DOI:10.19651/j.cnki.emt.2212385

基于像素注意力特征融合的城市街景语义分割算法研究*

李利荣[1,2] 丁 江[1] 梅 冰[1] 戴俊伟[1] 巩朋成[3]
(1.湖北工业大学电气与电子工程学院 武汉 430068; 2.湖北工业大学湖北省电网智能控制与装备工程技术研究中心 武汉 430068; 3.武汉工程大学计算机科学与工程学院 武汉 430205)

摘 要: 针对城市街景数据集中存在小目标和大量长条形状物体,分割难度大,虽然目前编码解码结构的网络能细化分割结果,但大多数都没有充分利用空间和上下文信息,因此本文提出一种基于像素注意力特征融合的语义分割算法。首先以ResNet50作为骨干网络,利用空洞空间卷积池化金字塔和条状池化进行初步特征融合,获得多尺度特征的同时规避无用信息;然后利用像素融合注意力模块,聚合上下文信息并恢复空间信息,最后利用注意力特征细化模块消除冗余信息。该算法在CamVid数据集上进行实验,结果表明该算法在验证集上能达到75.22%的mIoU,在测试集上也能达到67.21%。相比于DeepLabv3+网络分别提升了2.51%和2.86%。

关键词: 城市街景;像素融合;注意力机制;条状池化;语义分割
中图分类号: TP391.4 **文献标识码:** A **国家标准学科分类代码:** 520.99

Semantic segmentation method for urban street scenes based on pixel attention feature fusion

Li Lirong[1,2] Ding Jiang[1] Mei Bing[1] Dai Junwei[1] Gong Pengcheng[3]
(1. School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China; 2. Hubei Power Grid Intelligent Control and Equipment Engineering Technology Research Center, Hubei University of Technology, Wuhan 430068, China; 3. Shool of Computer Science and Engineering, Wuhan Engineering University, Wuhan 430205, China)

Abstract: For the presence of small targets and a large number of long bar-shaped objects in urban streetscape datasets, segmentation is difficult, and although current networks with coding and decoding structures can refine segmentation results, most of them do not make full use of spatial and contextual information, so this paper proposes a semantic segmentation algorithm based on pixel attention feature fusion. Firstly, using ResNet50 as the backbone network, the initial feature fusion is carried out using the null space convolutional pooling pyramid and strip pooling to obtain multi-scale features while circumventing useless information; then the pixel fusion attention module is used to aggregate contextual information and recover spatial information, and finally the attention feature refinement module is used to eliminate redundant information. The algorithm was experimented on the CamVid dataset and the results showed that the algorithm was able to achieve 75.22% mIoU on the validation set and 67.21% on the test set. This is an improvement of 2.51% and 2.86% respectively compared to the DeepLabv3+ network.

Keywords: urban streetscape;pixel fusion;attention mechanism;strip ponding;semantic segmentation

0 引 言

随着自动化水平的不断提高,车辆驾驶技术逐渐向无人驾驶领域发展,如何利用现有技术对城市街景进行有效且精准的分割成为一项具有挑战性的任务。

近年来,随着深度学习的快速发展,各种新颖有效的图像分割模型和算法被提出,并且应用到计算机视觉的无人驾驶领域,开始大放异彩。作为分割任务的开山鼻祖,全卷积神经网络(fully convolutional networks, FCNs[17])对图像分割的意义和影响尤为深远,全卷积层的结构使网络能

