激光与光电子学进展 先进成像

多尺度特征对齐聚合的语义分割方法

徐兆忠[1]，彭力[1,2*]，戴菲菲[3]

[1]江南大学物联网工程学院物联网技术应用教育部工程研究中心，江苏 无锡 214122；

[2]无锡太湖学院江苏省物联网应用技术重点建设实验室，江苏 无锡 214122；

[3]台州市产品质量安全监测研究院，浙江 台州 318000

摘要 卷积神经网络在对图像进行语义分割时，高层特征经过降采样和padding操作和低层特征之间容易产生错位。为了解决高低层特征之间的错位问题，更好地聚合多尺度特征信息，提出了一种带有多尺度特征对齐聚合(MFAA)模块的语义分割方法。MFAA模块采用一种可学习插值策略来学习像素的变换偏移，可以有效缓解不同尺度特征聚合的特征不对齐问题，同时模块内的注意力机制提高了解码器恢复重要细节特征的能力。该方法利用高层特征的语义信息和低层特征的空间信息，通过多个MFAA模块将高低层特征对齐之后聚合到一起，从而实现图像更加精细的语义分割效果。将所提网络结构在语义分割数据集PASCAL VOC 2012上进行了验证，使用ResNet-50作为骨干网络时在验证集上的平均交并比值达到了78.4%。实验结果表明，该方法与几种主流分割方法相比在评价指标方面存在优越性，可以有效提高图像分割的效果。

关键词 机器视觉；图像语义分割；特征对齐；多尺度特征；注意力机制

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP212814

Semantic Segmentation Method Based on Multiscale Feature Alignment and Aggregation

Xu Zhaozhong[1], Peng Li[1,2*], Dai Feifei[3]

[1]*Engineering Research Center of Internet of Things Technology Applications, School of IoT Engineering, Jiangnan University, Wuxi* 214122, *Jiangsu, China*;

[2]*Jiangsu Province Internet of Things Application Technology Key Construction Laboratory, Wuxi Taihu College, Wuxi* 214122, *Jiangsu, China*;

[3]*Taizhou Product Quality and Safety Monitoring Institute, Taizhou* 318000, *Zhejiang, China*

Abstract During semantic segmentation of images, a convolutional neural network easily misplaces the high-level features with low-level features after down-sampling and padding operations. To solve the mismatch problem between high- and low-level features and better aggregate the multiscale feature information, this paper proposes a semantic segmentation method with a multiscale feature alignment aggregation (MFAA) module. The MFAA module adopts a learnable interpolation strategy to learn pixel transform migration, thereby alleviating the feature-misalignment problem of feature aggregation at different scales. The module includes an attention mechanism that improves the decoder's ability to recover the important details. Using multiple MFAA modules, the semantic information of high-level features, and the spatial information of low-level features, this method aligns and aggregates the high- and low-level features to refine the semantic segmentation effect. The proposed network structure was validated on PASCAL VOC 2012. Using a ResNet-50 backbone network, the mean intersection-over-union reached 78.4% on the validation set. Experimentally, the proposed method achieved better evaluation indices than several mainstream segmentation methods and effectively improved the image segmentation effect.

Key words machine vision; image semantic segmentation; feature alignment; multiscale feature; attention mechanism

收稿日期：2021-10-26；**修回日期**：2021-11-15；**录用日期**：2021-11-29；**网络首发日期**：2021-12-10

基金项目：国家自然科学基金(61873112)、国家重点研发计划(2018YFD0400902)

通信作者：*penglimail2002@163.com

1　引　　言

作为计算机视觉的基础任务之一，图像语义分割是目前计算机视觉的热点研究方向[1]。语义分割是一个像素级的分类任务，可以对图像中每一个像素点按设定的语义标签进行分类[2]。图像语义分割有许多应用场景：在医疗图像[3]领域，通过语义分割可以精准找出医疗图像中的肿瘤等病变部位，减少医生的负担；在自动驾驶[4]领域，可以帮助掌握驾驶时汽车周围的环境信息，识别道路与障碍物；在地理信息系统中，可以识别出卫星遥感影像中的道路、建筑、河流等信息，并对其分别标注。在深度学习尚未应用到计算机视觉领域时，对图像的分割主要分为基于阈值、边缘和区域的方法。随着计算机性能的提高及 GPU 加速技术的出现，以卷积神经网络（CNN）为代表的深度学习方法取得了较大的进展。Long 等[5]在 2015 年提出了一种将全卷积神经网络（FCN）用于图像语义分割的方法，该方法被认为是卷积神经网络用于语义分割的基石之作。FCN 将 VGG16[6]中的全连接层换成了卷积层，通过上采样得到高分辨率的深层特征后与浅层特征直接相加得到密集的预测结果，实现了端到端的图像分割。随后出现了一大批以 FCN 为基础架构的图像语义分割方法。

