入，采用卷积操作来产生注意力图，相比于使用全连接层，卷积层能更好地考虑相邻行之间的关系。由 N 个卷积层得到注意力图 $\boldsymbol{A}$ 可以用式(5)表示：

$$\boldsymbol{A} = G_{\mathrm{up}}(\sigma(G^N_{\mathrm{Conv}}(\cdots\delta(G^1_{\mathrm{Conv}}(\hat{\boldsymbol{Z}}))))), \tag{5}$$

式中，σ 表示 Sigmoid 函数，δ 表示 ReLU 激活函数，G^i_{Conv} 表示第 i 个一维卷积层。本文中令超参数 $N=3$，有 3 个卷积层操作。第一层卷积将通道压缩 r 倍 $G^1_{\mathrm{Conv}}(\hat{\boldsymbol{Z}})=\boldsymbol{Q}^1\in\mathscr{R}^{\frac{C_l}{r}\times\hat{H}}$，第二层卷积将通道拉伸 2 倍 $G^2_{\mathrm{Conv}}(\delta(\boldsymbol{Q}^1))=\boldsymbol{Q}^2\in\mathscr{R}^{2\times\frac{C_l}{r}\times\hat{H}}$，最后一层卷积将通道恢复到 C_h，$G^3_{\mathrm{Conv}}(\delta(\boldsymbol{Q}^2))=\hat{\boldsymbol{A}}\in\mathscr{R}^{C_h\times\hat{H}}$。(e)位置编码：由于人在驾驶观察时对物体具有位置先验知识，启发于此，本文在中间层特征图 Q^i 加入正弦位置编码[10]，位置编码可定义为式(6)、(7)：

$$PE_{(p,2i)} = \sin(p/100^{2i/C}), \tag{6}$$

$$PE_{(p,2i+1)} = \cos(p/100^{2i/C}), \tag{7}$$

式中，p 代表整张图垂直方向上的位置因子，i 为垂直位置的数量，令 $i=\hat{H}$。新特征图 $\widetilde{\boldsymbol{Q}}$ 由式(8)产生：