收稿日期:2022-12-13
* 基金项目:国家自然科学基金(62071172,62202148)项目资助

够接受任意大小的输入,但 FCNs 仅使用深层特征对像素进行分类,而被认为具有丰富空间信息的浅层特征却没有被合理的利用起来,导致最终的分割效果较为粗糙,分割边界模糊。为了解决这一问题,并对重要的空间信息加以利用,采用编码解码结构的 U-Net[2] 和 SegNet[3] 随后被提出,该类结构利用解码器逐渐恢复空间信息来捕捉清晰的目标边界,同时 U-Net 也使用拼接方式进行特征融合,通过通道数的拼接,可以形成更深的特征。如果将差异较大的浅层特征和深层特征直接拼接必然会加大网络的学习难度,因此为了减少语义差别,Zhou 等[4] 在 U-Net 直接拼接的基础之上增加了类似于密集连接结构的卷积层,并融合了下一阶段卷积的特征。为了更好的对上下文信息进行利用,聚合不同区域的语义信息,使网络更具表达能力,Zhao 等[5] 提出了一种金字塔场景分析网络(pyramid scene parsing network,PSPNet),将像素级特征利用全局金字塔池化模块进行多次下采样,让全局信息和局部信息进一步融合[6]。DeepLab 系列分割网络均采用空洞卷积结构,该结构能在不改变图像输出特征图尺寸的情况下增大感受野。Chen 等[7] 改进的 DeepLabv3+网络结构利用空洞卷积的同时应用了编码解码结构,恢复了空间信息,使分割能力进一步提升,但由于空洞卷积会不可避免的造成局部细节信息损失,远距离获取的信息也没有相关性,影响分类结果,因此 DeepLabv3+网络依然存在个别目标分割不够精细、目标边界分割精度低等问题。白艳琼等[8] 以 DeepLabv3+为基本网络模型,为了能捕获更多的空间上下文信息和高级语义信息,在编解码中引入 PAM 和 CAM,两个注意力模块采用平行结构,在空间和通道维度上捕捉更多的语义信息,而杜梓维[9] 为了改善小尺寸目标的识别率和分割精度,采用模型迁移和加权损失函数的方法并结合当前主流神经网络进行特征提取,但这两种方法对浅层特征都缺乏利用。赵迪等[10] 提出在 DeepLabv3+网络的基础上引入高度有效驱动注意力机制,并将其嵌入到特征提取网络与空洞空间金字塔池化中,使其能关注到更多垂直方向上的空间位置提高分割效果,该方法有效的利用了浅层特征,但同时也带来了一定的信息冗余,缺乏细化能力。

为了更好的利用空间信息和上下文信息,提升语义分割的准确率,本文提出了一种基于 DeepLabv3+网络的像素注意力特征融合的算法。该算法引入了一个像素融合注意力模块(pixel fusion attention module,PFAM),并融合改进之后的注意力特征细化模块[11](attention feature refinement module,AFRM)。同时针对城市街景数据集中存在大量的长条形带状结构物体,常用的方形池化核会引入与目标不相关的污染信息,影响目标预测,在此本文采用了一种条状池化[12](strip pooling,SP),聚合全局和局部上下文信息的同时,去除多余信息的干扰,提高有效特征的提取率,最后所有实验都在 CamVid 数据集上进行验证与测试。

1 基于像素注意力特征融合的城市街景语义分割总体框架

本文提出的城市街景语义分割总体网络框架如图 1 所示,其主要由骨干网络 ResNet50、空洞空间卷积池化金字塔(atrous spatial pyramid pooling,ASPP)[13]、SP、PFAM 以及 ARFM 5 部分构成。