特征融合是语义分割最近研究进展主要遵循的策略之一，然而特征融合的方法融合了不同尺度卷积块的特征，这可能引起特征错位的问题。为了解决这个问题，Lu 等[7]提出 IndexNet 来学习池化和上采样操作的索引。Jaderberg 等[8]提出了一个新的可学习模块来提高卷积神经网络的空间不变性。Mazzini 等[9]提出了一个可以被引导的上采样模块来学习每个像素位置的二维变换偏移量。SFNet[10]和 AlignSeg[11]用光流的方式进行配准，来计算每个像素的运动偏移并进行校正。

受以上方法的启发，本文提出了一种多尺度特征对齐聚合的语义分割方法，旨在更好地利用各层级特征信息对齐和融合多尺度特征，从而实现更精细的分割效果。利用骨干网络不同层级的特征有助于恢复图像边缘信息和纹理信息，提高网络的细节表征能力。通过多个特征对齐聚合模块逐步将低分辨率的深层特征与高分辨的浅层特征相融合，逐步挖掘不同分辨率的特征信息。在高级特征对齐融合前加入空间注意力模块，增大重要的空间细节的权重，减少噪声干扰的同时强化网络的学习能力。使用空洞空间金字塔池化（ASPP）模块捕获上下文信息，在不降低特征分辨率的情况下扩大感受野。此外，在解码器中使用了一种平滑的激活函数 Mish[12]。Mish 具有平滑、非单调、无上界、有下界等特点，在深度神经网络中表现出了比 ReLU 更好的效果。

2　相关工作

图像语义分割常用编码器-解码器[13]结构预测端到端的像素级分类任务。编码器用于提取图像的高级语义特征，解码器则通过反卷积、插值等方式恢复原图尺寸，最终获得图像分割结果。目前分割较好的网络结构往往在解码阶段采用融合高分辨的低层特征来获得图像的空间信息，实现更精细的分割效果。