$$\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}\oplus PE, \tag{8}$$

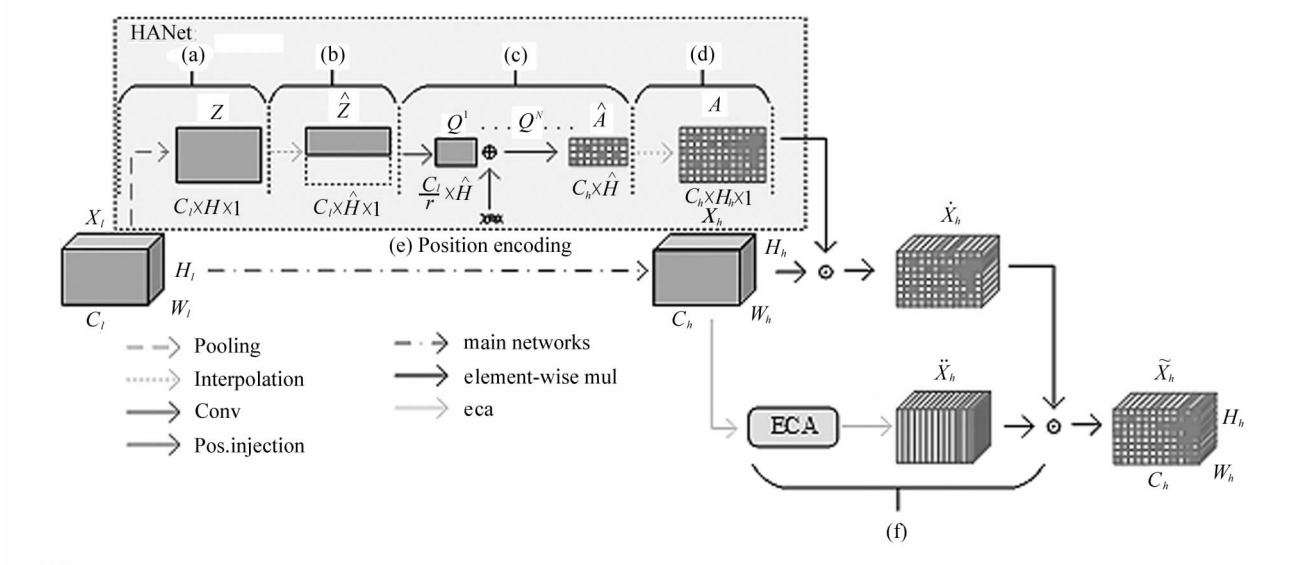

图 3　HEAM

Fig. 3　Height-drive effective attention module

2.2.2　改进的特征金字塔网络

浅层特征对应于图像的局部信息，利用丰富的局部信息可以区分简单目标；深层特征对应于图像的全局信息，利用颜色、纹理和形状等全局信息有助于对更精细的复杂目标进行区分。特征金字塔网络可以融合骨干网络多层级特征，实现对不同大小的多目标分割。改进的金字塔网络如图 4 所示，自下而上分别是输入图像尺寸 1/4、1/8 和 1/16 大小的特征图，通过 FPN 网络可以将 1/4 和 1/8 大小的特征图融合为 1/4 大小尺寸的新特征图。虚线框里详细介绍了特征图多尺度融合的过程，本文融合骨干网络的第一层和第二层特征，第二层作为深层特征首先经过 1×1 的卷积操作对其先降低维度，然后采用双线性插值的方法做上采样操作，使深层特征的尺寸扩张到第二层特征尺寸大小；第一层作为浅层特征做通道降维处理，然后将两层特征都经过 ECA 关

注重要特征信息，最后使用通道拼接的方法融合特征图。应用改进后的特征金字塔模块，在融合过程加入 ECA，使其侧重于相对重要特征，获取具有丰富语义信息和空间信息的特征，进行有效增强网络的预测精度。

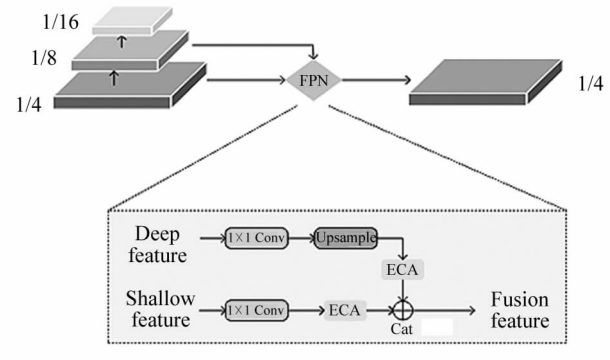

图 4　改进的 FPN 多尺度融合

Fig. 4　Improved FPN multiscale fusion

3　实验分析

3.1　数据集与训练策略

本文主要使用线上公开城市街景数据集 CamVid，CamVid 是剑桥驾驶标签视频数据集，数据图片包含训练集 367 张、验证集 101 张和测试集 233 张，分辨率大小为 480×360，目标分割类别为 11 类。

为了提高网络的收敛速度训练稳定性，首先将特征提取网络在 Imagenet 分类数据集上进行预训练，然后使用预训练模型对改进的网络模型进行迁移训练。网络模型超参数基本与原网络保持一致，batch size 设置为 4，使用随机梯度下降（stochastic gradientdescent，SGD）计算梯度，初始学习率 lr 为 1×10^{-2}，动量 momentum 为 0.9，权重衰减 weight decay 为 5×10^{-4}，学习率衰减采用 poly 策略。

3.2　实验环境

实验操作平台为 Ubuntu18.04 的 Linux 系统，CPU 为酷睿 i5-10400F，内存 16 G，GPU 为 Nvidia GeForce RTX 2060S，深度学习框架为 Pytorch1.8，Cuda11.1，Python3.6。