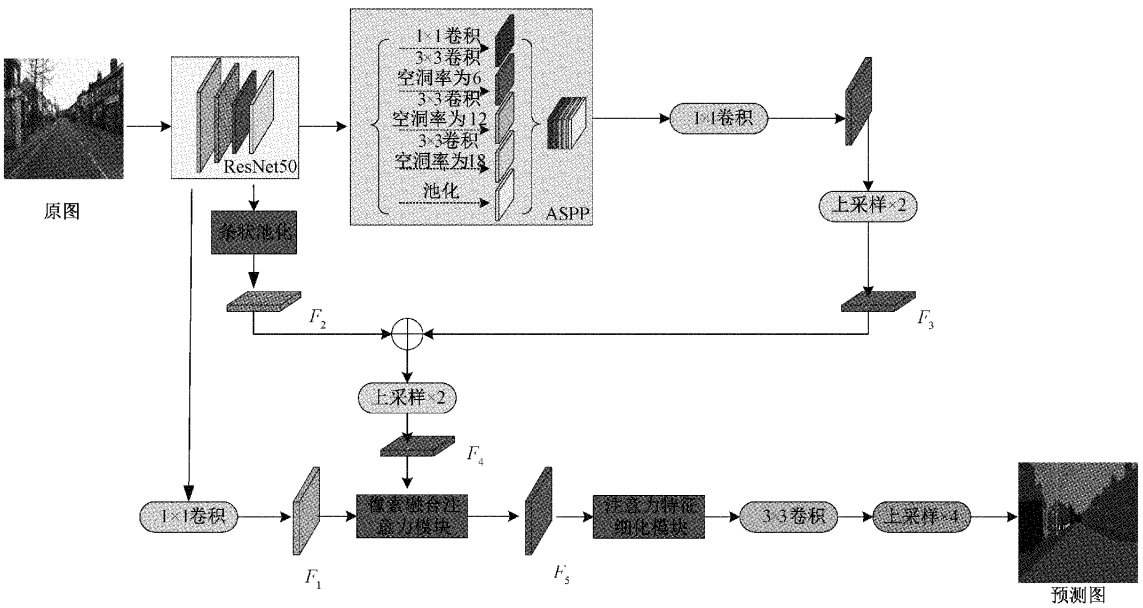


图 1 城市街景语义分割总体框架

骨干网络采用 ResNet50,以 ResNet50 的 stage4 特征输出作为 ASPP 的输入,由于 stage4 经过了多次的卷积,网络层数较深,具有较为丰富的语义信息,让其经过 ASPP 之后得到多尺度特征 $\boldsymbol{F}_3$,$\boldsymbol{F}_3$ 融合了浅层特征的空间信息

和深层特征的语义信息，对网络的表征能力更强。其次将 stage2 输出进行条状池化得到特征 $\boldsymbol{F}_2$，特征 $\boldsymbol{F}_2$ 聚合全局和局部上下文信息，避免无用信息的干扰，提高了有效特征的利用率。接着将特征 $\boldsymbol{F}_2$ 和 $\boldsymbol{F}_3$ 进行逐元素相加，得到融合之后的特征 $\boldsymbol{F}_4$，此时的特征 $\boldsymbol{F}_4$ 具备较为丰富的上下文信息，能更好的进行分割定位。然后将具有丰富空间信息的 stage1 的输出特征 $\boldsymbol{F}_1$ 作为浅层特征，和深层特征 $\boldsymbol{F}_4$ 同时输入到 PFAM 中进行像素注意力融合，得到具有丰富局部信息以及上下文信息的特征 $\boldsymbol{F}_5$。最后将特征 $\boldsymbol{F}_5$ 通过 AFRM 进行特征细化，过滤掉其中因为直接融合而产生的较多冲突信息[14]，保留细节信息，从而提高分割预测能力。本文的整体网络架构依旧采用编码解码结构，通过逐步恢复空间信息来补全目标轮廓信息，完善了分割效果，同时本文网络架构的输入和输出保持同尺寸大小。

1.1　像素融合注意力模块

DeepLabv3＋中的 ASPP 模块能够通过不同的空洞率获取多尺度卷积特征[15]，虽然具有较为丰富的语义信息也扩大了感受野，但由于空洞卷积的性质，每一层得到的卷积结果是相对独立的，没有相互依赖关系，如此各层的卷积结果之间就没有相关性，导致局部信息丢失；此外，空洞卷积稀疏的采样输入信号，使得远距离卷积得到的信息之间也没有关联性，最终会影响分类效果，因此采用 ASPP 不足以完成像素密集预测任务。所以本文提出了一个有效的像素融合注意力模块(pixel fusion attention module，PFAM)，将其用作解码器模块的一部分来增强局部细节信息和空间信息。如图 2 所示，特征 $\boldsymbol{F}_1$ 包含丰富的全局细节信息，而特征 $\boldsymbol{F}_4$ 包含较为丰富的上下文信息。PFAM 由两个分支组成：长跳跃连接分支和逐像素注意分支。长跳跃连接是通过两个输入特征 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_4$ 进行逐元素相加来实现的，能够进行像素之间的融合，恢复了部分经过不断卷积而损失的边界轮廓信息，同时保证了特征信息的连续性和远距离获取信息的相关性。分支二是计算像素级空间注意力图($\boldsymbol{M}_{att}$)，它能精确感知编码位置信息[16]，而空间位置信息在语义分割中尤为重要，是目前语义分割算法取得有效进展的关键技术之一。$\boldsymbol{M}_{att}$ 实现的方法是对特征 $\boldsymbol{F}_4$ 应用一个残差块和线性嵌入一个 1×1 卷积，其中残差块(residual block，RB)是由 2 个连续 3×3 卷积之后加一个 $softmax$ 激活函数构成的，然后通过 $softmax$ 函数来进行目标分类的概率预测。这个注意力图 $\boldsymbol{M}_{att}$ 帮助经过逐元素相加之后的特征 $\boldsymbol{M}_{add}$ 转换为空间细节增强的特征 $\boldsymbol{M}_1$。其公式如下：