所提方法同样基于编解码结构，提出了多尺度特征对齐聚合（MFAA）模块。所提模型结构如图 1 所

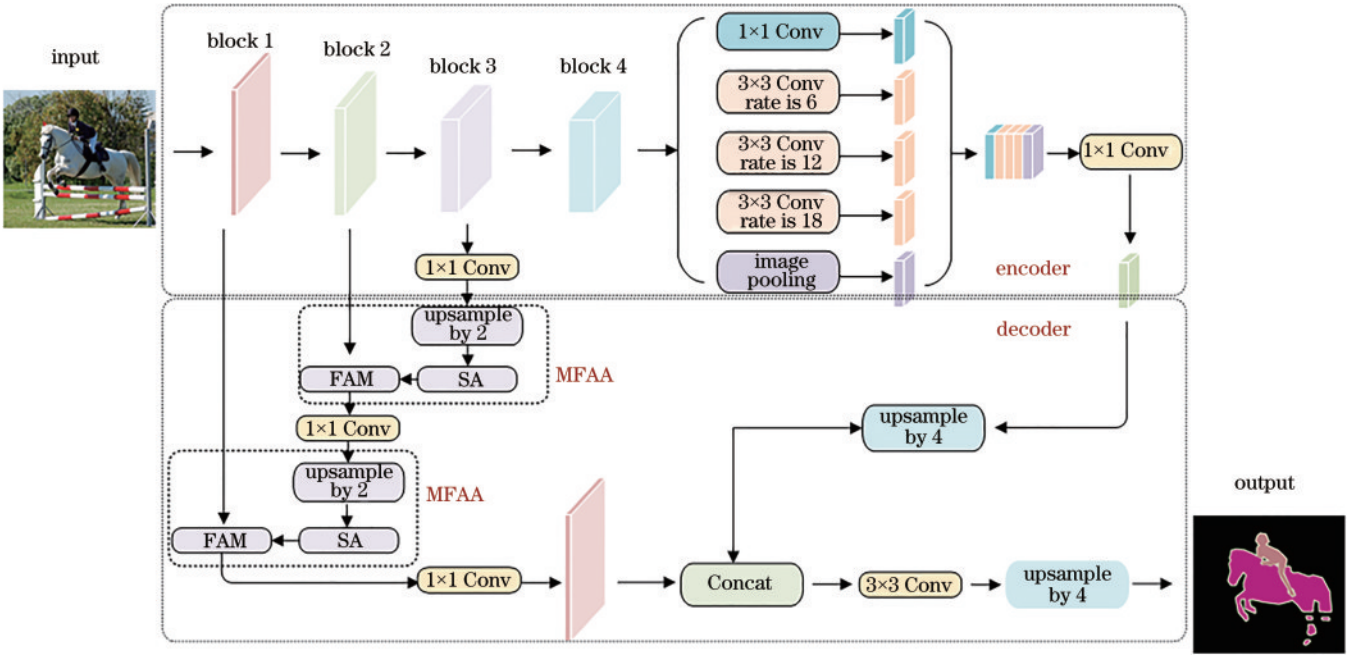

图 1　所提模型结构图

Fig. 1　Structure diagram of proposed model

示，编码器使用ResNet-50作为骨干网络来提取特征，下采样步幅为16，输入图像经过ResNet-50提取特征后得到尺寸为输入图像尺寸1/16的高级语义特征，然后经过上下文模块ASPP获得多种尺度的上下文特征。解码器中使用两个MFAA模块对骨干网络中不同分辨率的特征图进行对齐聚合，得到的输出特征与ASPP输出的高级特征进行级联，最后通过上采样获得密集的图像分割结果。

2.1　MFAA

MFAA由特征对齐模块(FAM)和空间注意力(SA)模块组成，是整个网络解码器的重要组成部分。MFAA通过SA模块突出高层输入的重要空间细节，使高层特征在保留语义信息的情况下尽可能激活更多的空间信息。FAM有两个输入，用于高层特征与低层特征之间的对齐与融合。高层特征首先进行二倍上采样，之后经过SA模块后与低层特征通过FAM模块对齐再进行相加融合。输入图像经过骨干网络后有不同分辨率的特征输出，为了获得更好的融合效果，可以使用多个MFAA模块串联进行多尺度的特征融合。所提方法使用两个MFAA模块串联实现不同层的特征融合。

2.1.1　FAM

FAM采用一种可学习插值策略来学习像素的变换偏移，用于精确对齐高分辨和低分辨率的特征图，之后聚合高级特征和低级特征。特征融合的错位来自两个输入特征之间的偏移，特征经过FAM学习到高级特征与低级特征的偏移量，之后与各自的输入特征经过对齐函数获得矫正后的特征信息再进行融合。FAM模块如图2所示，高级特征上采样后经过空间注意力模块得到F_{h}，F_{h}与低级特征F_{l}通过级联(Concat)来建立两个特征之间的相关性，之后经过1×1卷积与批量归一化层后分成两个支路，每个支路使用1×1卷积将通道维度降为2，用来预测该支路特征的二维偏移$\Delta\in\mathbf{R}^{2\times H\times W}$，$\Delta$的两维分别代表了特征的横向偏移与纵向偏移。两条支路输出的二维偏移Δ_{h}和Δ_{l}分别用于对齐高级特征和低级特征，通过函数U获得对齐之后的特征信息：

$$A_{\mathrm{out}}=U(F_{\mathrm{h}},\Delta_{\mathrm{h}})+U(F_{\mathrm{l}},\Delta_{\mathrm{l}}),\tag{1}$$

式中：A_{out}是对齐之后的输出特征；F_{h}和F_{l}是需要对齐的两个输入特征；U是对齐函数。假设要对齐的特征图F的大小为$H\times W$，F上像素点F_{hw}的坐标为(h,w)，Δ_{1hw}和Δ_{2hw}分别是对F_{hw}预测的纵向偏移与横向偏移，则像素点F_{hw}对齐之后的期望输出坐标是$(h+\Delta_{1hw},w+\Delta_{2hw})$，由函数$U$可得该像素点的期望输出为

$$U_{hw}=\sum_{h'=1}^{H}\sum_{w'=1}^{W}F_{h'w'}\cdot\max\left(0,\ 1-\left|h+\Delta_{1hw}-h'\right|\right)\cdot\max\left(0,\ 1-\left|w+\Delta_{2hw}-w'\right|\right)。\tag{2}$$

由式(2)可知，对齐函数的输出在全图范围内求和，但将公式中的绝对值展开后发现，只有点$(h+\Delta_{1hw},w+\Delta_{2hw})$周围最近的4个点(左上、左下、右上、右下)对权重有贡献，对齐之后的新像素点是根据期望输出坐标附近的4个点按距离权重双线性插值后得到的。未采用更简单的最近邻插值是因为二维偏移量在大多数情况下不是整数，如果根据最近邻直接赋值给新像素点那么该映射不可导，无法满足反向传播的条件。

