激光与光电子学进展 先进成像

面向无人机航摄图像语义分割的双路特征融合网络

李润增[1]，史再峰[1,3*]，孔凡宁[1]，赵向阳[1]，罗韬[2]
[1]天津大学微电子学院，天津 300072；
[2]天津大学智能与计算学部，天津 300072；
[3]天津市成像与感知微电子技术重点实验室，天津 300072

摘要 针对无人机航摄图像中目标尺寸差异大导致的感受野难以同时兼顾不同尺寸物体分割效果的问题，提出了利用两路分支分别提取浅层和深层信息的双路特征融合网络(DSFA-Net)。在编码器中，浅层分支利用三个串行ConvNeXt模块提取高通道数的浅层特征以保留更多空间细节；深层分支利用坐标注意力空洞空间金字塔池化(CA-ASPP)模块为特征图重新分配权重，使网络更加关注尺寸各异的分割目标，获得深层多尺度特征。在解码过程中，网络利用双边引导融合模块为两层特征建立通信以进行分辨率融合，提高层级特征的利用率。所提方法在AeroScapes和Semantic Drone航摄图像数据集上进行了实验，其平均交并比分别达到83.16%和72.09%、平均像素准确率分别达到90.75%和80.34%。与主流的语义分割方法相比，所提方法对于具有较大尺寸差异的目标，分割能力更强，更适用于无人机航摄图像场景下的语义分割任务。

关键词 语义分割；特征融合；双路网络；坐标注意力空洞空间金字塔池化；多尺度特征提取
中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP230955

Dual-Stream Feature Aggregation Network for Unmanned Aerial Vehicle Aerial Images Semantic Segmentation

Li Runzeng[1], Shi Zaifeng[1,3*], Kong Fanning[1], Zhao Xiangyang[1], Luo Tao[2]
[1]*School of Microelectronics, Tianjin University, Tianjin* 300072, *China*;
[2]*College of Intelligence and Computing, Tianjin University, Tianjin* 300072, *China*;
[3]*Tianjin Key Laboratory of Imaging and Sensing Microelectronic Technology, Tianjin* 300072, *China*

Abstract Large object size difference in unmanned aerial vehicle (UAV) aerial photography makes it difficult to take into account the segmentation effect of objects of different sizes in the receptive field. A dual-stream feature aggregation network (DSFA-Net) with two branches to extract low-level and high-level features separately, is proposed for such problems. In the encoder, a low-level information extraction branch with three serial ConvNeXt modules is used to preserve more low-level features by generating more channels of features. In the deep feature branch, the coordinate attention atrous spatial pyramid pooling (CA-ASPP) module reassigns weights to feature maps in the channel dimension. It makes the module focus on segmentation objects of different sizes and deep-level multi-scale features are obtained. During the decoding process, the bilateral guided aggregation module performs resolution aggregation between the low-level and deep-level features. Our method is evaluated on the AeroScapes and Semantic Drone datasets, the mean intersection over union is 83.16% and 72.09% respectively, and the mean pixel accuracy is 90.75% and 80.34% respectively. The proposed method is more capable of segmenting objects with large difference sizes compared to mainstream methods. It is suitable for semantic segmentation tasks for UAV aerial images.

Key words semantic segmentation; feature aggregation; dual-stream architecture; coordinate attention atrous spatial pyramid pooling; multi-scale feature extraction

1 引 言

随着无人机(UAV)的广泛应用，对UAV航摄图像的分析也越来越重要，航摄图像分析依赖语义分割技术，语义分割的目的是将图像中的像素分类并赋予对应类别的标签[1]，以便进行下游图像分析任务。近年来随着卷积神经网络的发展和图形处理器(GPU)算力的提升，基于深度学习的语义分割技术取得了巨

收稿日期：2023-03-27；**修回日期**：2023-04-15；**录用日期**：2023-04-23；**网络首发日期**：2023-05-03
基金项目：国家自然科学基金(62071326)、天津市自然科学基金(22JCYBJC00140)
通信作者：*shizaifeng@tju.edu.cn

大进展[2]。全卷积神经网络(FCN)[3]用卷积层替代全连接层,使用端到端的方式训练并进行语义分割。U-Net[4]采用了U型的编解码器结构,并在其中使用跳跃连接填补浅层信息。SegNet[5]使用了与编码器空间尺寸和通道数对称的解码器,利用最大池化索引来保存位置信息并在上采样时使用。

UAV航摄图像由于拍摄视角相对开阔,因此场景复杂、目标尺寸差异大。简单编解码器结构网络的感受野大小固定,对于大目标来说,上下文信息获取不够,容易导致大目标被分割为具有不同标签的多个部分;对于小目标来说,浅层的空间信息保留太少,导致小目标缺少细节甚至被忽略[6]。多尺度特征提取的方法可以一定程度上兼顾不同尺寸的目标。PSPNet(pyramid scene parseing network)[7]和RefineNet[8]分别使用相对独立的模块进行多尺度的特征提取;Panoptic FPN(feature pyramid network)[9]使用专门的多尺度特征提取网络来获取上下文信息;DeepLabV3+[10]提出了空洞空间金字塔池化(ASPP)模块,在保持分辨率的基础上获取了不同尺寸的感受野;Panoptic-Deeplab[11]使用两个解耦的ASPP提取上下文信息;在ASPP基础上,DenseASPP[12]引入密集连接,提高了其特征提取能力;ECANet(efficient channel attention network)[13]利用金字塔空间注意力模块捕获全局上下文;MagNet[14]采用了多尺度框架,每个阶段以不同分辨率进行由粗到细的特征传播。虽然多尺度特征提取一定程度上提高了网络对不同尺寸目标的分割能力,但是多尺度模块也有局限性:第一,模块通常处理主干网络输出的深层特征图,而在下采样过程中大量浅层空间信息就已丢失,因此在目标尺寸差异大的UAV图像的分割任务中,难以使小物体得到细粒度的分割效果;第二,随着多尺度层数的递增,效果会逐渐趋于饱和[15],且参数量的增加导致信息过载,使网络难以重点关注目标区域。