本文采用平均交并比（mean intersection over union，*MIoU*）作为语义分割任务的性能评价指标。交并比（intersection over union，*IoU*）计算的是预测值与真实值交集和并集的比值，*IoU* 越高，两者重合度越高，网络预测性能也越好。*MIoU* 为所有类别 *IoU* 的均值，计算如式（9）所示：

$$MIoU=\left(\sum_{p=1}^{M}\frac{X_{pp}}{T_p+\sum_{q=1}^{N}(X_{qp}-X_{pp})}\right)/M,\quad(9)$$

式中，M 表示像素的类别，T_p 表示第 p 类别的像素，X_{qp} 表示真实类别为 p、预测类别为 q 的像素。

3.3　实验结果对比

本文在 CamVid 城市街景数据集上对改进网络进行了训练、验证和测试，通过定量分析（包括消融实验）证明了改进网络的有效性和广泛适用性。实验部分对 HEAM 进行了可视化演示与分析，分割性能指标均采用 *MIoU* 指标。

3.3.1　不同骨干网络的比较

训练 CamVid 数据集时，输入尺寸 resize 为 720×720，batch size 设为 4，训练迭代次数为 14 000 次（300 epoch）。本文训练的网络模型是基于 DeepLabv3＋改进的，拥有的改进部分主要集中在 HEAM 和 MFFM 两个模块的嵌入。表 1 列出了基于 DeepLabv3＋语义分割网络在使用四种不同骨干网络模型时，基于 Baseline 和采用本文改进方法以及 *OS*（Output Stride）为 8 和 16 时的分割评价结果。*OS* 为图像的输入空间分辨率与输出的比值。表 1 实验数据表明，以 ResNet50 为骨干网络在 CamVid 验证集上分割结果效果最佳，*OS* 为 16 时，*MIoU* 为 75.4%，相较于 Baseline 分别提高了 2.3%；相比于其他骨干网络 ShuffleNetV2[11]、MobileNetV2[12]、ResNet18 的 *MIoU* 值分别提高了 7.0%、5.5%、3.7%；*OS* 为 8 时，以 ResNet50 为骨干网络能达到最佳效果，即 *MIoU* 分别能达到 76.9%，其分割效果表现最为优异，表明此网络相较于其他骨干网络具有更好的特征提取能力，因此本文采用 ResNet50 作为改进网络的骨干网络进行特征提取。

表 1　不同骨干网络在 CamVid 验证集上的比较

Tab. 1　Comparison of different backbone networks on CamVid verification set

Backbone	*OS*	Models	*MIoU*/%
ShuffleNetV2	8	Baseline	66.7
		＋MFFM＋HEAM	**68.5**
	16	Baseline	66.5
		＋MFFM＋HEAM	**68.4**
MobileNetV2	8	Baseline	70.1
		＋MFFM＋HEAM	**70.8**
	16	Baseline	69.1
		＋MFFM＋HEAM	**69.9**
ResNet18	8	Baseline	71.8
		＋MFFM＋HEAM	**72.9**
	16	Baseline	70.3
		＋MFFM＋HEAM	**71.7**
ResNet50	8	Baseline	74.9
		＋MFFM＋HEAM	**76.9**
	16	Baseline	73.1
		＋MFFM＋HEAM	**75.4**

3.3.2　消融实验分析

为验证 HEAM 和 MFFM 在改进网络结构中的实际效果,设计了逐次增加网络模块的消融实验,依次增加 ResNet50、MFFM、HEAM1、HEAM2 各个模块作对比。其中 HEAM1 和 HEAM2 分别是将 HEAM 嵌入到特征提取网络和 ASPP 之中。从消融实验结果表 2 可以看出,MFFM 对网络特征提取能力有很大提升,网络加入 MFFM 后,*MIoU* 相比于基线分别高出 0.9%;结果表明,HEAN 同时加在 1、2 两个位置时效果最佳,较之于单加 MFFM,*MIoU* 分别高出 1.4%。当 *OS* 设为 8 时,网络对城市街景的目标分割性能达到最佳,*MIoU* 可达到 76.9%。