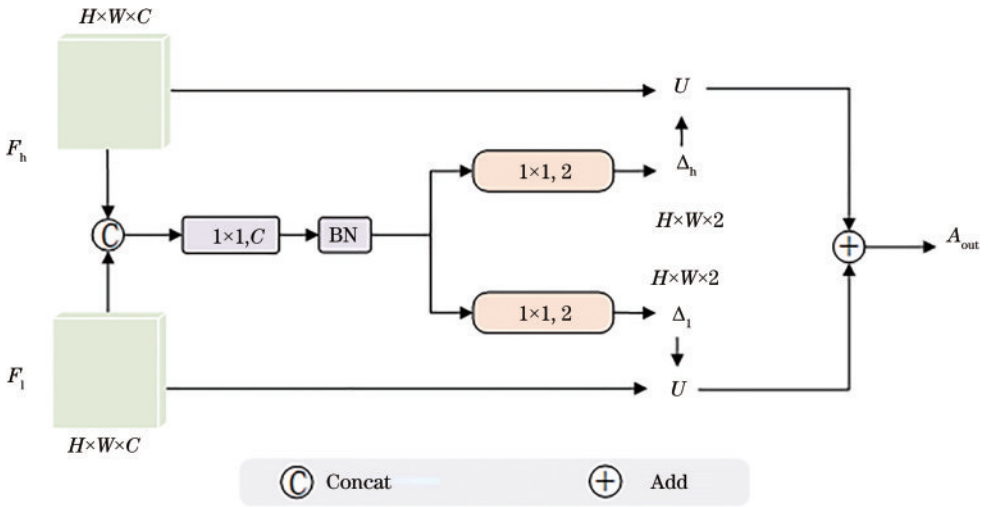

图2　FAM结构图

Fig. 2　FAM structure diagram

2.1.2　SA

SA模块通过加权的方式增强图像上有用的关键信息并抑制其他的信息，使模型聚焦于特征图中感兴趣的特征空间区域，突出高层输入的重要空间细节。

SA模块采用平均池化和最大池化来聚合通道信息，分别关注了特征图的全局特征和突出特征。SA模

块结构如图 3 所示，输入特征首先经过两个池化模块得到通道上的平均池化特征 F_{avg} 和最大池化特征 F_{max}，然后将 F_{avg} 和 F_{max} 在通道维度进行拼接，接着使用 3×3 卷积融合特征将通道维度降为 1，最后经过一个 Sigmoid 激活函数对特征的注意力权重归一化，对要强调或抑制的位置进行编码。经过 SA 模块的特征输出为

$$S(F)=\mathrm{Sigmoid}\left\{\mathrm{Conv}^{3\times3}\left\{\left[F_{avg}(F);F_{max}(F)\right]\right\}\right\},\tag{3}$$

式中：F 代表输入特征。

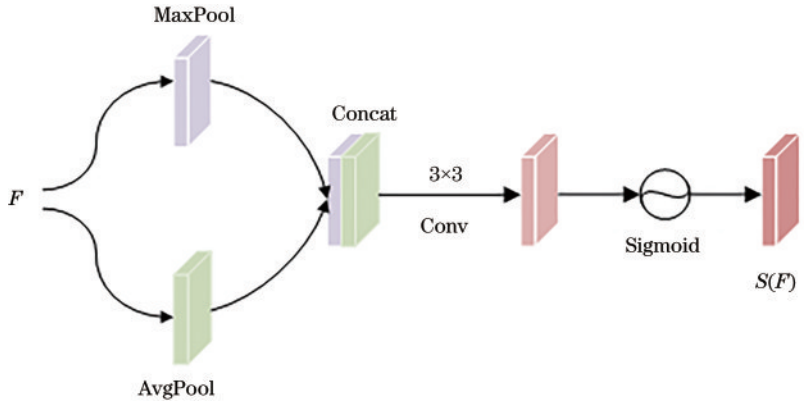

图 3　SA 模块结构图

Fig. 3　SA module structure diagram

2.2　ASPP

空洞卷积是 ASPP 模块的关键，在标准卷积中填充空洞从而增大卷积核的尺寸，如图 4 所示。空洞卷积可以在不降低特征图分辨率、不增加参数量的情况下增大感受野。感受野大小可以根据空洞率调节，便于提取不同尺度的特征。空洞率与感受野对应关系为