对于第一个问题,本文利用以ConvNeXt模块[16]为主体构成的浅层信息提取分支,对浅层特征进行提取,弥补主干网络下采样过程中造成的信息丢失,保留更多空间细节以提高网络对小目标的分割能力。对于第二个问题,提高多尺度特征的利用率是解决此问题的有效方法[17],将注意力机制引入ASPP,得到坐标注意力[18]空洞空间金字塔池化(CA-ASPP)模块,可以缓解信息过载问题,使模块重点关注目标所在区域,有利于兼顾不同尺寸的目标。在此基础上,利用双边引导融合(BGA)模块[19]将得到的浅层特征与CA-ASPP模块输出的深层特征进行多尺度融合,在深浅层特征之间建立通信并提高浅层信息的利用率,也缓解了因感受野尺寸限制带来的无法兼顾不同尺寸目标的问题。

2 双路特征融合语义分割网络

2.1 网络结构

针对UAV航摄图像目标尺寸差异大、感受野难以同时兼顾不同尺寸目标的问题,提出了双路特征融合语义分割网络(DSFA-Net),其主体为编码器-解码器结构,如图1(a)所示,其中k和s分别为卷积核尺寸和卷积步长。编码器由两条分支构成,第一条分支为主干网络Xception和CA-ASPP模块;第二条分支是以ConvNeXt模块为主体的浅层信息提取分支。输出特征图包含更大的通道数,专用于提取浅层特征。解码器包括双边引导融合模块和线性插值上采样模块,其中双边引导融合模块将提取到的浅层特征图与CA-ASPP模块输出的深层特征图进行更高效的融合。此结构针对多尺度提取、浅层特征获取等问题作出了改进,以提高网络模型在UAV航摄图像语义分割任务中的性能。

2.2 浅层信息提取分支

UAV航摄图像场景中,因为目标尺寸之间的差异大,除了利用多尺度特征提取模块以及多个尺寸各异的感受野之外,通过保留浅层空间信息的方式来提高小目标的分割效果也是简单且有效的。DeepLabV3+在主干网络的第二个模块提取了一个原图1/4尺寸的中间特征图作为浅层特征在编码器中与多尺度特征提取模块的结果进行合并。但是主干网络将输入图像下采样到1/2和1/4尺寸时分别只有64通道和128通道,较低的通道数也就意味着较少的浅层空间信息。为了获取更多的浅层特征,增加小目标的分割细节,将一个浅层信息提取分支作为主干网络的附属分支,用于提取浅层空间特征[图1(a)]。浅层信息提取分支以ConvNeXt模块为主体,输入特征图为原图的1/4尺寸,且通道数从下采样开始即保持在256,极大程度地保留了浅层空间信息。此分支的输出作为中间特征图,在解码器中与深层特征进行空间分辨率融合。

ConvNeXt模块参考ResNet(residual network)[20]和Transformer等模型的思想进行了一系列改进,使用ConvNeXt模块中的stem模块和ConvNeXt模块组成了浅层信息提取分支,ConvNeXt模块结构如图1(b)所示,其中p为填充大小。输入图像首先进入stem模块,这里改用一个步长为4的卷积,进行4倍下采样。紧随其后的是一个LN(layer normalization)层,取代了常用的BatchNorm层。接下来利用3个相同的ConvNeXt模块进行特征提取。在ResNet的每个残差块中,使用通道数先减小再增大的瓶颈结构(bottleneck),而在ConvNeXt模块中采用了MobileNetv2中的反转瓶颈结构(inverted bottleneck),通道数先增大后减小,这样做的目的是抵消激活函数层带来的信息损失[21]。同时在第一个卷积层进行深度可分离卷积(depth-wise convolution),其组数与通道数相等。第二个卷积层将通道数调整为原始的4倍,并在最后一个卷积层再将其还原。除此之外,ConvNeXt模块中,更为平滑的激活函数GELU(Gaussian error linear unit)取代了ReLU(rectified linear unit),利用Layer Scale层对特征图进行

