DOI:10.19651/j.cnki.emt.2417529

一种基于 VDSEC-UNet 的遥感影像建筑物提取方法*

张剑飞 王友为
(黑龙江科技大学计算机与信息工程学院 哈尔滨 150022)

摘 要:近年来卷积神经网络在遥感影像建筑物提取研究中取得了极大的成功,但其仍然面临着整体提取精度不高、错分、漏分和边界模糊等问题。针对以上问题,提出一种基于 VDSEC-UNet 的遥感影像建筑物提取方法。首先,使用 VGG-16 作为编码器,以提取建筑物特征信息;其次,使用动态上采样代替传统上采样,增强模型对细节的感知能力,从而提升建筑物边界的提取精度;接着,在编解码器中间嵌入一个多尺度上下文信息提取模块,以充分考虑建筑物周围其他对象影响,引入足够的上下文信息及不同感受野下的全局信息,减少空间信息损失,提升对不同尺度建筑物的提取效果;然后,在每个跳跃连接部分嵌入 ECA 注意力机制,提高模型对影像中建筑物特征的关注度;同时,使用联合损失函数缓解类别不平衡问题;最后,构造 CA-DPGHead 模块并加在解码器末尾,以增强建筑物与背景之间的区分,使模型更加精准地定位和识别图像中的建筑物信息,进而提升对小型建筑物的提取精度并细化建筑物边界的提取效果。实验结果表明,VDSEC-UNet 在 Massachusetts 和 Inria 数据集上的 mIoU 分别达到了 82.07%和 84.35%,F1 指数分别达到了 83.34%和 86.66%,优于其他经典方法。

关键词:建筑物提取;VDSEC-UNet;多尺度上下文;注意力机制;CA-DPGHead
中图分类号:TP391;TN911.73 **文献标识码**:A **国家标准学科分类代码**:520.20

A method of building extraction from remote sensing images based on VDSEC-UNet

Zhang Jianfei Wang Youwei
(School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China)

Abstract: In recent years, convolutional neural networks have achieved great success in the study of building extraction from remote sensing images, but they still face problems such as low overall extraction accuracy, misclassification, omission, and fuzzy boundaries. Aiming at the above problems, a building extraction method based on VDSEC-UNet for remote sensing images is proposed. Firstly, VGG-16 is used as the encoder to extract the building feature information. Secondly, dynamic up-sampling is used instead of traditional up-sampling to enhance the model′s ability to perceive the details so as to improve the extraction accuracy of the building boundaries. Next, a multi-scale context information extraction module is embedded in the middle of the coder and decoder in order to take the influence of other objects around the building into account and introduce sufficient context information and global information under different sensing fields to reduce the loss of spatial information and enhance the extraction effect of buildings at different scales. Then, the ECA attention mechanism is embedded in each jump connection part to improve the model′s attention to the building features in the image. At the same time, the joint loss function is used to alleviate the category imbalance problem. Finally, the CA-DPGHead module is constructed and added at the end of the decoder to enhance the distinction between buildings and background so that the model can locate and identify the building information in the image more accurately, which in turn improves the extraction accuracy of small buildings and refines the extraction effect of building boundaries. The experimental results show that the mIoU of VDSEC-UNet on Massachusetts and Inria datasets reaches 82.07% and 84.35%, respectively, and the F1 index reaches 83.34% and 86.66%, respectively, which is better than other classical methods.

Keywords: building extraction; VDSEC-UNet; multi-scale context; attention mechanism; CA-DPGHead

0 引 言

建筑物作为重要的人造地物,是城市化发展进程的重要标志。高效、自动且准确地从高分辨率遥感影像中提取建筑物信息,在建筑物变化检测[1]、灾难应急[2]和地图制图等领域具有重要作用,目前已成为研究的热点与难点。

收稿日期:2024-12-03
*基金项目:国家自然科学基金(61803148)、黑龙江省哲学社会科学研究规划项目(23YSD245)、黑龙江省属高等学校基本科研业务费项目(2024-KYYWF-1099)资助

早期的建筑物提取手段主要依赖于遥感图像技术和空间数据处理技术[3],包括面向对象的方法[4]、融合辅助信息的方法[5]等。虽然这些方法能够提取一些建筑物信息,但是往往需要大量的人工参与。随着技术的进步,出现了提取效果更好的机器学习算法,例如支持向量机和基于马尔科夫随机场模型等。尽管这些机器学习的方法带来了诸多优势,但其对影像中蕴含的丰富纹理和细节信息的挖掘能力有限,往往会导致大量信息的丢失,造成错分、漏分和提取不充分等问题。此外,这些技术的自动化水平不高,难以及时且高效地更新建筑物信息,从而无法满足在规定时间内完成信息更新的需求。