$$D=\sum_{i=1}^{n}D_i-(n-1),\tag{4}$$

式中：D_i 表示第 i 个卷积的感受野范围；n 代表级联卷积的个数。

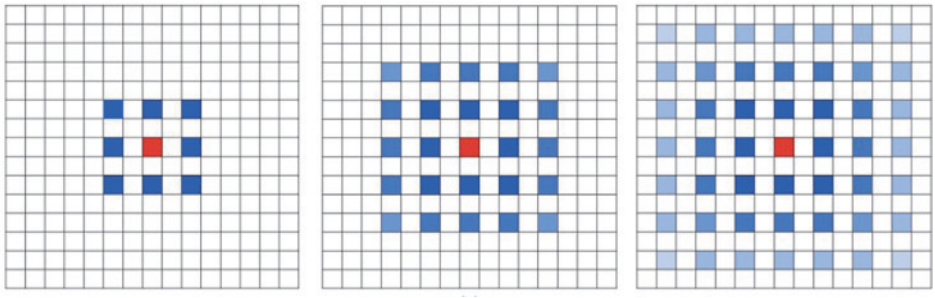

图 4　空洞卷积

Fig. 4　Dilated convolution

Deeplab[14-17]系列语义分割模型中的关键模块就是 ASPP，ASPP 最先在 Deeplab v2[15]中被提出，之后 Deeplab v3[16]和 Deeplab v3+[17]对 ASPP 进行了优化。ASPP 采用不同空洞率的空洞卷积来提取多尺度信息，由 5 条并行支路组成，如图 5 所示：第 1 层是 1×1 卷积，将通道数降为 256；第 2、3、4 层是拥有不同空洞率的 3×3 空洞卷积，对于 output_stride 为 16 的骨干网络，空洞率分别是 6、12、18；第 5 层首先是一个全局平均池化，然后通过双线性插值恢复原始大小，该层是为解决 ASPP 设置的空洞率过大时，有效的滤波参数减小，空洞卷积会退化为 1×1 卷积的问题。将 5 个并行支路得到的输出特征进行 Concat 后将通道数降为 256，就得到了拥有不同尺度的特征图。

2.3　Mish 激活函数

激活函数的作用是在神经网络中引入非线性。在深度学习中，常用的激活函数主要有 ReLU、Sigmoid、Tanh、PReLU[18]、Swish[19]等。所提方法在解码器中使用了一种新的激活函数 Mish，如图 6 所示。Mish 是一种自正则的非单调激活函数，与 ReLU 激活函数相比具有如下特点：1）Mish 函数允许较小的负梯度流入，因此保证了信息的流动，有效缓解了 ReLU 激活函数反向传播过程中梯度消失的问题；2）Mish 函数中每一个点都是平滑的，平滑的激活函数可更好地允许信息深入神经网络，梯度下降效果更好；3）Mish 函数有下界但没有上界，没有上界可以避免饱和，有下界保证了一定的正则化效果。Mish 函数的表达式为

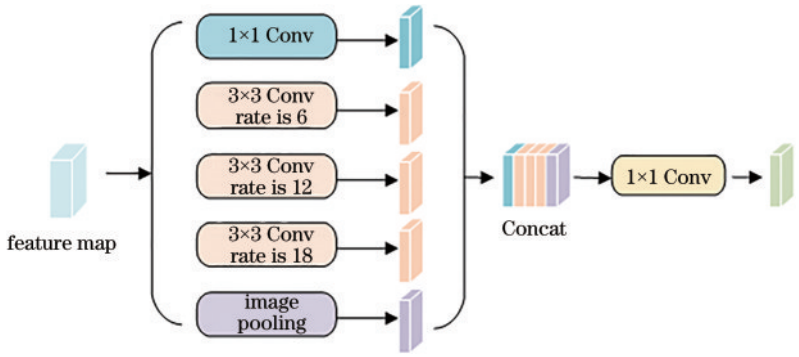

图 5　空洞空间金字塔池化模块

Fig. 5　Atrous spatial pyramid pooling

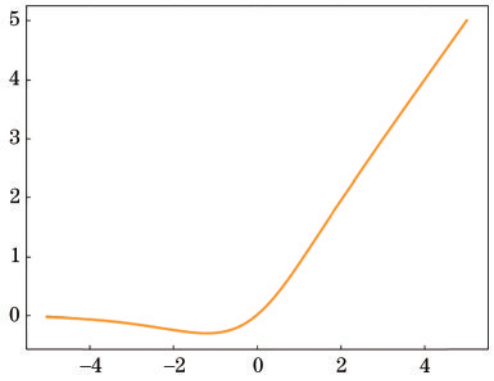

图 6　Mish 激活函数

Fig. 6　Mish activation function

$$\mathrm{Mish}(x)=x\times\tanh\left\{\ln\left[1+\exp(x)\right]\right\}。\tag{5}$$