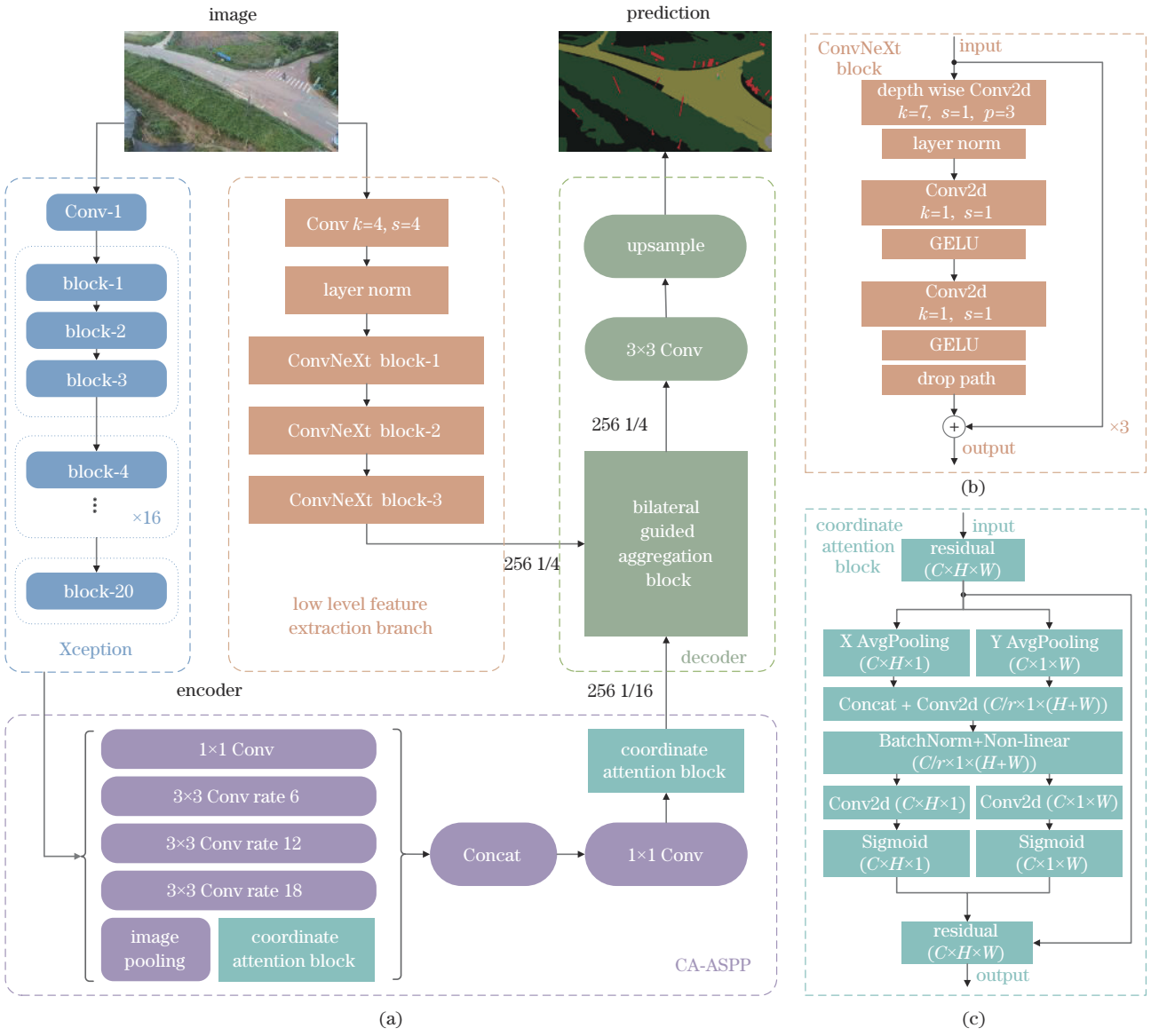

图 1　网络结构图及部分模块图。(a)网络整体结构;(b) ConvNeXt模块;(c)坐标注意力模块

Fig. 1　Network architecture and partial module. (a) Overall network architecture; (b) ConvNeXt block; (c) coordinate attention block

缩放,利用Drop Path进行正则化。最后把上述操作处理得到的特征图与整个模块的输入特征图叠加,作为此模块的输出。

经过浅层信息提取分支的处理,得到尺寸为原图的1/4、通道数为256的输出特征图。此特征图相较于从主干网络Xception中提取到的中间特征图,包含了更多浅层空间信息。将此特征图融合到解码过程中,有利于在预测结果中恢复更多的空间细节。

2.3　CA-ASPP模块

UAV航摄场景中目标尺寸差异大的问题对ASPP的多尺度特征提取能力提出了更高的要求。但是在ASPP中,层数的简单堆叠与其他的空洞率组合对其特征提取能力的提升已经达到饱和,甚至起到了反作用[22]。但是,ASPP本身具有较大的参数量,其中也包含了大量的重要性较低甚至无用的信息。而注意力机制可以计算输入信息的加权平均,为其分配不同的权重,使网络在接下来的特征提取中关注重点目标区域,解决ASPP性能饱和的问题。在ASPP中的全局池化层和合并卷积层后各使用一个坐标注意力机制模块,组成包含坐标注意力机制的CA-ASPP模块,此结构经过消融实验得到。相比传统的注意力机制,坐标注意力机制不仅可以获取跨通道的信息,而且能够获取两个方向的位置敏感信息,有助于模型定位和识别感兴趣的目标和区域。

坐标注意力模块的结构如图1(c)所示,其中:C、H和W分别为特征图的通道数、高度和宽度;r为下采样比例;AvgPooling表示平均池化。坐标注意力模块的原理主要分为两个部分,坐标信息嵌入和坐标注意力生成。全局池化提取的特征被压缩到通道注意力后,其位置信息是难以保留的,所以在坐标信息嵌入时,坐标注意力模块将全局池化操作分解为在x和y两个方向进行一维特征编码。对于输入$\boldsymbol{x}$,用尺寸为(H,

1)和(1, W)的池化核分别沿着水平和垂直方向对两个通道编码得到高度为 h 的第 c 个通道的输出

$$z_c^{(h)}(h)=\frac{1}{W}\sum_{0\leqslant i<W}x_c(h,i), \tag{1}$$

同理可以得到宽度为 w 的第 c 个通道的输出

$$z_c^{(w)}(w)=\frac{1}{H}\sum_{0\leqslant j<H}x_c(j,w)。 \tag{2}$$

由以上两式即可得到聚合了两个方向特征的特征图。同时,这两个转换还使得注意力模块沿一个方向获取了远程依赖,沿另一个方向保留了位置信息,完成了坐标信息的嵌入。想要有效生成注意力,既要充分获取感兴趣区域的位置信息,又要有效地捕捉通道之间的关系。首先,将式(1)、式(2)已经生成的特征图进行合并,再进行1×1卷积变换[用 $F_1(\cdot)$ 表示],得到

$$f=\delta\left\{F_1\left\{\left[z^{(h)},z^{(w)}\right]\right\}\right\}, \tag{3}$$