随着深度学习的快速发展,语义分割开始作为遥感影像分割领域的研究重点,越来越多的研究者对高分辨率遥感建筑物的语义分割进行了广泛的研究,并提出了大量的遥感图像语义分割方法。闫祎巧等[6]设计了一种适应多尺度变化的新型网络,在边缘、细节纹理和客观指标数据上均达到了更优的效果,具有更准确的图像分割精度,但在运行效率和网络泛化能力等方面仍存在提升空间。任远锐等[7]提出了一种增强注意力门控的 U 型网络(AA_UNet),不仅提高了提取精度,还改善了建筑物轮廓不清晰和小目标建筑物破碎的问题。Chen 等[8]提出了一种上下文特征增强网络(CFENet),能够有效地增强和融合建筑的低层和高层特征,从而提高建筑提取精度,但对困难样本的判别能力有待进一步提高。Ye 等[9]提出 FMAM-Net 网络,其使用 ResNet-34 作为编码器,解决深度网络训练时梯度消失的问题,在每个跳跃连接处嵌入特征细化补偿模块和串联注意力模块,提取深层建筑物信息,但感受野小,导致对小目标提取能力较差。Qiu 等[10]在 U-Net 模型中引入空洞空间金字塔池化(atrous spatial pyramid pooling,ASPP)模块,以捕获并融合多尺度特征,并增强信息之间的关联性,进而提高模型对不同尺度建筑物的感知能力和提取能力,但在建筑物提取的质量和准确性方面仍有待进一步提高,尤其是在边界区域和不规则建筑物的形状。Wang 等[11]提出了一种残差 U-Net 方法,该方法结合了 U-Net、残差结构和空间金字塔等方法,提高了建筑物的提取精度,但存在边界精度不高和相邻建筑物识别效果差等方面的问题。Yang 等[12]提出了一种用于建筑提取的复杂场景自适应网络(CSANet),其由分层上下文特征提取模块、全局局部特征交互模块和多尺度自适应特征融合模块组成,可以提高复杂场景中建筑物提取的准确性,但其在建筑边界的准确性方面仍有不足。

以上方法虽然在一定程度上可实现对建筑物的提取。但是由于遥感影像建筑物类间差异大、尺度不一和复杂性高等特点,仍然面临着整体提取精度不高、错分、漏分和边界模糊等问题。为此,提出一种基于 VDSEC-UNet 的遥感影像建筑物提取方法。本文主要工作:

1)使用 VGG-16 作为编码器,充分提取建筑物特征信息。

2)引入动态上采样代替传统上采样,增强模型对细节的感知能力,从而提升建筑物边界的提取精度。

3)设计多尺度上下文信息提取模块(multi-scale contextual information extraction module,MCIEM)并嵌入在编解码器中间,以充分考虑建筑物周围其他对象影响,引入足够的上下文信息及不同感受野下的全局信息,减少空间信息损失,提升对不同尺度建筑物目标的提取效果。

4)在每个跳跃连接部分嵌入 ECA(efficient channel attention,ECA)注意力机制,以抑制遥感影像中复杂背景的干扰,提高模型对影像中建筑物特征的关注度。

5)使用联合损失函数在训练过程中进行联合优化,以缓解遥感图像中存在的类别不平衡问题。

6)构造 CA-DPGHead(coordinate attention-dynamic prototype guided head,CA-DPGHead)模块并加在解码器末尾,以增强建筑物与背景之间的区分,使模型更加精准地定位和识别图像中的建筑物信息,进而提升对小型建筑物的提取精度并细化建筑物边界的提取效果。

1 研究方法

1.1 VDSEC-UNet 模型结构

VDSEC-UNet 模型结构如图 1 所示,该模型主要由编码器、MCIEM 和解码器组成。

编码器采用 VGG-16 的前 5 个块对图像特征进行提取,实现对建筑物和背景的区分。每个块内包含多层卷积操作,这些卷积层的首层输出通道数依次为 64、128、256、512 和 512(且同一块内各层通道数相同),每个卷积层的卷积步长为 1,填充为 1。每个块后有一个尺寸为 2 的最大池化层,用于逐步缩减特征图的空间维度。其工作原理是:首先,输入图像经过编码器编码后会得到 4 个由浅层到深层的不同尺度的特征信息;然后,将每个特征信息通过一个 ECA 注意力机制,以减少遥感影像中复杂背景的干扰,提高模型对影像中建筑物特征的关注度;最后,通过“跳跃”连接与解码器模块中相同尺度的特征图像在通道方向上进行“融合”,弥补了图像在下采样过程中丢失的建筑物信息。

在编解码器中间嵌入一个多尺度上下文信息提取模块(MCIEM),以充分考虑建筑物周围其他对象影响,引入足够的上下文信息及不同感受野下的全局信息,减少空间信息损失,提升对不同尺度建筑物目标的提取效果。

解码器主要由上采样模块、卷积层和 CA-DPGHead 模块构成,其中上采样模块采用动态上采样,其通过学习偏移量动态调整上采样的每个点,可以提高模型对细节的感知能力,进而提高模型对建筑物边缘的提取精度。将经过 MCIEM 模块处理后的特征传入解码器,经解码器处理后,得到与输入图像等同大小的特征图(512×512)。最后,使用模型的最后一层 CA-DPGHead 模块得到最终的提取结果。