3　实验与分析

3.1　实验分析

实验使用 PASCAL VOC 2012 增强版数据集。PASCAL VOC 2012 增强版数据集是图像语义分割领域最常用的公共数据集之一，总共包含 21 个类别，20 个前景类和 1 个背景类。数据集总共包含 10582 张训练集，1449 张验证集和 1456 张测试集。

实验程序基于深度学习框架 PyTorch 实现。实验硬件配置为 AMD EPYC 7302 处理器，NVIDIA GeForce RTX 3090 显卡。

3.2　实验参数设置及评价指标

实验选用 ResNet-50 作为网络主干，输入图片大小设置为 513×513，batch size 设置为 16，使用交叉熵损失函数。学习率衰减采用“poly”学习策略：

$$R_{\mathrm{lr}}=R_{\mathrm{base\_lr}}\left(1-\frac{N_{\mathrm{iter}}}{N_{\mathrm{max\_iter}}}\right)^{\mathrm{power}},\tag{6}$$

式中：power 设置为 0.9；初始学习率 $R_{\mathrm{base\_lr}}$ 设置为 0.01；N_{iter} 代表训练时每一次迭代的迭代次数；$N_{\mathrm{max\_iter}}$ 代表总的迭代次数，设为 30000。

采用平均交并比（mIoU）作为模型的评价指标。mIoU 是真实值和预测值两个集合的交集和并集之比，表示分割结果与其真值的重合度，是图像语义分割最常用的评价指标。mIoU 的表达式为

$$R_{\mathrm{mIoU}}=\frac{1}{K}\left(\sum_{x=1}^{K}\frac{T_{xx}}{\sum_{y=1}^{K}T_{xy}+\sum_{y=1}^{K}T_{yx}-T_{xx}}\right),\tag{7}$$

式中：K 是图像语义分割标签的总类别数；T_{xy} 代表像素类别是 x 类却被预测为 y 类的像素总数；T_{yx} 代表像素类别是 y 类却被预测为 x 类的像素总数，T_{xx} 代表像素类别是 x 类预测类别也是 x 的像素总数。

3.3　特征选择与融合实验结果

为验证 MFAA 模块相较于一般特征融合方法的有效性，在 PASCAL VOC 2012 增强版数据集上进行了 5 组对比实验，每组实验分别使用 MFAA 模块、特征相加（Add）和特征级联（Concat）的特征融合策略进行对比。此外，骨干网络中不同层提取的特征之间存在差异，为获得最优的融合效果，从 ResNet-50 的 4 个 Block 中选取不同尺度 Block 输出的特征进行对比实验，结果如表 1 所示。

表 1　不同特征选择的 mIoU

Table 1　mIoU with different feature selection　unit: %

Block selection	Add	Concat	MFAA
1,2	77.5	77.5	**77.6**
1,3	77.7	77.6	**78.0**
1,4	77.5	77.3	**77.9**
1,2,3	77.6	77.8	**78.4**
1,2,4	77.5	77.7	**78.4**

由表 1 可知，在相同特征进行融合的情况下，无论使用 Add 还是 Concat 特征融合方法，其分割精度皆低于使用 MFAA 模块进行融合的精度。在第 1 组实验中，使用骨干网络 Block 1 和 Block 2 的输出特征进行融合，MFAA 融合方法的 mIoU 为 77.6%，相比 Add 和 Concat 提升了 0.1 个百分点。随着融合的多尺度特征越多，特征之间的距离越远，MFAA 模块的优势就越明显。在第 4 组实验中，MFAA 相比 Add 和 Concat 的分割精度分别提升了 0.8 个百分点和 0.6 个百分点，原因是骨干网络中距离相近的特征之间经过的下采样

和padding操作较少，特征不对齐问题还不明显，随着融合的特征越多，特征之间的距离越远，特征融合时这种错位就越严重，而使用MFAA模块可以有效缓解特征不对齐问题，提高分割效果。由实验结果可知，MFAA融合方法相比Add和Concat的分割精度最高提升了0.8个百分点。

表1中的5组实验分别选取了不同尺度的特征进行融合，由实验结果可知，使用Add和Concat的方法进行多尺度特征融合时，5组实验之间的mIoU分数相近，说明简单的特征融合方法由于特征不对齐的原因，不能充分聚合特征。在使用MFAA模块时，融合两个不同尺度低层特征的mIoU分数为77.6%，融合3个不同尺度特征时MFAA方法的mIoU分数最高达到了78.4%。由于第4组和第5组实验的分割精度相同，且Block 4输出特征的通道维度是Block 3输出特征的两倍，为了降低参数量和计算量，所提方法选取第4组中的特征组合，使用MFAA模块融合骨干网络中前三个Block的输出特征。