表 2　改进网络在 CamVid 验证集上的消融实验

Tab. 2　Ablation experiment of improved network on CamVid verification set

ResNet50	MFFM	HEAM		*OS*	*MIoU*/%
		1	2		
√				16	73.1
√	√			16	74.0
√	√	√		16	74.8
√	√		√	16	75.0
√	√	√	√	16	75.4
√	√	√	√	8	**76.9**

针对提出的 MFFM,为验证效果最佳的特征融合方式,在保持 HEAM 不变的情况下,对骨干网络四个网络层的不同融合方式做消融实验。MFFM 主要分为两条特征融合支路,支路 A 对骨干网络的浅层特征融合,支路 B 对骨干网络的深层特征融合,然后跳跃连接到解码部分。从表 3 可以看出,单用骨干网络第一层作为支路 A,*MIoU* 为 74.2%;将第四层作为支路 B 加入后,*MIoU* 提升了 0.2%;而当将第一层和第二层作为支路 A,第三层和第四层作为支路 B 使四个网络层都加入到网络中,实验效果有明显提升,*MIoU* 提升了 0.9%;最后做特殊改变,将第二层特征同时加入支路 A 和 B 中,实现对骨干网络四个网络层的多阶段特征融合,*MIoU*达到了 75.4%。

针对提出的 HEAM,该注意力机制主要是在高度位置上的驱动,为验证其最佳的结构表现,在嵌入 MFFM 的基础上,主要对 HEAM 中 HANet 和 ECA 做并联和串联两种不同连接方式对比实验。如图 5 所示,图 5(a)在 HEAM 中只加入了 HANet;图 5(b)在 HANet 后串联 ECA;图 5(c)中的特征图在经过 HANet 处理的同时,经过 EACM 操作实现并联方式;图 5(d)为本文提出的方法,即将经过 HANet 处理的特征图与经过 ECA 处理的特征图实现并联后融合。实验结果如表 4 所示,HEAM 中仅有 HANet 时 *MIoU* 为 74.6%;HEAM 并联形式的分割性能要优于串联形式,最后本文提出的并联方式如图 5(d)显示的分割性能最佳,*MIoU* 能达到 75.4%。

表 3　MFFM 在 CamVid 验证集上的消融实验

Tab. 3　Ablation experiment of MFFM on CamVid verification set

Layers				*OS*	*MIoU*/%
1	2	3	4		
Branch1				16	74.2
Branch1			Branch2	16	74.4
Branch1	Branch1	Branch2	Branch2	16	75.1
Branch1	Branch1＋Branch2	Branch2	Branch2	16	**75.4**

表 4　HEAM 在 CamVid 验证集上的消融实验

Tab. 4　Ablation experiment of HEAM on CamVid verification set

HEAM	Picture	*OS*	*MIoU*/%
HANet	Fig. 5(a)	16	74.6
Series connection (HANet＋ECA)	Fig. 5(b)	16	74.9
Parallel connection (HANet＋ECA)	Fig. 5(c)	16	75.1
	Fig. 5(d)	16	**75.4**