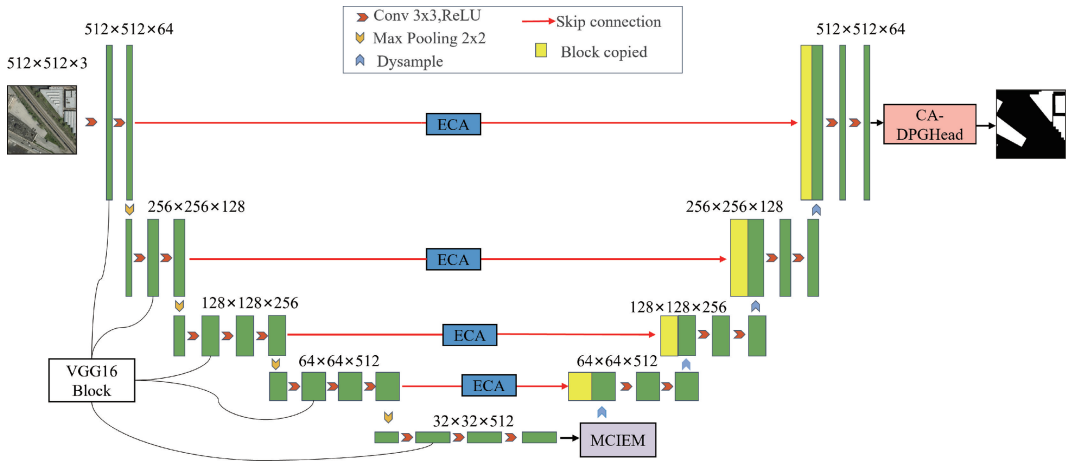

图 1　VDSEC-UNet 结构

Fig. 1　The architectures of VDSEC-UNet

1.2　引入动态上采样

传统上采样方式，例如转置卷积方法易产生棋盘格伪影问题，降低特征图的质量；而最近邻插值方法，仅利用少量周围像素进行预测，会忽略像素之间的平滑过渡，且只关注空间信息而忽略特征图的语义信息。因此，引入 DySample 方法代替传统上采样方法，DySample 通过学习偏移量动态调整上采样的每个点，可以提高模型对细节的感知能力，进而提高模型对建筑物边缘的提取精度。

DySample[13] 是一种动态上采样方法，通过动态调整采样点实现高效且高质量的上采样。该方法专注于点采样，而不是基于卷积核的上采样，这使其可以有效地保持目标的几何信息。相较于其他动态上采样器，该方法显著减少了参数量和 GPU 内存占用，同时在全景分割和单目深度估计等密集预测任务上优于其他上采样器，其工作流程如图 2 所示。

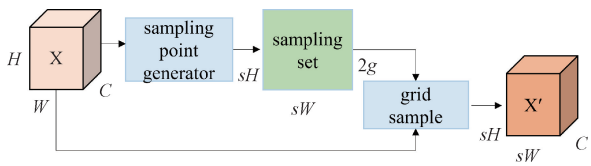

图 2　DySample 上采样流程

Fig. 2　DySample upsampling process

给定一个特征图 $X \in R^{C\times H\times W}$ 和一个 $2g\times sH\times sW$ 的点采样集 S，其中 C、H 和 W 分别是特征图的通道数、高和宽，$2g$ 表示 x 和 y 坐标，grid_sample 函数使用点采样集 S 中的位置对 X 重新采样，生成大小为 $C\times sH\times sW$ 的特征图 X'，如式(1)所示。

$$X' = Gridsample(X,S) \tag{1}$$

其中，点采样集 S 的工作流程如图 3 所示，点采样集 S 采用“线性＋像素重组”的方式生成，偏移范围由静态和动态范围因子决定。

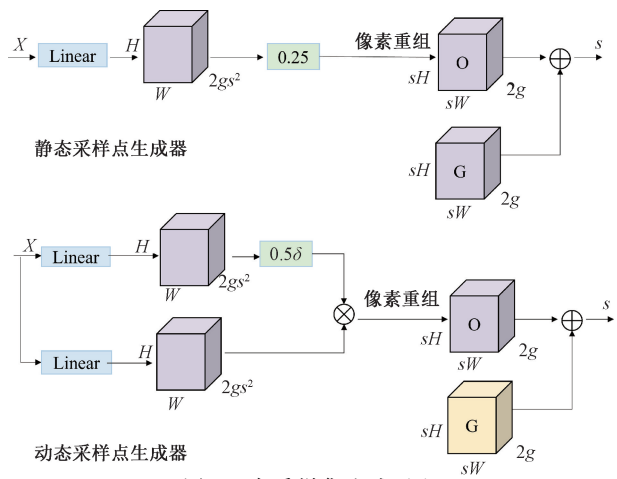

图 3　点采样集生成过程

Fig. 3　Point sampling set generation process

1.3　多尺度上下文信息提取模块设计

遥感影像中目标众多，建筑物区域与水域、植被等背景元素之间存在一定程度的粘连，容易导致误分割现象。为了充分考虑建筑物周围其他对象影响，引入足够的上下文信息及不同感受野下的全局信息，减少空间信息损失，提升对不同尺度建筑物目标的提取效果，在编解码器中间嵌入一个多尺度上下文信息提取模块(MCIEM)。

该模块受空洞空间金字塔池化(ASPP)[14] 启发设计，由 6 个分支组成，具体包含一个平均池化分支，用于获得特征图的全局特征，五个膨胀率分别为 3、6、12、18 和 24 的空洞卷积分支，用于扩大特征图感受野，综合提取多尺度建筑物信息。其中，膨胀率为 3 的空洞卷积分支，可以更精准地捕获图像中的小尺度目标，而膨胀率为 24 的空洞卷积分支则具备更大的感受野，有助于更好地捕获大尺度目标，其结构如图 4 所示。