3.4 FAM和SA模块实验结果

MFAA由FAM和SA模块组成，为了分析MFAA中两个模块的有效性，选用在第3.3节最后提到的特征选择，融合骨干网络中前三个Block的输出特征在PASCAL VOC 2012增强版数据集上进行了4组实验。第1组实验去掉了MFAA模块，第2、3组实验分别保留MFAA中的SA和FAM模块，最后1组使用完整的MFAA模块进行实验。

表2 MFAA模块拆分实验结果

Table 2 Experimental results of MFAA module disassembly

Module selection	mIoU /%
Baseline	77.2
SA	77.7
FAM	77.8
MFAA(SA+FAM)	78.4

由表2可知，加入SA和FAM的mIoU精度分别为77.7%和77.8%，比第1组的分割精度提升了0.5个百分点和0.6个百分点，这说明两个模块都促进了不同尺度特征之间的融合。将两个模块同时加入网络时，mIoU达到了78.4%，在第2、3组的基础上进一步提高了0.7个百分点和0.6个百分点的分割精度，这是因为FAM可以对齐特征，SA模块可以增大重要空间信息的权重，两个模块同时使用时去除了高低层特征之间的错位偏差，突出特征重要空间信息，更好地聚合了高层语义特征和浅层低级特征。

3.5 消融实验

为检验各模块对模型分割精度的影响，基于相同的实验环境和实验参数，使用ResNet-50作为主干网络在PASCAL VOC 2012增强版数据集上进行消融实验。由前4组实验结果可知，在基础网络上增加Mish激活函数、MFAA模块及ASPP模块皆能有效增加模型的分割精度。ASPP使用6、12、18的空洞率组合，可以有效扩大感受野，提高分割精度。增加MFAA模块可以有效聚合多尺度特征，进一步改善分割效果，ASPP+MFAA的组合相比单独使用ASPP模块的分割精度增加了1.2个百分点。Mish与ReLU相比更加平滑，梯度下降效果更好，由后4组实验可知，在模型中使用Mish激活函数可以使mIoU进一步提升0.3个百分点。

表3 消融实验

Table 3 Ablation results

Group	mIoU /%
Baseline	63.8
MFAA	66.7
ASPP	76.9
ASPP+Mish	77.2
ASPP+MFAA	78.1
ASPP+MFAA+Mish	78.4

3.6 与其他分割方法的对比

自从FCN被提出以来，出现了一大批优秀的语义分割网络。将所提方法与几种主流的分割方法进行了对比，来进一步验证其有效性，结果如表4所示。FCN是深度学习用于语义分割的开创之作，确立了图像语义分割通用网络模型架构。SegNet[20]在FCN的基础上搭建了一个对称的网络结构，是典型的编-解码分割网络之一。由表4可知，所提方法的分割精度相比FCN和SegNet分别提升了8.7个百分点和6.8个百分点。Deeplab v3和PSPNet[21]分别将空洞卷积和池化层引入上下文模块，融合了多尺度特征，是基于多尺度预测的代表性分割算法[22]。所提方法与拥有更深网络主干的Deeplab v3和PSPNet相比，使用ResNet-50提取特征，分割精度分别提升了1.2个百分点和0.4个百分点，证明了所提方法的优越性。

表4 对比实验结果

Table 4 Results of comparative experiments

Algorithm	FCN	SegNet	Deeplab v3	PSPNet	Proposed algorithm
mIoU /%	69.7	71.6	77.2	78.0	78.4

为了更直观地展示模型的分割效果，将所提方法与上述分割方法在PASCAL VOC 2012数据集上进行了进一步实验，部分可视化结果如图7所示。从图中可以看出，对于轮廓分明且不受遮挡的物体，各方法的分割效果都尚可，例如在第1幅图中，4种方法对猫的分割都比较准确。但与其他几种方法相比，所提方法对物体细节的分割更为清晰。在第3幅图中，FCN受马后背阴影的影响未能将马精确分割，DeepLab v3和PSPNet虽然分割出了马的轮廓，但对边缘的处理

图 7　不同分割方法的可视化结果。(a)输入图片;(b)标签;(c)FCN 分割结果;(d)Deeplab v3 分割结果;(e)PSPNet 分割结果;(f)所提方法分割结果

Fig. 7　Visualization results of different segmentation methods. (a) Input images; (b) labels; (c) segmentation results of FCN; (d) segmentation results of Deeplab v3; (e) segmentation results of PSPNet; (f) segmentation results of proposed algorithm