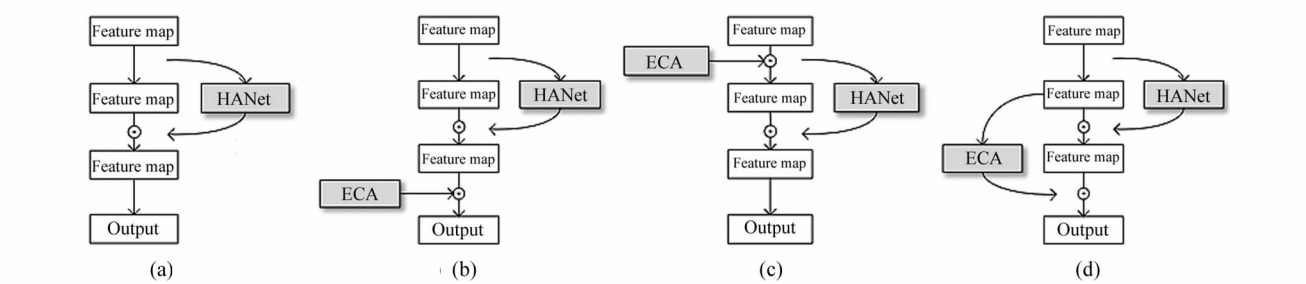

图 5　HEAM 串并联消融实验

Fig. 5　Ablation experiments of HEAM in series and parallel

为了更进一步地增强网络分割性能，本文还采用了常用的技巧如左右翻转、多尺度(scales＝{0.5，1.0，2.0})和滑动推理方式。通过此种方式，网络最佳模型在 CamVid 验证集上 *MIoU* 能达到 77.4%，如表 5 所示。

表 5　推理技巧在 CamVid 验证集上的对比

Tab. 5　Comparison of reasoning skills on CamVid verification set

Inference techniques	*OS*	Baseline	Ours
None	8	74.9	76.9
Multiscale，sliding，flipping	8	**75.6**	**77.4**

3.3.3　不同算法的比较

为了与目前其他先进网络进行比较，本文在 CamVid 数据集上对其他网络进行了训练并在测试集上做了结果测试。本文方法与其他先进网络模型在 CamVid 测试集上分割性能如表 6 所示。结果表明：本文提出的模型对城市街景的分割效果优于其他算法，体现了最佳的分割性能。

3.4　实验结果分析

3.4.1　HEAM 注意力可视化

城市街景数据集如图 6(a)所示，明显地可以分成上中下3部分，天空固定出现在上部，路面固定出现在下部，而中部多为复杂多变的小目标物体。图 6(b)是 HEAM2 的可视化注意力图，它计算了 ASPP 层的注意力权重，ASPP 结构通过不同的空洞卷积操作使注意力图的不同通道获得了多尺度特征。对比像素色彩，可以看出图 6(b)中虚线框内显示的是图 6(a)中部区域更低关注度的通道，而实线框内显示的通道拥有更多的关注度，对应于图 6(a)下部区域的大目标，如路面。

表 6　其他算法在 CamVid 测试集上的对比

Tab. 6　Comparison of other algorithms on CamVid test set

Model	*MIoU*	sky	build.	pole	road	swalk.	tree	sym.	fence	car	ped.	bicy.
ENet[13] ('16)	61.3	95.1	74.7	35.4	95.1	86.7	77.8	51.0	51.7	82.4	67.2	34.1
SegNet[2] ('17)	55.6	92.4	88.8	27.5	97.2	84.4	87.3	20.5	51.7	82.1	57.1	30.7
BiSeNet1[14] ('18)	65.6	91.9	82.2	25.4	93.3	77.3	74.4	42.8	49.7	80.8	53.8	50.0
DeepLabv3＋[5] ('18)	63.5	89.8	81.3	21.6	93.3	78.9	74.6	38.1	34.8	82.8	48.7	54.7
HRNet[15] ('19)	58.4	89.5	86.3	14.8	94.2	81.2	84.4	40.2	34.6	84.5	63.2	50.2
NDNet54-FNC8-LF[16] ('20)	57.5	88.8	85.5	17.6	90.4	83.8	80.6	39.2	37.3	82.6	60.1	53.7
LBN-AA[17] ('21)	68.0	92.5	83.2	36.3	93.0	82.1	70.5	51.6	53.2	81.7	55.6	47.9
Ours	**68.2**	91.1	84.0	30.1	94.7	80.6	75.7	49.5	40.5	86.9	57.8	58.9