1.4　引入 ECA 注意力机制

为抑制遥感影像中复杂背景的干扰，同时提高模型对

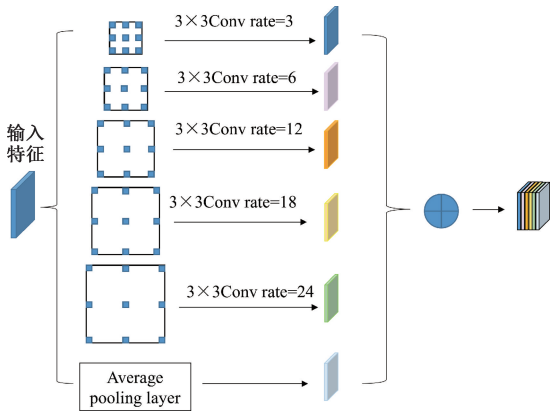

图 4 MCIEM 模块结构

Fig. 4 Structure of MCIEM module

影像中建筑物特征的关注度,在 VDSEC-UNet 模型的每个跳跃连接部分嵌入 ECA 注意力机制。ECA 注意力机制[15]是在 SE-Net(squeeze-and-excitation network,SE-Net)[16]的基础上改进的新型轻量级通道注意力机制。

ECA 结构如图 5 所示,其中 C、H 和 W 分别是特征图的通道数、高度和宽度,GAP 为全局平均池化,σ 为 Sigmoid 激活函数,K 代表卷积内核的尺寸。

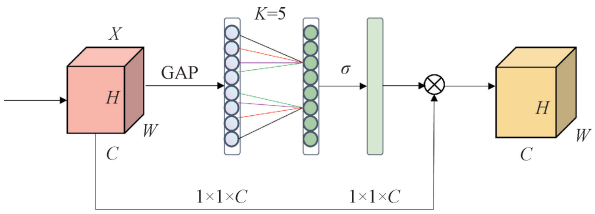

图 5 ECA 注意力模块结构

Fig. 5 Structure of ECA attention module

其工作流程为:首先,对输入的特征图进行全局平均池化,将其维度从[H,W,C]转换为[1,1,C];其次,按照实际特征图中通道的数量,确定 K 的大小;然后,利用 Sigmoid 激活函数处理一维卷积的输出,以得到每个通道的权重系数,其值在[0,1]范围内;最后,将归一化处理获得通道权重与原始特征进行加权乘积运算得到新的特征。

ECA 结构中的 K 与通道数 C 关系如式(2)所示,其中 $|*|_{odd}$ 表示取最近的奇数,γ 和 b 通常设置为 2 和 1。

$$K=\Phi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{\text{odd}} \tag{2}$$

1.5 联合损失函数

为解决遥感图像中存在类别不平衡问题,本文采用 CE Loss、Dice Loss 和 Focal Loss 在训练过程中进行联合优化。CE Loss 是一种常用的图像分割任务损失函数,通常用于衡量模型预测结果与真实标签之间的差异,并作为优化目标进行模型训练。但是在遥感图像中,存在类别不平衡问题,即背景远多于目标,如果只使用 CE Loss 会导致模型更倾向于背景类,会降低建筑物的提取精度。Dice Loss 通过计算预测结果和真实标签的交集与并集之比来衡量相似度,能够更关注前景目标,缓解样本不平衡带来的消极影响,但是在 Dice Loss 计算中,如果预测值和真实值相差很大,则梯度会变得非常小,使模型难以学习。Focal Loss 也是一种针对类别不平衡问题而设计的一种损失函数,通过调整样本的权重,使模型更加关注难以分类的样本,提高对少数类别的学习能力,同时引入了可调参数,用于调节难易样本的关注程度。所以将 3 种函数相结合,可以优势互补,更好地解决遥感图像中存在的类别不平衡问题。

1.6 CA-DPGHead 模块设计

由于建筑物目标所占像素小且容易受到背景因素影响,为了增强建筑物与背景之间的区分,使模型更加精准地定位和识别图像中的建筑物信息,构造了 CA-DPGHead 模块并加在解码器末尾,以提升模型对小型建筑物的提取精度并细化建筑物边缘的提取效果。CA-DPGHead 模块由坐标注意力机制(coordinate attention,CA)[17]和动态原型引导头(dynamic prototype guided head,DPG Head)[18]组成,其结构如图 6 所示。

CA 注意力机制可以进行局部特征增强,使网络更加精准地定位和识别图像中的建筑物信息。与传统注意力机制 CBAM(convolutional block attention module,CBAM)[19]等相比,CA 是针对通道注意力提出的一种新的注意力模块,其将位置信息融合到通道信息中,在捕获通道特征的同时进行方向和位置的信息捕捉。

DPG Head 通过将类嵌入特征投影到像素特征空间中,对像素特征进行加权,以增强不同类之间的区分,其结构如图 6 所示,工作原理如下:

1)首先,将特征投影到类特征空间中。然后,将类空间中的特征与像素空间中的特征相乘,得到动态原型。动态原型能反映每幅图像上不同类别的特征分布,并能跟随输入的动态变化而变化。生成动态原型的过程如式(3)所示:

$$F_p=\delta_{D\to C}(F_X)\otimes F_X \tag{3}$$

其中,$\otimes$表示矩阵乘法,F_p 表示原型,F_X 是输入特征,$\delta_{D\to C}$ 表示将 F_X 投影到类特征空间,通道数从 D 投影到 C。