式中:$\delta(\cdot)$ 为非线性激活函数;生成的 $f\in\mathbb{R}^{[C/r\times(H+W)]}$ 为编码了水平和垂直两个方向的中间特征图,其中下采样比例 r 用于控制模块的尺寸。然后,再次沿着空间维度将 f 拆分为两个独立的张量 $f^{(h)}\in\mathbb{R}^{(C/r\times H)}$ 和 $f^{(w)}\in\mathbb{R}^{(C/r\times W)}$,再分别进行1×1卷积 $F_h(\cdot)$、$F_w(\cdot)$ 以及激活函数 $\delta(\cdot)$ 计算,将特征图 $f^{(h)}$ 和 $f^{(w)}$ 变换到与输入 x 通道数相同,表达式为

$$g^{(h)}=\delta\left\{F_h\left[f^{(h)}\right]\right\}, \tag{4}$$

$$g^{(w)}=\delta\left\{F_w\left[f^{(w)}\right]\right\}。 \tag{5}$$

最后,$g^{(h)}$ 和 $g^{(w)}$ 分别被扩展并用作注意力权重,得到整个坐标注意力模块的输出 y:

$$y_c(i,j)=x_c(i,j)\times g_c^{(h)}(i)\times g_c^{(w)}(j)。 \tag{6}$$

通过以上两个部分的操作即可实现使用精确的位置信息对通道之间的关系和远程依赖编码。此模块既是一种通道注意力,同时也包含了两个方向上的位置信息。

CA-ASPP 相比原来的 ASPP,能够重点关注感兴趣的目标和区域,更高效地利用多尺度提取到的特征,解决了 ASPP 性能饱和的问题,有利于兼顾不同尺寸的分割目标。

2.4　双边引导融合模块

在浅层特征与深层特征的融合方法上,逐元素相加与通道维度上的合并是融合多种类型特征图的常用方式。然而深层特征比较抽象,包含大量语义信息,浅层特征分辨率较高但缺少抽象的语义信息。以上两种融合方式将深层和浅层特征直接融合,没有在二者之间建立联系,忽略了两组特征的差异,导致融合效率有限[19]。相比之下,通过深层特征的语义信息来指导融合过程对浅层空间信息的利用效率会更高[15]。受到 BiSeNet V2[19]的启发,在解码器中引入了 BGA 模块来融合 CA-ASPP 的深层特征和浅层信息提取分支的浅层特征,如图1(a)所示。BGA 模块取代了通道维度上的简单合并,在深浅层特征之间建立通信,利用深层语义信息和上下文信息指导浅层进行特征融合,其结构如图2所示,其中,DWConv 表示深度卷积。

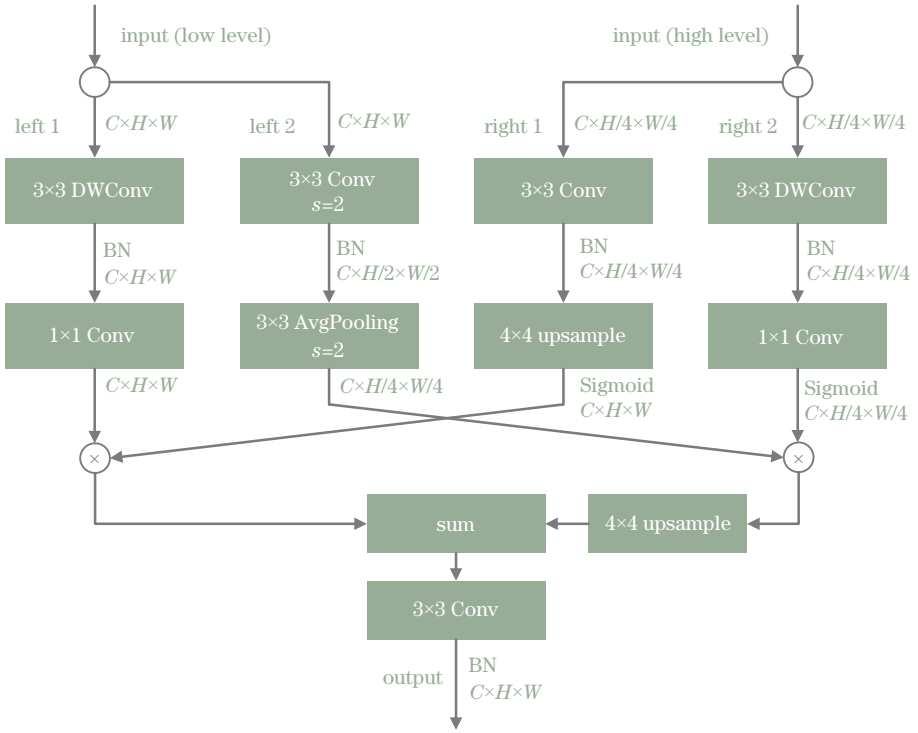

图2　BGA模块

Fig. 2　BGA module

BGA模块包含两个输入：一个是以浅层信息提取分支的输出特征图作为浅层输入；另一个是以CA-ASPP输出的特征图作为深层输入。每个输入分别经过两条分支处理，模块结构及分支编号如图2所示，左侧分支输入的是浅层信息提取分支的输出，其尺寸为原图的1/4。左侧left 1分支由一个深度可分离卷积层和一个1×1普通卷积层构成，用于特征提取，不改变特征图尺寸和通道数量；left 2分支分别通过一个卷积层和一个最大池化层进行两次步长为2的下采样，通道数不变但是尺寸缩小为原来的1/4，与CA-ASPP输出的深层特征图尺寸相等。右侧right 2分支与left 1分支一样进行特征提取但不改变特征图尺寸和通道数；right 1分支通过一个3×3卷积层和一个4倍上采样层得到一个输出与left 1分支大小相同的特征图。得到的4张特征图按浅层与深层两两组合的方式分别相乘，再进行逐元素相加和卷积操作即可得到高效融合的特征图。

BGA模块中的多张特征图的交叉融合操作通过不同尺度之间的相互引导获得了多尺度特征，而且深层与浅层的两两组合实现了通过深层语义信息引导浅层空间特征融合的作用。相比简单的融合方式，该方式实现了深层与浅层的高效通信，充分融合浅层信息提取分支得到的空间细节信息，有利于提高小目标的分割效果。