$$M_{add}=F_1+F_4 \tag{1}$$

$$M_{att}=Softmax(\mathrm{Conv1}\times1(\mathrm{RB}(F_4))) \tag{2}$$

$$M_1=M_{add}\otimes M_{att} \tag{3}$$

为了避免在文献[17]中报告的训练初期对特征 $\boldsymbol{M}_1$ 的过度强调或忽略其的重新缩放问题，并进一步加强像素融合的局部细节信息，本文将另一个注意力块应用于 $\boldsymbol{M}_1$，其中注意力块是在残差块原始结构中增加一个 CoT[18]注意力模块，同时将 $\boldsymbol{M}_1$ 与 $\boldsymbol{M}_{add}$ 进行特征融合，最后在其后应用一个双线性上采样和 1×1 卷积输出解码器特征 $\boldsymbol{F}_5$。

通过 PFAM 可以增强特征图的精细细节，鼓励捕获局部细节，并使解码器特征图包含丰富的全局和局部上下文信息，生成精细且连贯的空间预测。

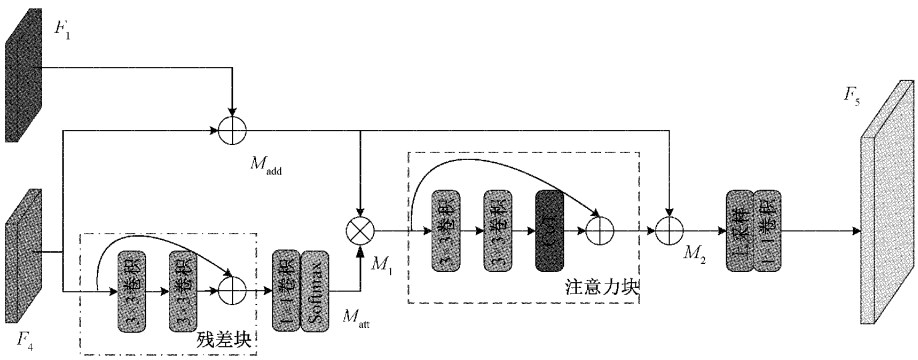


图 2　像素融合注意力模块(PFAM)

在传统的自注意力机制[19]中，所有成对的查询键关系都是通过孤立的查询键独立学习的，没有考虑丰富的上下文信息，这严重限制了基于 2D 特征图的自我注意学习能力。而 CoT 模块可以在单一的架构中充分利用相邻键之间的上下文信息，将键之间的上下文挖掘和二维特征图上的自注意力学习统一起来，增强输出聚合特征图的表达能力。该模块同时捕获输入键的静态上下文和基于上下文自注意力的动态上下文信息，以促进视觉表示学习。

图 3 是 COT 模块的结构图，首先对输入特征 $\boldsymbol{X}$ 分成三条支路输出，其中一条支路进行 $k\times k$ 的分组卷积，来获得具备局部上下文信息的表示，记作 $\boldsymbol{K}^1$，$\boldsymbol{K}^1$ 可以看作是在局部信息上进行了静态的建模。其次将 $\boldsymbol{K}^1$ 和 $\boldsymbol{X}$ 进行通道维度的拼接，增加特征数，然后对拼接的结果进行两次连续的 1×1 卷积操作得到 $\boldsymbol{A}$：

$$A=[K^1,Q]W_\theta W_\delta \tag{4}$$

其中，W_θ 中带有 Relu 激活函数，W_δ 中不带激活函数。

不同于传统的自注意力机制，此模块中的 $\boldsymbol{A}$ 是由输入

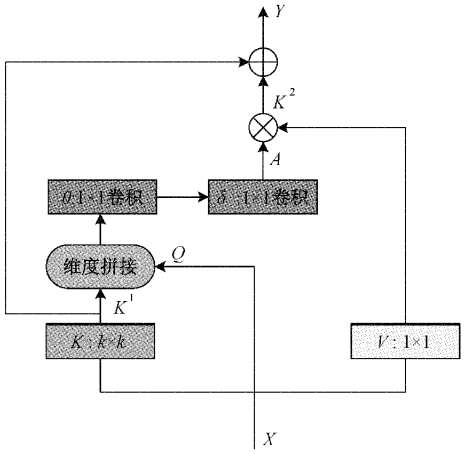


图 3 CoT 模块

信息 $\boldsymbol{X}$ 和局部上下文信息 $\boldsymbol{K}^1$ 拼接得到的,而不只是建模了传统自注意力机制中的 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的关系,从而通过局部上下文建模的引导,增强了自注意力机制。最后将 $\boldsymbol{A}$ 和 $\boldsymbol{V}$ 逐元素相乘,得到动态上下文建模的 $\boldsymbol{K}^2$:

$$K^2 = V \otimes A \tag{5}$$

CoT 的输出 $\boldsymbol{Y}$ 为局部静态上下文建模的 $\boldsymbol{K}^1$ 和全局动态上下文建模的 $\boldsymbol{K}^2$ 逐元素相加之后的结果,即

$$Y = K^1 + K^2 \tag{6}$$

1.2 条状池化

在某些情况下,目标对象中可能具有长条形带状结构或离散分布,因此方形核的池化操作不可避免的会包含来自与目标不相关区域的污染信息,所以使用大的方形池化窗口无法在离散分布区域之间建立远程的依赖关系,而条状池化的长方形池化核在一个维度上能够捕获孤立区域的远程关系,在另一个维度上又有助于捕获局部上下文信息并防止不相关区域干扰标签预测,集成这种长且窄的池化核可以帮助网络同时聚合全局和局部上下文信息,合理利用有效信息,有利于大尺度的目标任务。

如图 4 所示,令 $x \in \mathbb{R}^{C\times H\times W}$ 为输入张量,$y \in \mathbb{R}^{C\times H\times W}$ 为过渡张量,$z \in \mathbb{R}^{C\times H\times W}$ 为输出张量,因此条状池化的计算公式可为:

$$z = x \otimes \sigma(Conv(y)) \tag{7}$$

其中,C 表示通道数,$H\times W$ 为特征图尺寸大小,$\otimes$ 为逐像素相乘,σ 为 $sigmoid$ 函数,$Conv$ 为 1×1 卷积。$y \in \mathbb{R}^{C\times H\times W}$ 则可由下式产生:

$$y_{c,i,j} = y^h_{c,i} + y^v_{c,j} = \frac{1}{W}\sum_{0\leqslant j<W} x_{i,j} + \frac{1}{H}\sum_{0\leqslant i<H} x_{i,j} \tag{8}$$

其中,$y^h_{c,i}$ 表示经过 $H\times 1$ 池化核之后的输出,$y^v_{c,j}$ 表示经过 $1\times W$ 池化核之后的输出。

本文从骨干网络 ResNet50 的 stage2 引出一条支路,因为该支路在空间信息和细节信息上有一个较好的平衡,在该支路上应用条状池化能更好的捕获孤立的远程关系,聚合上下文信息。

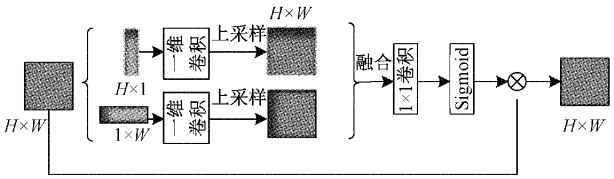


图 4 条状池化结构图

1.3 注意力特征细化模块

AFRM 分为两个分支,如图 5 所示,第 1 个分支使用全局平均池化来捕获全局上下文信息,而第 2 个分支保持原始的特征输入。该设计可以细化每个阶段的输出功能,可以轻松联系全局上下文信息,而无需进行额外的上采样操作。将经过全局平均池化和 1×1 卷积等一系列操作之后的细化特征和原始输入特征进行哈达玛积计算,能更加有效的恢复损失的像素边界信息,对目标边界信息的利用更加充分。

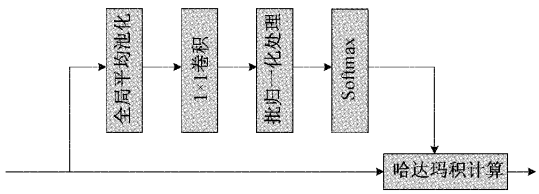


图 5 注意力特征细化模块

2 实验过程与分析

2.1 数据集与训练策略

本文使用的是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,其中划分训练集、验证集和测试集分别为 367 张、101 张以及 233 张,分辨率大小为 480×360,目标分割类别为 11 类。同时为了加快模型训练收敛速度,此次实验中运用了 Imagenet 图像分类数据集的预训练模型。为了合理利用显存资源和训练时间,本文采用冻结与解冻方式进行训练[20]。一共训练 200 个周期得到检测模型,验证和训练的损失曲线如图 6 所示。