2)为了进一步嵌入类信息以增强不同类之间的区分,将原型通过一个全连接层和 Softmax 激活函数,全连接层将原型压缩到 $C\times 1$ 维,以有效地嵌入类信息并得到一个类嵌入向量,Softmax 用于将取值限制在[0,1]。类嵌入向量 $C\times 1$ 表示每个类中的全局信息。类嵌入向量的计算过程如式(4)所示。

$$F_{gp}=\text{Softmax}(\delta_{D\to 1}(F_p)) \tag{4}$$

其中,δ 表示全连接层。

3)为了将类嵌入向量投影到像素特征空间中,将类嵌入向量与转置原型相乘。随着类别嵌入的增强,新的注意力向量具有更强的类别区分能力。然后,将类嵌入向量与

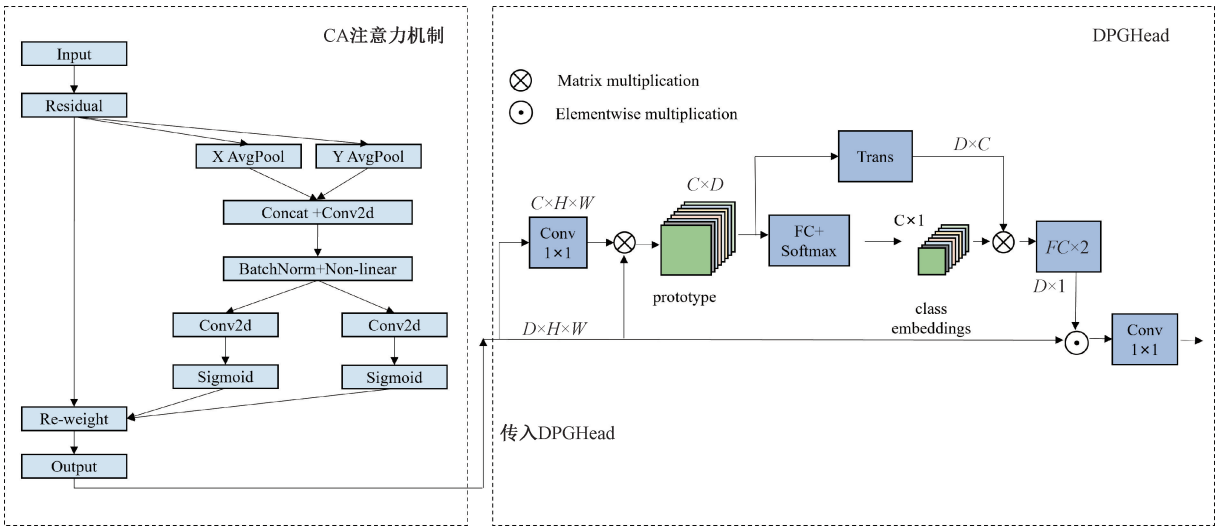

图 6　CA-DPGHead 模块结构

Fig. 6　Structure of CA-DPGHead module

转置原型相乘的结果通过两个全连接层，以捕获注意力向量中的上下文。接着，使用 LN 对两层之间的特征进行归一化处理。最终，将注意力向量对像素特征进行加权，以强调重要特征，增强特征表示。计算过程如式(5)所示。

$$F_O = \delta(\mathrm{ReLU}(\mathrm{LN}(\delta(F_p \otimes F_{gp})))) \odot F_X \tag{5}$$

其中，LN 和 ReLU 分别是层归一化和 ReLU 激活函数，δ 表示全连接层，$\odot$表示 Hadamard 积。

2　实验与分析

2.1　数据集介绍

本文选用 Massachusetts 数据集(massachusetts building dataset)和 Inria 数据集(inria aerial image labeling dataset)进行实验。

Massachusetts 数据集由 137 张训练、4 张验证和 10 张测试图像组成，每张图像的大小为 1 500×1 500 像素，提供了别墅、商业中心等不同色调和纹理或复杂形状的建筑。为了获得更多数据训练，按照 500 的步长，将原始图像切割成尺寸为 512×512 像素的小块。最终得到训练集、验证集、测试集的数量分别为 1 233、36 和 90 张。

Inria 数据集包含 Austin、Chicago、Kitsap、WesternTyrol 和 Vienna5 个城市的航空图像。由于设备资源有限，选取 Chicago 和 Vienna 两个地区的 72 张影像进行实验，选取每个地区的前 5 张作为测试集，第 12～16 张作为验证集，其余为训练集。再通过滑动窗口的形式裁剪为 512×512 的图像，滑动步长为 512。最终得到训练集、验证集、测试集的数量分别为：4 212、810、810 张。

2.2　实验设置

本文实验平台配置：Ubuntu20.04 操作系统、Intel® Xeon® Platinum 8352V 处理器、NVIDI RTX4090 显卡、Python3.8、Cuda11.3。深度学习框架为 Pytorch1.11.0，动量参数为 0.9，初始学习率为 0.007，通过余弦退火函数作为衰减策略动态更新学习率，Batchsize 为 16，网络在 Massachusetts 和 Inria 数据集上最大训练的迭代数分别为 150 和 100。