3 实验与分析

3.1 实验设置

实验全部在PyTorch框架下使用NVIDIA RTX3090 GPU完成。实验中将batch size设为16，整个训练过程历经400个epoch，输入图像的尺寸设定为320 pixel×320 pixel。主干网络使用Xception在ImageNet上的预训练权重，并将adam算法作为优化器，其平滑常数β_1和β_2分别设为0.9和0.999，初始学习率设为5×10^{-4}，其最小值限定为5×10^{-6}。

在训练采用的损失函数方面，交叉熵函数(CE)适用于大多数语义分割场景，但是它为所有样本分配了相同的重要性，在类不平衡数据集上训练不稳定，模型的学习会偏向像素较多的类[23]。而目标尺寸差异大也就是数据不平衡的情况，Dice Loss(DL)[24]相较于CE有更优的表现。DL的计算方法为

$$L_{\mathrm{D}}=1-D=1-\frac{2\left(P_{\mathrm{pd}}\cap P_{\mathrm{gt}}\right)}{P_{\mathrm{pd}}\cup P_{\mathrm{gt}}},\tag{7}$$

式中：D为Dice系数；P_{pd}为预测结果；P_{gt}为标签。以CE为主损失函数，将DL引入训练过程来优化在数据不平衡场景下的训练效果，并为其分配较小的权重以防止其对训练的稳定性产生影响，构成复合损失函数为

$$L=L_{\mathrm{CE}}+\alpha L_{\mathrm{DL}}。\tag{8}$$

实验证明，当系数α=0.6时，DL可以在对反向传播造成不利影响和对数据不平衡情况的作用之间达到平衡。

3.2 数据集与评价指标

实验在AeroScapes[25]和Semantic Drone[26]这两个UAV低空航摄语义分割数据集上完成。AeroScapes是由卡内基梅隆大学的Ishan Nigam团队制作的一个2D语义分割数据集，由UAV在距离地面垂直高度为5～50 m内采集得到，共3269张低空场景图片和同样数量的密集标注的标签文件。标签包含了11个目标种类：人、自行车、汽车、UAV、船、动物、障碍物、建筑物、植物、路和天空。Semantic Drone数据集是2D城市场景语义分割数据集，由UAV在距离地面垂直高度为5～30 m内拍摄得到，包含400张6000 pixel×4000 pixel的图片及标签。UAV可以在三维空间中自由飞行，因此这两个数据集包含了更丰富和更多样化的视角和视觉尺度。而且这两个数据集的类分布具有典型的数据不平衡的特点，比如在AeroScapes中，人、自行车、汽车、UAV、船、动物和障碍物这7个种类总共只占了所有像素分布的1.51%[25]。

将平均交并比(mIoU)和平均像素准确率(mPA)作为评价指标。二者都是由混淆矩阵计算得到的：mIoU为所有类别预测结果和真实标签的交并比的平均值；mPA计算的是所有类别中被正确分类的像素点的平均值。

$$M_{\mathrm{mIoU}}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}\left(p_{ij}+p_{ji}\right)-p_{ii}},\tag{9}$$

$$M_{\mathrm{mPA}}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}},\tag{10}$$

式中：k为数据集的分类数；$k+1$为添加了背景的类数；p_{ij}为将i类别预测为j类别的像素点；p_{ii}和p_{ji}同理，因此p_{ii}为预测正确的像素点。

3.3 在AeroScapes数据集上的实验

将AeroScapes数据集按80%、10%和10%的比例划分出训练集、验证集和测试集，并在测试集上进行mIoU和mPA的计算。在此数据集上对所提方法与已有方法进行了对比实验，结果如表1所示。在几个经典语义分割算法中，U-Net的表现相对较好，其结构中包含了连接编码器和解码器的跳跃连接结构，能够从编码器中保留一部分空间信息，达到了75.84%的mIoU和83.31%的mPA。DeepLabV3+由于其多尺度提取模块和重新设计的解码器结构，也达到了77.49%的mIoU和85.03%的mPA。模型DADA[27]和DSRL[28]在AeroScapes数据集上的性能相比前几个模型有较大提高，如表1所示。所提方法的mIoU和mPA分别达到了83.16%和90.75%，说明所提CA-ASPP、浅层特征提取分支和双边引导聚合模块对于提高网络模型在UAV航摄场景下的分割表现是有效的。

表 1 不同模型在 AeroScapes 数据集上的评估结果对比

Table 1 Comparison of evaluation results of different models on AeroScapes dataset

Method	Backbone	mIoU /%	mPA /%
FCN[3]	VGG-16[29]	67.59	74.53
U-Net[4]	ResNet-50[20]	75.84	83.31
PSPNet[7]	MobileNetV3[30]	58.15	63.86
PSPNet[7]	ResNet-50[20]	60.57	66.72
RefineNet[8,31]	ResNet-101[20]	63.09	70.82
DeepLabV3+[10]	MobileNetV3[30]	78.01	84.3
DeepLabV3+[10]	Xception[32]	77.49	85.03
DADA[27,31]	DeepLabV2[33]	81.53	88.75
DSRL[28,31]	ResNet-101[20]	82.48	89.72
Proposed	Xception[32]	**83.16**	**90.75**

图 3 是部分网络模型的预测结果对比，同时给出了图片在数据集中的编号。可以看出，所提方法的分割结果中小目标没有被忽略且具有更多细节，比如图 3(b)中的大熊猫、图 3(c)中的 UAV 和图 3(d)中的人；相邻的小目标没有被错分成同一类或相邻的其他类别，比如图 3(e)中骑自行车的人；大目标分割完整，且距离较近的同类目标之间与内部有他类目标的情况中很少出现粘连问题，比如图 3(a)中的植物和图 3(b)中的天窗。

在 AeroScapes 数据集上针对所提方法进行了消融实验，以验证各个改进的有效性及其相互关系。在构建 CA-ASPP 模块时，对不同位置使用坐标注意力模块的效果进行了实验，如表 2 所示，在不同改进点之间的消融实验结果如表 3 所示。