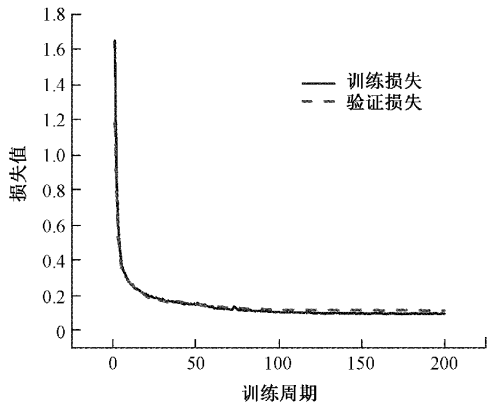


图 6 损失函数

2.2　实验环境

本文的实验环境如下：深度学习框架为 Pytorch1.2，编程语言采用 Python3.6，实验平台采用 Ubuntu20.10 操作系统，并配置有 GeForce RTX 2080 ti 显卡(11 GB 内存)。此外网络模型训练相关参数如表 1 所示。

表 1　训练参数设置

参数	数值/名称
优化器	SGD
冻结批次大小	8
解冻批次大小	4
训练周期	200
初始学习率	7×10^{-3}
动量值	0.9
权重衰减	1×10^{-4}
下采样因子	16

学习率衰减采用 CosineAnnealingLR 策略，以 cos 函数为周期，以初始学习率为最大学习率，$2\times T_{\max}$ 为周期，在一个周期内先下降后上升。公式如下：

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1+\cos\left(\frac{T_{cur}}{T_{\max}}\pi\right)\right) \tag{9}$$

其中，η_t 为当前学习率，$\eta_{\min}$ 为学习率最小值，$\eta_{\max}$ 为学习率最大值，T_{cur} 为当前训练周期，$T_{\max}$ 为最大训练周期。

3　相关实验

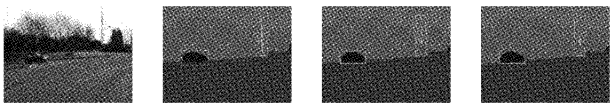
3.1　消融实验

DeepLabv3＋网络以 DeepLabv3 网络为编码器，以骨干网络的部分输出作为解码器的一部分输入，用来实现更好的融合，将具有分辨率更高，包含更多位置、细节信息的浅层特征和经过 ASPP 之后的具有更强语义信息的深层特征进行融合，提高分割能力，因此本文以 ResNet50 为骨干网络，探讨了将其不同特征层作为浅层特征送入解码器中的不同效果。结果如表 2 所示，ResNet50 的 stage1 作为浅层特征进行编码效果最好，相比于 stage2 和 stage3 分别提升了 0.69%和 3.63%，归其原因，stage1 的特征距离输入特征比较近，包含更多的像素点信息以及一些细粒度信息，例如图像的一些颜色、纹理、边缘、棱角信息，而 stage3 经过了更多的卷积操作，空间信息和像素信息损失很多，所以和深层特征融合之后也无法恢复损失的部分信息，导致没有充分利用空间信息，分割效果不佳。

表 2　ResNet50 不同层之间的对比　%

所用模块	mIoU(测试)	mIoU(验证)
stage1	**64.35**	**72.71**
stage2	63.66	71.05
stage3	60.72	65.83

可视化结果如图 7 所示，可以明显发现，以 stage1 为浅层特征输入，对目标边界分割更加地清晰，对小目标的特征提取也更加有效。



(a) 原图　(b) stage1　(c) stage2　(d) stage3

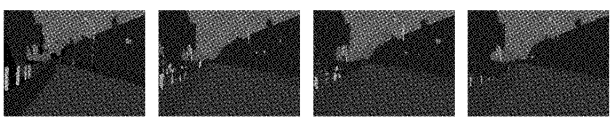
图 7　ResNet50 不同层的可视化

为了验证在 PFAM 中 CoT 模块对像素融合效果的影响，本文做了相关的消融实验，结果如表 3 所示，采用 CoT2(把 CoT 模块加在第 2 个残差块中，即图 2 中的注意力块)的效果最好，在 CamVid 测试集上可以达到 66.42%的 mIoU 和 74.91%的 mPA，加在 CoT1(把 CoT 模块加在第 1 个残差块)处 mIoU 也有 0.64%的提升，但同时加入两个 CoT 模块，则在 mIoU 上下降了 0.29%。通过进一步分析可以推测，由于 CoT 模块是一个改进的自注意力机制，这种设计超越了传统的自注意力机制，通过额外利用输入键之间的上下文信息来促进自我注意学习，最终提高了深层网络的表征特性。但自注意力机制的计算量相对较大，在 PFAM 中单独使用一个 CoT 模块，可以更好的联系上下文信息，提高分割效果，但同时引入两个 CoT 模块，虽然加强了对上下文信息的利用，捕获图像特征中的远程依赖关系，不同尺度的特征进行整合后也有效的缩小了语义差异，但过多利用自注意力机制，整合后的特征中会不可避免的带来部分信息的冲突，造成信息的冗余，反而加大了计算量，带来了部分信息的混乱，对分割预测造成了干扰。