2.3　评价指标

为对所构建模型的性能进行定量描述，选用精准率(Precision)、召回率(Recall)、F1 指数(F1 score)和均交并比(mIoU)作为评价指标，其具体计算公式如式(6)～(9)所示。

$$mIOU = \frac{1}{class}\sum_{i=1}^{class}\frac{TP}{TP+FP+FN}(class=2) \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{9}$$

其中，TP 表示为正确预测出建筑物的数量，FP 表示将非建筑物预测为建筑物的数量，FN 表示将建筑物预测为非建筑物的数量。F1 是 Precision 和 Recall 的加权平均值，F1 越高，模型的性能越好。mIoU 是语义分割的标准度量，可以表示不同类别的预测值和真实值之间的重叠程度，其值越大代表网络分割精度越高。

2.4　模型训练结果

初始学习率和衰减策略对模型的收敛速度及建筑物提取性能有一定影响。在其他条件相同的情况下，采用 0.07、0.007 和 0.000 7 三种不同的学习率初始值并结合 cos 和 step 两种不同的衰减策略进行模型训练。

表 1 是基准模型的实验结果，由实验结果可以看出，基准模型在初始学习率为 0.007 且采用 cos 衰减策略的条件下取得最优结果，因此后续的改进工作均在此条件下展

开。此外,在初始学习率为 0.07 时,提取精度很差,故本文模型不再做该条件下的对比实验。表 2 是本文模型的实验结果,由实验结果可以看出,本文模型也在初始学习率为 0.007 且采用 cos 衰减策略的条件下取得最优结果。

表 1 基准模型在不同学习率和衰减策略下实验结果

Table 1 Experimental results of benchmark model under different learning rates and attenuation strategies

数据集	学习率	衰减策略	mIoU/%	F1/%
Massachusetts	0.07	cos	76.00	76.33
	0.007	**cos**	**80.77**	**82.06**
	0.000 7	cos	76.47	77.51
	0.007	step	79.76	80.95
Inria	0.07	cos	76.63	79.06
	0.007	**cos**	**82.77**	**85.20**
	0.000 7	cos	80.45	82.84
	0.007	step	82.13	84.52

表 2 VDSEC-UNet 模型在不同学习率和衰减策略下实验结果

Table 2 Experimental results of VDSEC-UNet model under different learning rates and attenuation strategies

数据集	学习率	衰减策略	mIoU/%	F1/%
Massachusetts	**0.007**	**cos**	**82.07**	**83.34**
	0.000 7	cos	79.08	80.23
	0.007	step	81.34	82.69
Inria	**0.007**	**cos**	**84.35**	**86.66**
	0.000 7	cos	82.00	84.45
	0.007	step	83.40	85.77

2.5 消融实验

为保证各个模块的有效性,本文在 Massachusetts 和 Inria 数据集上进行了消融实验,具体结果如表 3 所示。其中,实验 1 为使用 VGG-16 作为编码器的 U-Net 网络;实验 2 为在实验 1 基础上引入动态上采样;实验 3 为在实验 2 基础上加入多尺度上下文信息提取模块;实验 4 为在实验 3 基础上引入 ECA 注意力机制;实验 5 为在实验 4 基础上使用联合损失函数;实验 6 为在实验 5 基础上加入 CA-DPGHead 模块。

由表 3 所得到的数据进行分析可知,在 Massachusetts 数据集上,引入动态上采样后,mIoU、F1 相较于实验 1 分别提高了 0.35%、0.32%;实验 3 在实验 2 的基础上加入了多尺度上下文信息提取模块,mIoU、F1 相较于实验 2 分别提高了 0.26%、0.27%;实验 4 在实验 3 的基础上加入了 ECA 注意力机制,mIoU、F1 相较于实验 3 分别提高了 0.07%、0.07%;实验 5 在实验 4 的基础上加入了联合损失函数,mIoU、F1 相较于实验 4 分别提高了 0.30%、0.28%;

表 3 消融实验分析

Table 3 Analysis of ablation experiments %

实验	Massachusetts		Inria	
	mIoU	F1	mIoU	F1
1	80.77	82.06	82.77	85.20
2	81.12	82.38	82.97	85.32
3	81.38	82.65	83.74	86.04
4	81.45	82.72	83.79	86.10
5	81.75	83.00	84.23	86.58
6	82.07	83.34	84.35	86.66

实验 6 在实验 5 的基础上加入了 CA-DPGHead 模块,mIoU、F1 相较于组 5 分别提高了 0.32%、0.34%。

同样由表 3 数据可知,在 Inria 数据集上,引入动态上采样后,mIoU、F1 相较于实验 1 分别提高了 0.20%、0.12%;实验 3 在实验 2 的基础上加入了多尺度上下文信息提取模块,mIoU、F1 相较于实验 2 分别提高了 0.77%、0.72%;实验 4 在实验 3 的基础上加入了 ECA 注意力机制,mIoU、F1 相较于实验 3 分别提高了 0.05%、0.06%;实验 5 在实验 4 的基础上加入了联合损失函数,mIoU、F1 相较于实验 4 分别提高了 0.44%、0.48%;实验 6 在实验 5 的基础上加入了 CA-DPGHead 模块,mIoU、F1 相较于组 5 分别提高了 0.12%、0.08%。