表 2 在 CA-ASPP 模块不同位置使用坐标注意力模块的评估结果对比

Table 2 Comparison of evaluation results using coordinate attention block at different locations of CA-ASPP

Location	mIoU /%	mPA /%
4	79.35	87.29
5	79.41	87.53
6	79.55	87.67
4,6	**79.71**	**87.90**
1,2,3,4	78.86	87.68
1,2,3,4,6	79.62	87.88

在表 2 中，数据全为仅在主干网络为 Xception 的 DeepLabV3+ 上改进了 ASPP 模块的结果。表 2 中 Location 一栏的序号 1～6 分别表示在 3 个空洞率为 6、12、18 的卷积层、全局池化层、通道合并层和最后的 1×1 卷积层后使用注意力模块。数据表明，当在全局池化层和最后的 1×1 卷积层之后各使用一个坐标注意力模

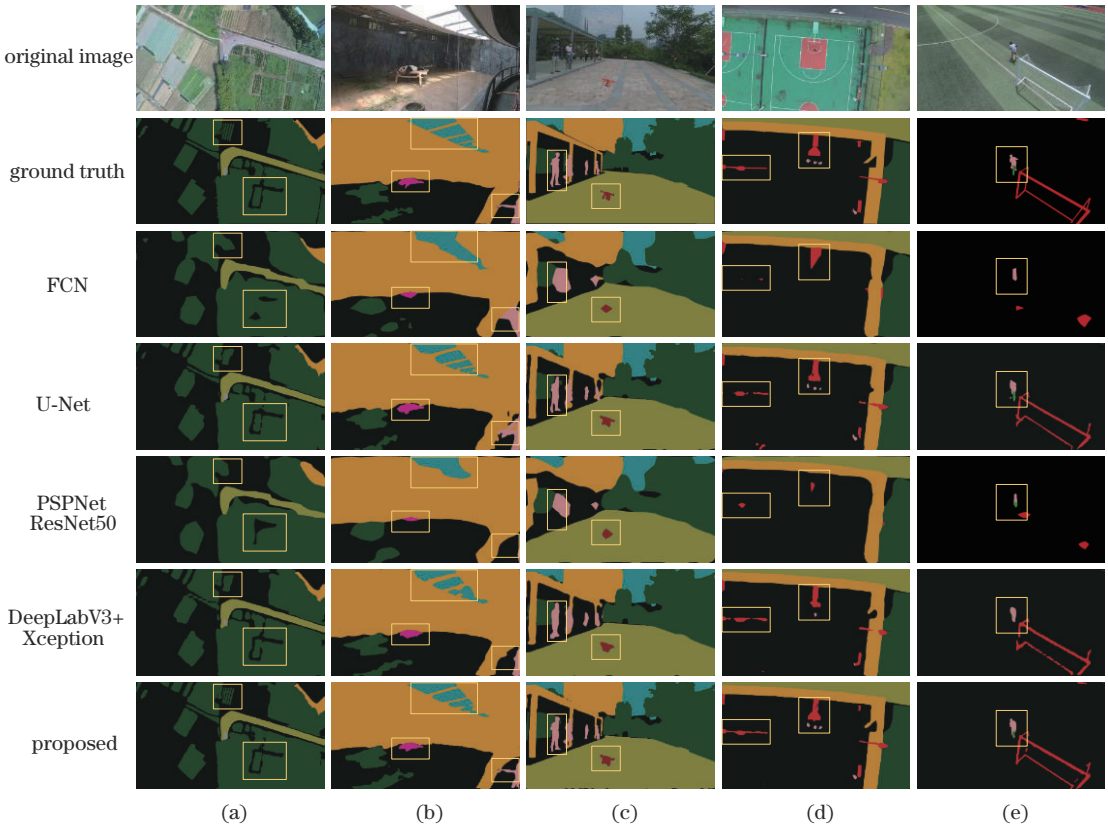

图 3 不同模型在 AeroScapes 数据集上的分割结果对比。(a)002001_049 号图片；(b)038032_032 号图片；(c)045002_049 号图片；(d)310019_016 号图片；(e)311000_004 号图片

Fig. 3 Comparison of prediction maps of different models on AeroScapes dataset. (a) Picture 002001_049; (b) picture 038032_032; (c) picture 045002_049; (d) picture 310019_016; (e) picture 311000_004

表 3　不同改进点的消融实验评估结果

Table 3　Evaluation results of ablation experiments with different improved methods

Method	mIoU /%	mPA /%
-	77.49	85.03
CA-ASPP	79.71	87.90
ConvBranch	78.96	86.84
BGAModule	79.45	87.38
ConvBranch, BGAModule	80.85	88.04
CA-ASPP + ConvBranch + BGAModule	81.53	88.96
ConvBranch + BGAModule + Multi-loss	82.84	90.31
CA-ASPP + ConvBranch + BGAModule + Multi-loss	**83.16**	**90.75**

块时效果最好，其 mIoU 和 mPA 分别达到了 79.71% 和 87.90%。因此在其他实验中均采用这个方案的 CA-ASPP。

在表 3 中，Method 一栏列举了当次实验所使用的改进点，其中：ConvBranch 为浅层信息提取分支；Multi-loss 为采用复合损失进行训练，其中参数 α=0.6。由表 3 中数据可以看出，仅使用 CA-ASPP 作为改进点时，网络性能有了显著提升；仅使用 ConvBranch 时，提升效果并不明显，且低于仅使用 BGA 模块时的效果；而当同时使用 ConvBranch 和 BGA 模块时，mIoU 和 mPA 分别达到了 80.85% 和 88.04%，相较于改进前的 77.49% 和 85.03% 有较大提升，分别提升了 3.36 个百分点和 3.01 个百分点。说明不仅要获取更多浅层信息，也要相应保证深浅层特征高效融合才能有效提高浅层信息的利用率。此外，Multi-loss 对网络性能的提升也有明显效果。在所有改进点共同作用时，网络性能最终达到了 83.16% 的 mIoU 和 90.75% 的 mPA。