表 3　不同位置的 CoT 模块在 CamVid 测试集上对比　%

所用模块	mIoU	mPA
CoT1	65.99	74.21
CoT2	**66.42**	**74.91**
CoT1＋CoT2	65.06	73.89

该消融实验的可视化结果如图 8 所示。



(a) 标签　(b) CoT1　(c) CoT2　(d) CoT1+2

图 8　不同位置 CoT 模块的可视化

为了检验本文提出的模块对整体网络性能的影响，进行了不同模块之间的消融实验，实验结果如表 4 所示。

在骨干网络中引出一条支路进行条状池化，聚合全局和局部上下文信息，过滤多余信息，提高有效特征的提取率，在验证集上相比于基准网络，mIoU 提升了 1.3%；增加

表 4 不同模块在 CamVid 数据集上的比较 %

所用模块	mIoU(测试/验证)	mPA(测试/验证)	Acc(测试/验证)
R50	64.35/72.71	74.55/81.2	90.64/94.26
R50+SP	65.35/74.01	75.58/81.23	90.82/94.69
R50+SP+PFAM	66.42/74.56	74.91/81.78	91.17/94.83
R50+SP+PFAM+AFRM	**67.21/75.22**	**76.18/83.37**	**91.20/94.94**

PFAM 能够增强像素融合,充分利用特征的空间位置信息,自注意力的结构也能捕获远程依赖关系,探索更一般的上下文信息,因此在验证集上的 mIoU 相比基准网络提高了 1.85%;增加 AFRM 实现特征细化后,能消除部分冗余信息,细化分割效果,对比基准网络在验证集上 mIoU 提升了 2.51%,测试集上也有 2.86%的提升,同时 mPA 也提升了 1.63%。

3.2 对比实验

为了验证本文算法的优势,将本文算法同当前主流语义分割算法做对比试验。表 5 为不同算法的分割效果对比。从实验结果可以发现,本文算法相比于多尺度特征语义分割算法 PSPNet 和 DeepLabv3+在测试集上的 mIoU 分别提升了 4.85%和 2.86%,mPA 分别提升了 5.03%和 1.63%,与 BiSeNet1 相比在测试集上 mIoU 也提升了 1.59%,与其他作者基于 DeepLabv3+改进后的算法相比也有 0.61%的提升,说明本文算法对空间和上下文信息的利用有效,特征细化对算法精度的提升也起到了一定的作用。

表 5 不同算法在测试集上的对比 %

算法模型	mIoU	mPA
PAM+CAM[8]	58.79	72.29
ENet[21]	61.32	70.06
PSPNet[5]	62.36	71.15
HRNet[22](W32)	63.85	73.59
HRNet[22](W48)	64.97	75.35
DeepLabv3+[7]	64.35	74.55
BiSeNet1(resnet18)[11]	65.62	75.43
DeepLabv3+'[9]	66.60	**79.25**
Ours	**67.21**	76.18

为了从视觉角度验证本文算法对城市街景语义分割的有效性,分别对不同算法的分割结果在原图上进行可视化,图 9 所示为部分数据集的可视化结果图展示,本文算法对目标边界的分割效果更加细化,边界轮廓对比于其他算法更加的清晰。同时针对细小目标,比如路标和红绿灯,本文算法虽然也有待完善,但整体上的效果却优于其他算法。