为探讨 MCIEM 模块中分支数对模型精度的影响,尤其是用于更好地捕获小尺度目标的空洞率为 3 的分支和捕获大尺度目标的空洞率为 24 的分支,实验结果如表 4 所示,其中,实验 1 为表 3 中的实验 2;实验 2 为去除空洞率为 3 的分支的 MCIEM 模块;实验 3 为去除空洞率为 24 的分支的 MCIEM 模块;实验 4 为本文 MCIEM 模块,实验结果证明了 MCIEM 模块设计的有效性。

表 4 不同 MCIEM 模块的性能比较

Table 4 Performance comparison of different MCIEM modules %

实验	Massachusetts		Inria	
	mIoU	F1	mIoU	F1
1	81.12	82.38	82.97	85.32
2	81.15	82.40	83.52	85.89
3	81.14	82.36	83.51	85.86
4	**81.38**	**82.65**	**83.74**	**86.04**

2.6 对比实验

1)定量分析

为了验证 VDSEC-UNet 的有效性,本文选取了当前主流的 HRNetV2[20]、Segformer[21]、SwinUnet[22]、MMB-Net[23]和 OAU-net[24]等模型在 Massachusetts 和 Inria 数据集上进行对比实验。

各网络模型在 Massachusetts 数据集上的对比结果如表 5 所示，可以看出，VDSEC-UNet 与其他几种模型相比 4 种指标都有提高。其中，mIoU 分别提升了 2.48%、12.07%、2.26%、3.29%、13.56%、3.78%、3.48% 和 2.37%；F1 分别提升了 2.54%、13.85%、2.43%、3.66%、18.68%、4.10%、3.77% 和 2.52%，证明了本文所提出的网络模型的可行性。

表 5　不同模型在 Massachusetts 数据集的实验对比

Table 5　Experimental comparison of different model on Massachusetts dataset　%

模型	mIoU	Precision	Recall	F1
UNet-ResNet50	79.59	79.22	82.45	80.80
PSPNet-ResNet50	70.00	68.62	70.38	69.49
HRNetV2	79.81	81.45	80.37	80.91
Segformer	78.78	81.40	78.04	79.68
SwinUnet	68.51	54.24	80.05	64.66
DeepLabv3＋	78.29	79.39	79.10	79.24
MMB-Net	78.59	79.78	79.37	79.57
OAU-net	79.70	80.76	80.89	80.82
Ours	**82.07**	**83.79**	**82.90**	**83.34**

各网络模型在 Inria 数据集上的对比结果如表 6 所示，可以看出，VDSEC-UNet 与其他几种模型相比四种指标都有提高。其中，mIoU 分别提升了 2.40%、1.08%、2.24%、1.18%、6.70%、0.65%、3.11% 和 2.00%；F1 分别提升了 2.30%、1.04%、2.18%、1.11%、7.13%、0.58%、3.02% 和 1.96%，证明了本文所提出的网络模型的可行性。

表 6　不同模型在 Inria 数据集的实验对比

Table 6　Experimental comparison of different model on Inria dataset　%

模型	mIoU	Precision	Recall	F1
UNet-ResNet50	81.95	84.50	84.23	84.36
PSPNet-ResNet50	83.27	86.16	85.08	85.62
HRNetV2	82.11	85.30	83.67	84.48
Segformer	83.17	85.52	85.58	85.55
SwinUnet	77.65	73.91	86.07	79.53
DeepLabv3＋	83.70	85.72	86.44	86.08
MMB-Net	81.24	84.23	83.05	83.64
OAU-net	82.35	85.81	83.61	84.70
Ours	**84.35**	**86.74**	**86.59**	**86.66**

2）定性分析

图 7 和图 8 给出了本文方法与其他模型的可视化分割结果。

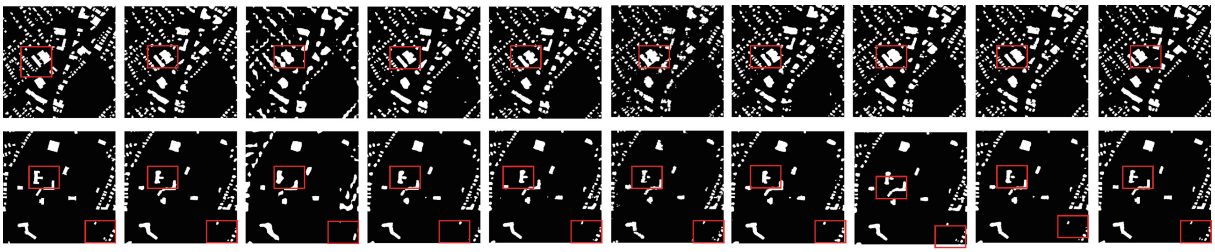

图 7　在 Massachusetts 数据集上提取结果对比

Fig. 7　Comparison of extraction results on Massachusetts dataset

图 8　在 Inria 数据集上提取结果对比

Fig. 8　Comparison of extraction results on Inria dataset

由图 7(b)上半部分和图 7(b)下半部分可以看出，U-Net-ResNet50 提取小型建筑物边缘清晰，但存在漏分现象。由图 8(b)上半部分和图 8(b)下半部分可以看出，U-Net-ResNet50 提取大型建筑物时效果极差，不仅存在错分现象，还存在空洞现象。

由图 7(c)上半部分和图 7(c)下半部分可以看出，

PSPNet-ResNet50 提取小型建筑物时效果极差,不仅有边缘模糊、粘连现象,而且还存在漏分现象。由图 8(c)下半部分可以看出,PSPNet-ResNet50 提取大型建筑物时有空洞现象。