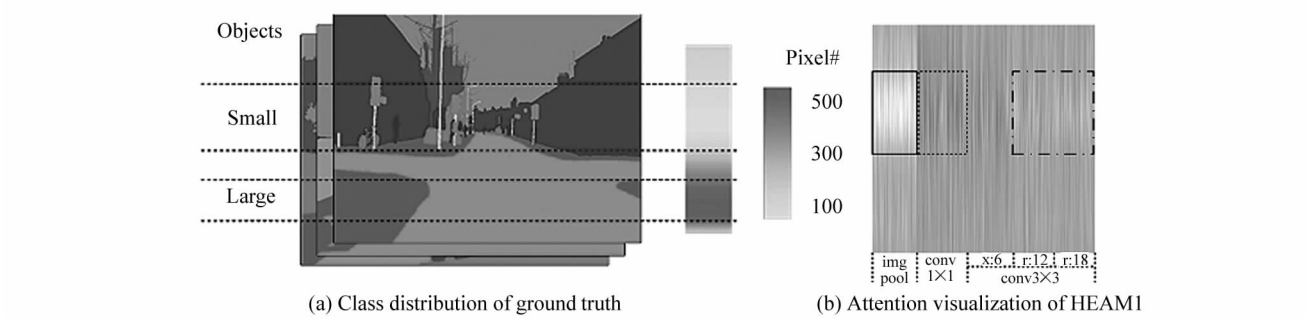

图 6　GT 图像高度分布和 HEAM 可视化：(a) GT 图像的类分布；(b) HEAM1 注意力的可视化

Fig. 6　Height distribution of ground truth images and heat visualization: (a) Class distribution of ground truth images; (b) Attention visualization of HEAM1

图 7(b)显示了 HEAM 的注意力分布可视化，它遵循了 CamVid 数据集中类别目标的实际高度分布，如图 7(a)所示。图中每个类别目标根据垂直位置给出相应权重大小，注意力图中类别 0(sky)和类别 1(building)的权重集中聚集在中上部，类别 3(road)和类别 8(car)的权重主要聚集在中下部，实验结果与真实数据类别分布基本一致，证明了 HEAM 确实关注了高度方向上的空间位置信息，并对个别依赖空间结构的大类别物体获取了更多的关注度，有效运用了数据的位置先验性。

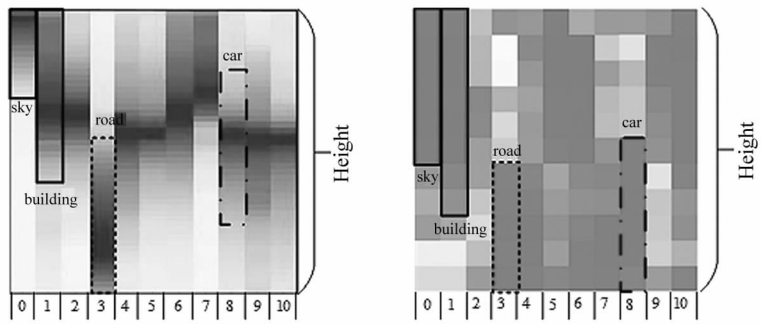

图 7　HEAM 注意力可视化:(a) 实际类分布; (b) 测试类分布

Fig. 7　Attention visualization of HEAM: (a) Real class distribution; (b) Test of class distribution

3.4.2　不同网络预测可视化

图 8 是 ENet、DeepLabv3+、BiSeNet1 和本文算法在 CamVid 测试集上的预测结果的可视化。结果表明,本文方法的预测结果更接近 GT 图像,预测的第一张图片中对树木(tree)和指示牌(symbol)的分割明显更清晰;第二张图片中对建筑物(building)和路灯(pole)的分割能力更强;第三张图片中对底部路面(road)分割更精准。显然,本文方法的 HEAM 和 MFFM 确实令底部大目标类别路面(road)和中部的小目标类别如指示牌(symbol)、路灯(pole)分割性能更好。