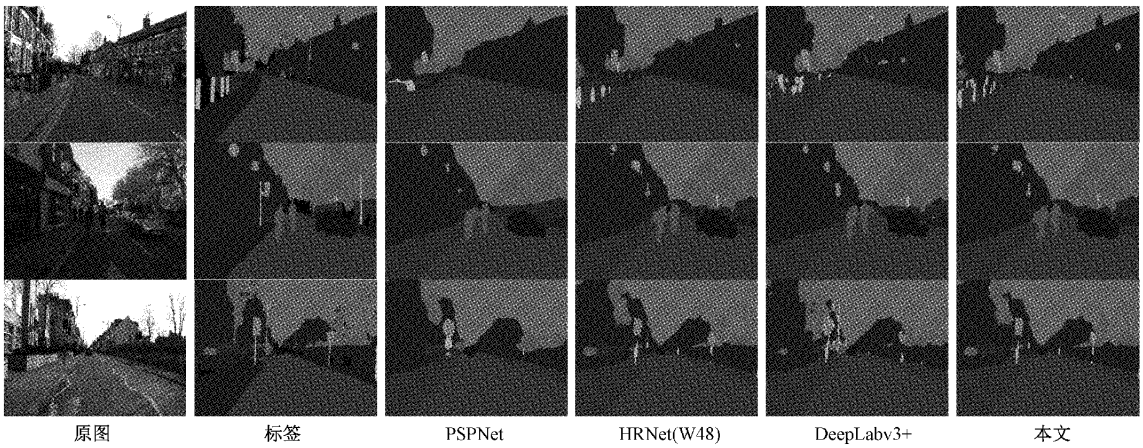


图 9 不同算法分割结果可视化

4 结 论

通过一系列实验表明,本文提出的基于像素注意力特征融合的城市街景语义分割算法能够很好的利用自注意力思想,聚合上下文信息和全局信息,又针对数据集存在大量长条形状的特点应用了条状池化,规避了无用的干扰信息,加强了有效特征的提取效率,最后利用注意力细化模块对融合之后产生的冗余信息进行了过滤,对分割效果至关重要的边界细节信息进行了细化。从最终实验可以清楚得到,本文提出的这种算法可以在 CamVid 数据集的验证集上达到 75.22% 的 mIoU,在其测试集上也能够达到 67.21%。

在未来的工作中可以考虑对数据集中小目标的有效特征提取进行研究,提高网络的特征提取能力和泛化能力,同时也可以参考 transformer 结构进行网络性能的改进。

参考文献

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

[2] 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. Springer, Cham, 2015: 234-241.

[3] BADRINARAYANAN V, HANDA A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling [J]. ArXiv Preprint,2015, ArXiv:1505.07293.

[4] ZHOU Z, RAHMAN S M M, TAJBAKHSH N, et al. Unet++: A nested u-net architecture for medical image segmentation [M]. Springer: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018.

[5] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2881-2890.

[6] 李芬,李志鹏,冯祥胜,等.基于语义分割的城市内涝检测算法[J].国外电子测量技术,2022,41(7):45-49.

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

[8] 白艳琼,郑玉甫,田宏.基于 Deeplabv3+和注意力机制的道路场景语义分割方法(英文)[J]. Journal of Measurement Science and Instrumentation, 2021, 12(4):412-422.

[9] 杜梓维. 基于 DeepLabV3+模型的街景影像语义分割方法研究[D].上海:东华理工大学,2022.

[10] 赵迪,孙鹏,陈奕博,等.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[J].光电子·激光,2022,33(10):1038-1046.

[11] YU C, WANG J, PENG C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 325-341.

[12] HOU Q, ZHANG L, CHENG M M, et al. Strip pooling: Rethinking spatial pooling for scene parsing[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 4003-4012.

[13] 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, 2017, 40(4): 834-848.

[14] 李利荣,张云良,陈鹏,等.基于上下文信息增强与特征细化的绝缘子破损检测方法[J].高电压技术:2022:1-10,DOI:10.13336/j.1003-6520.hve.20220547.

[15] 黄聪,杨珺,刘毅,等.基于改进 DeeplabV3+的遥感图像分割算法[J].电子测量技术,2022,45(21):148-155.

[16] HAN D, SHIN J, KIM N, et al. TransDSSL: Transformer based depth estimation via self-supervised learning[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 10969-10976.

[17] HUANG S, LU Z, CHENG R, et al. FaPN: Feature-aligned pyramid network for dense image prediction [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 864-873.

[18] LI Y, YAO T, PAN Y, et al. Contextual transformer networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1489-1500.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[20] 李利荣,陈鹏,张云良,等.基于多尺度特征编码和双重注意力融合的绝缘子缺陷检测[J].激光与光电子学进展,2022,59(24):81-90.

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

[22] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349-3364.

作者简介

李利荣,博士,讲师,主要研究方向为图像分析、计算机视觉与模式识别。

E-mail:Rongli@hubt.edu.cn