在 AeroScapes 数据集上的对比试验和消融实验的结果表明，所提分割方法有利于网络在分割时兼顾不同尺寸的目标，提高网络在数据不平衡场景下的性能，更加适用于 UAV 航摄图像的语义分割场景。

3.4　在 Semantic Drone 数据集上的实验

为了更泛化地评估所提方法在 UAV 航摄图像语义分割场景下的表现，在 Semantic Drone 数据集上进行了部分模型的对比实验。相比 AeroScapes 数据集，Semantic Drone 数据集中图像的分辨率更高，在有的图像中目标尺寸差异更大且场景更复杂、小目标更多。实验结果如表 4 所示，分割结果如图 4 所示，同时给出图片在数据集中的编号。

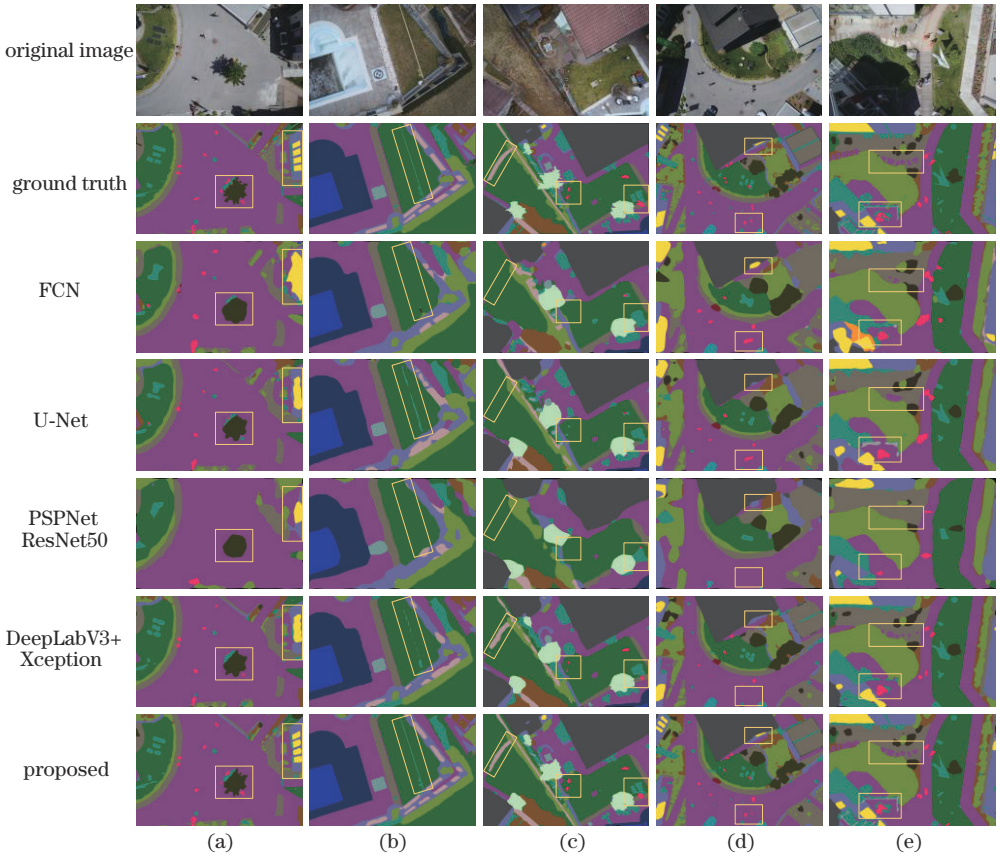

图 4　不同模型在 Semantic Drone 数据集上的分割结果对比。(a) 002 号图片；(b) 056 号图片；(c) 119 号图片；(d) 311 号图片；(e) 412 号图片

Fig. 4　Comparison of prediction maps of different models on Semantic Drone dataset. (a) Picture 002; (b) picture 056; (c) picture 119; (d) picture 311; (e) picture 412

表 4 不同模型在 Semantic Drone 数据集上的评估结果对比

Table 4 Comparison of evaluation results of different models on Semantic Drone dataset

Method	Backbone	mIoU /%	mPA /%
FCN[3]	VGG-16[29]	54.61	63.63
U-Net[4]	ResNet-50[20]	57.38	68.45
PSPNet[7]	MobileNetV3[30]	45.43	54.08
PSPNet[7]	ResNet-50[20]	42.81	51.55
DeepLabV3+[10]	MobileNetV3[30]	55.31	64.56
DeepLabV3+[10]	Xception[32]	55.48	64.00
Proposed	Xception[32]	**72.09**	**80.34**

因为 Semantic Drone 数据集中图片分辨率更高、场景更复杂、尺寸差异更大等因素，所有网络模型在这个数据集上的表现相较于 AeroScapes 数据集都有一定差距。但是表 4 中的数据表明，所提方法依然优于几个主流语义分割方法，mIoU 和 mPA 分别达到了 72.09% 和 80.34%。

由图 4 可以看出，相较于对比的几种模型，所提方法依然有一定优势：大目标分割完整且不会导致其内部的错分，比如图 4(b)中的细线和图 4(c)中左侧大面积草坪上的围栏；在大尺寸目标内部的狭窄形状目标中很少产生错误分割的情况，比如图 4(b)中的细线；小目标得到了关注，几乎没有小目标被忽略，比如图 4(a)、(c)、(e)中的人；距离较近的小目标很少出现分割粘连的情况，比如图 4(a)中的窗口和图 4(e)中的石板路。由 Semantic Drone 数据集的实验结果可以看出，所提方法在 UAV 航摄图像的语义分割中具有较强的泛化能力。