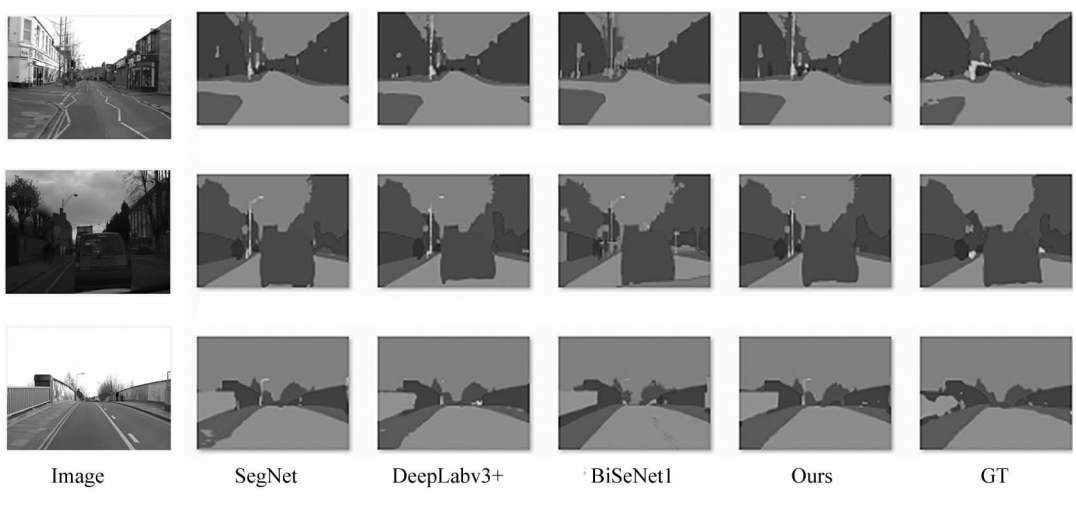

图 8　不同网络在 CamVid 数据集上的预测结果可视化

Fig. 8　Visualization of prediction results of different networks on CamVid dataset

4　结　论

研究表明,本文提出的基于 HEAM 与 MFFM 的城市街景分割网络,较好地运用了城市街景数据集固有的空间先验性,从而提高了垂直方向上个别大目标类别的分割性能;而且由于网络加强了对浅层特征的融合,使网络能更好地提取全局信息,基本消除了大目标物体分割有孔洞的缺陷,对大目标边缘分割也更精准。通过实验表明,本文提出的方法在 CamVid 数据集验证集上 $MIoU$ 能达到77.4%,在测试集上 $MIoU$ 能达到 68.2%。在后续的研究中,会考虑对损失函数进行改进,并引入上下文依赖来提高网络对图像的分割性能。

参考文献:

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 3431-3440.

[2] 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.

[3] 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.

[4] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [EB/OL]. (2017-06-17)[2022-01-05]. http://arxiv.org/abs/1706.05587.

[5] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin: Springer, 2018: 801-818.

[6] CHOI S, KIM J T, CHOO J. Cars can't fly up in the sky: improving urban-scene segmentation via height-driven attention networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 9373-9383.

[7] HE K, ZHANG X, REN S, 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.

[8] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2117-2125.

[9] WANG Q, WU B, ZHU P, et al. Eca-net: efficient channel attention for deep convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 11534-11542.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. Red Hook, NY: Curran Associates Inc., 2017: 5998-6008.

[11] MA N, ZHANG X, ZHENG H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]// European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin: Springer, 2018: 116-131.

[12] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2022-01-05]. http://arxiv.org/abs/1704.04861.

[13] PASZKE A, CHAURASIA A, KIM S, et al. Enet: a deep neural network architecture for real-time semantic segmentation[EB/OL]. (2016-06-07)[2022-01-05]. http://arxiv.org/abs/1606.02147.

[14] YU C, WANG J, PENG C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation [C]//15th European Conference on Computer Vision, ECCV, September 8-14, 2018, Munich, Germany. Berlin: Springer, 2018: 334-349.

[15] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 5693-5703.

[16] YANG Z, YU H, FU Q, et al. Ndnet: narrow while deep network for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(9): 5508-5519.

[17] DONG G, YAN Y, SHEN C, et al. Real-time high-performance semantic image segmentation of urban street scenes [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(6): 3258-3274.

作者简介：

熊　炜（1976—），男，博士，副教授，硕士生导师，主要从事数字图像处理和计算机视觉方面的研究.