不够充分,而所提方法精确分割出了马的轮廓,对边缘的处理也较好。在第 5 幅图中,由于受到马鞍的影响,其他方法无法将人的腿部精确分割,所提方法分割出了少部分腿部,但对人头的分割不够精确。在第 4 幅图与第 6 幅图中,其他方法虽然都分割出了飞机的机翼,但对机翼上导航灯的分割较为模糊,而所提方法精确分割出了导航灯。综上可知,所提方法对图像边缘轮廓等细节特征的分割较准确,有效改善了图像的分割效果。

4　结　　论

针对现有语义分割算法进行不同尺度特征融合时出现的特征不对齐问题,提出了一种多尺度特征对齐聚合的语义分割方法。该方法利用 MFAA 模块获得不同尺度特征之间的偏移量,经过矫正后再进行融合,充分利用各层级特征,提高了特征融合的效果;同时使用 ASPP 提取上下文信息,在不降低特征分辨率的情况下扩大感受野。解码网络采用了一种平滑、性能更好的 Mish 激活函数,进一步提高了分割效果。实验结果表明,所提方法可以有效聚合不同尺度特征,提高图像分割精度。在未来的研究中考虑设计更轻量的网络用于图像分割,在保证分割精度的前提下尽可能减少模型参数。

参　考　文　献

[1] 王龙飞, 严春满. 道路场景语义分割综述[J]. 激光与光电子学进展, 2021, 58(12): 1200002.
Wang L F, Yan C M. Review on semantic segmentation of road scenes[J]. Laser & Optoelectronics Progress, 2021, 58(12): 1200002.

[2] Csurka G, Perronnin F. An efficient approach to semantic segmentation[J]. International Journal of Computer Vision, 2011, 95(2): 198-212.

[3] 张欢, 仇大伟, 冯毅博, 等. U-Net 模型改进及其在医学图像分割上的研究综述[J]. 激光与光电子学进展, 2022, 59(2): 1200002.
Zhang H, Qiu D W, Feng Y B, et al. Improved U-net models and its applications in medical image segmentation: a review[J]. Laser & Optoelectronics Progress, 2022, 59(2): 1200002.

[4] 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.
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.

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

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-05-10)[2021-10-10]. https://arxiv.org/abs/1409.1556.

[7] Lu H, Dai Y T, Shen C H, et al. Indices matter: learning to index for deep image matting[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 3265-3274.

[8] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]//2015 Conference and Workshop on Neural Information Processing Systems, December 7-12, 2015, Montreal, Quebec, Canada. [S. l.: s. n.], 2015: 2017-2025.

[9] Mazzini D. Guided upsampling network for real-time semantic segmentation[C]//2018 British Machine Vision Conference(BMVC), September 3-6, 2018, Newcastle, UK. London: BMVA Press, 2018: 117.

[10] Li X T, You A S, Zhu Z, et al. Semantic flow for fast and accurate scene parsing[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 775-793.

[11] Huang Z L, Wei Y C, Wang X G, et al. AlignSeg: feature-aligned segmentation networks[EB/OL]. (2021-05-02)[2021-10-10]. https://arxiv.org/abs/2003.00872.

[12] Misra D. Mish: a self regularized non-monotonic activation function[EB/OL]. (2019-08-23) [2021-10-10]. https://arxiv.org/abs/1908.08681.

[13] 张哲晗, 方薇, 杜丽丽, 等. 基于编码-解码卷积神经网络的遥感图像语义分割[J]. 光学学报, 2020, 40(3): 0310001.
Zhang Z H, Fang W, Du L L, et al. Semantic segmentation of remote sensing image based on encoder-decoder convolutional neural network[J]. Acta Optica Sinica, 2020, 40(3): 0310001.

[14] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[EB/OL]. (2014-12-22) [2021-10-10]. https://arxiv.org/abs/1412.7062.

[15] 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.

[16] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05) [2021-10-10]. https://arxiv. org/abs/1706.05587.

[17] 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.

[18] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1026-1034.

[19] Ramachandran P, Zoph B, Le Q V. Swish: a self-gated active function[EB/OL]. (2017-10-27) [2021-10-10]. https://arxiv.org/abs/1710.05941v1.

[20] 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.

[21] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6230-6239.

[22] 赵霞, 白雨, 倪颖婷, 等. 基于深度学习的语义分割算法综述[J]. 上海航天, 2019, 36(5): 71-82.
Zhao X, Bai Y, Ni Y T, et al. A review of semantic segmentation algorithm based on deep learning[J]. Aerospace Shanghai, 2019, 36(5): 71-82.