4 结 论

针对 UAV 航摄图像场景中目标尺寸差异大导致的感受野难以同时兼顾不同尺寸目标分割效果的问题，提出了双路特征融合网络 DSFA-Net。其中：浅层特征提取分支通过生成比主干网络初期通道数更多的特征图以保留更多浅层特征，弥补了主干网络在多次下采样过程中对浅层信息保留不足的问题；BGA 模块以多尺度引导的方式进行深层和浅层特征的融合，相比通道维度的简单合并，提高了深层语义信息和浅层空间信息的融合效率，在分割结果中还原了更多细节；针对 ASPP 多尺度提取层的堆积导致的性能饱和、信息过载等问题引入坐标注意力机制，所实现的 CA-ASPP 结构在多尺度提取时可更加关注目标区域，提升了网络对不同尺寸目标的分割能力。实验证明，所提方法提高了网络模型兼顾不同尺寸分割目标的能力，在 UAV 航摄图像的语义分割任务中获得了更高精度的分割结果。所提方法是基于卷积神经网络的，与之相比，Transformer 及其自注意力机制可以建立全局范围的依赖关系以获得更大的感受野，这对于 UAV 航摄场景下的语义分割任务是有利的。因此下一步将重点关注 Transformer 及其自注意力机制在 UAV 航摄图像的语义分割方面的研究和应用。

参 考 文 献

[1] 徐兆忠，彭力，戴菲菲. 多尺度特征对齐聚合的语义分割方法[J]. 激光与光电子学进展，2023，60(2)：0215004.
Xu Z Z，Peng L，Dai F F. Semantic segmentation method based on multiscale feature alignment and aggregation[J]. Laser & Optoelectronics Progress，2023，60(2)：0215004.

[2] 唐璐，万良，王婷婷，等. DECANet：基于改进 DeepLabV3+的图像语义分割方法[J]. 激光与光电子学进展，2023，60(4)：0410002.
Tang L，Wan L，Wang T T，et al. DECANet：image semantic segmentation method based on improved DeepLabV3+ [J]. Laser & Optoelectronics Progress，2023，60(4)：0410002.

[3] Long J，Shelhamer E，Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，June 7-12，2015，Boston，MA，USA. New York：IEEE Press，2015：3431-344.

[4] Ronneberger O，Fischer P，Brox T. U-net：convolutional networks for biomedical image segmentation[M]//Navab N，Hornegger J，Wells W M，et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham：Springer，2015，9351：234-241.

[5] 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.

[6] 邱云飞，温金燕. 基于 DeepLabV3+与注意力机制相结合的图像语义分割[J]. 激光与光电子学进展，2022，59(4)：0410008.
Qiu Y F，Wen J Y. Image semantic segmentation based on combination of DeepLabV3+ and attention mechanism[J]. Laser & Optoelectronics Progress，2022，59(4)：0410008.

[7] 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 Press，2017：6230-6239.

[8] Lin G S，Milan A，Shen C H，et al. RefineNet：multi-path refinement networks for high-resolution semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，July 21-26，2017，Honolulu，HI，USA. New York：IEEE Press，2017：5168-5177.

[9] Kirillov A，Girshick R，He K M，et al. Panoptic feature pyramid networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 15-20，2019，Long Beach，CA，USA. New York：IEEE Press，2020：6392-6401.

[10] 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.

[11] Cheng B W, Collins M D, Zhu Y K, et al. Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 12472-12482.

[12] Yang M K, Yu K, Zhang C, et al. DenseASPP for semantic segmentation in street scenes[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 3684-3692.

[13] Yang K L, Zhang J M, Reiß S, et al. Capturing omni-range context for omnidirectional segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 1376-1386.

[14] Huynh C, Tran A T, Luu K, et al. Progressive semantic segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 16750-16759.

[15] Zhang Z, Zhang X, Peng C, et al. ExFuse: enhancing feature fusion for semantic segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11214: 273-288.

[16] Liu Z, Mao H Z, Wu C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 11966-11976.

[17] 徐聪, 王丽. 基于改进DeepLabv3+网络的图像语义分割方法[J]. 激光与光电子学进展, 2021, 58(16): 1610008.
Xu C, Wang L. Image semantic segmentation method based on improved DeepLabv3+ network[J]. Laser & Optoelectronics Progress, 2021, 58(16): 1610008.

[18] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13708-13717.

[19] Yu C Q, Gao C X, Wang J B, 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.

[20] He K M, Zhang X Y, Ren S Q, 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 Press, 2016: 770-778.

[21] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4510-4520.

[22] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2023-03-16]. https://arxiv.org/abs/1706.05587.

[23] Kervadec H, Bouchtiba J, Desrosiers C, et al. Boundary loss for highly unbalanced segmentation[EB/OL]. (2018-12-17)[2023-03-16]. https://arxiv.org/abs/1812.07032.

[24] Milletari F, Navab N, Ahmadi S A. V-net: fully convolutional neural networks for volumetric medical image segmentation[C]//2016 Fourth International Conference on 3D Vision (3DV), October 25-28, 2016, Stanford, CA, USA. New York: IEEE Press, 2016: 565-571.

[25] Nigam I, Huang C, Ramanan D. Ensemble knowledge transfer for semantic segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV), March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 1499-1508.

[26] Mostegel C, Maurer M, Heran N, et al. Semantic drone dataset[EB/OL]. (2019-01-25) [2023-03-16]. https://www.tugraz.at/index.php?id=22387.

[27] Vu T H, Jain H, Bucher M, et al. DADA: depth-aware domain adaptation in semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 7363-7372.

[28] Wang L, Li D, Zhu Y S, et al. Dual super-resolution learning for semantic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 3773-3782.

[29] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, May 7-9, 2015, San Diego, USA. [S.l.: s.n.], 2015.

[30] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 1314-1324.

[31] Zhang H, Hui J X. A new semantic segmentation network with FPFN and dense ASPP[C]//2021 International Conference on Control, Automation and Information Sciences (ICCAIS), October 14-17, 2021, Xi'an, China. New York: IEEE Press, 2021: 799-804.

[32] 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 Press, 2017: 1800-1807.

[33] 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.