由图 7(d)上半部分和图 7(d)下半部分可以看出,HRNetV2 提取小型建筑物边缘清晰,但存在漏分现象。由图 8(d)上半部分和图 8(d)下半部分可以看出,HRNetV2 提取大型建筑物时不仅存在错分现象还存在空洞现象。

由图 7(e)上半部分和图 7(e)下半部分可以看出,Segformer 提取小型建筑物时有漏分现象。由图 8(e)上半部分和图 8(e)下半部分可以看出,Segformer 提取大型建筑物时存在边缘不清晰和少量空洞现象。

由图 7(f)上半部分和图 7(f)下半部分可以看出,SwinUnet 提取小型建筑物时有边缘模糊现象。由图 8(f)上半部分和图 8(f)下半部分可以看出,SwinUnet 提取大型建筑物时,存在边缘不清晰和错分现象。

由图 7(g)上半部分和图 7(g)下半部分可以看出,Deeplabv3+提取小型建筑物时边缘不够清晰且存在漏分现象。由图 8(g)上半部分和图 8(g)下半部分可以看出,Deeplabv3+提取大型建筑物时边缘不够清晰。

由图 7(h)上半部分和图 7(h)下半部分可以看出,MMB-Net 提取小型建筑物时边缘不够清晰且存在错分现象。由图 8(h)上半部分和图 8(h)下半部分可以看出,MMB-Net 提取大型建筑物时有错分和空洞现象。

由图 7(i)上半部分和图 7(i)下半部分可以看出,OAU-net 提取小型建筑物时边缘不够清晰且存在漏分现象。由图 8(i)上半部分和图 8(i)下半部分可以看出,OAU-net 提取大型建筑物时有错分和空洞现象。

由图 7(j)上半部分、图 7(j)上半部分、图 8(j)下半部分和图 8(j)下半部分可以看出,本文方法提取结果与标签近似,优于其他网络模型,不仅对大型建筑物、小型建筑物都能准确提取,而且其所提取的小型建筑物边缘较为清晰,减少了粘连现象。

3 结 论

近年来卷积神经网络在遥感影像建筑物提取研究中取得了极大的成功,但其仍然面临着整体提取精度不高、错分、漏分和边界模糊等问题。针对以上问题,提出一种基于 VDSEC-UNet 的遥感影像建筑物提取方法,并在 Massachusetts 和 Inria 数据集上进行训练和测试。通过与多个相关语义分割模型进行对比,得到以下结论。首先,VDSEC-UNet 在 Massachusetts 和 Inria 数据集上的 mIoU 分别达到了 82.07% 和 84.35%,F1 指数分别达到了 83.34%和 86.66%,均优于经典方法。其次,加入的多尺度上下文信息提取模块,提升了对不同尺度建筑物目标的提取效果;同时嵌入的 ECA 注意力机制,抑制了遥感影像中复杂背景的干扰,提高了模型对影像中建筑物特征的关注度;此外,联合损失函数解决了类别不平衡问题。以上改进使得 VDSEC-UNet 对大型建筑物、小型建筑物都能准确提取。最后,加入的动态上采样模块可以提高模型对细节的感知能力,进而提高模型对建筑物边缘的提取精度;此外,加入的 CA-DPGHead 模块可以强建筑物与背景之间的区分,使模型更加精准地定位和识别图像中的建筑物信息,细化建筑物边缘的提取效果。以上改进使得 VDSEC-UNet 所提取的小型建筑物边缘较为清晰,减少了粘连现象。未来将进一步减少模型的参数量,以适用于资源受限环境或实时处理场景,从而实现更高效的建筑物提取。

参考文献

[1] XIAO W, CAO H, TANG M, et al. 3D urban object change detection from aerial and terrestrial point clouds: A review[J]. International Journal of Applied Earth Observation and Geoinformation, 2023, 118: 103258.

[2] ZHANG Y P, YANG G, GAO AI, et al. An efficient change detection method for disaster-affected buildings based on a lightweight residual block in high-resolution remote sensing images[J]. International Journal of Remote Sensing, 2023, 44(9): 2959-2981.

[3] MA Y CH, CHEN SH, ERMON S, et al. Transfer learning in environmental remote sensing[J]. Remote Sensing of Environment, 2024, 301: 113924.

[4] 安文,杨俊峰,赵羲,等.高分辨率遥感影像的建筑物自动提取[J].测绘科学, 2014, 39(11): 80-84.
AN W, YANG J F, ZHAO X, et al. Buildings automatic extraction from high resolution RS images[J]. Science of Surveying and Mapping, 2014, 39(11): 80-84.

[5] 何曼芸,程英蕾,廖湘江,等.融合光谱特征和几何特征的建筑物提取算法[J].激光与光电子学进展, 2018, 55(4): 380-387.
HE M Y, CHENG Y L, LIAO X J, et al. Building extraction algorithm by fusing spectral and geometrical features[J]. Laser & Optoelectronics Progress, 2018, 55(4): 380-387.

[6] 闫祎巧,王宏生,赵怀慈,等.融合多尺度卷积和注意力机制的场景提取方法[J].电子测量技术, 2023, 46(16): 172-178.
YAN Y Q, WANG H SH, ZHAO H C, et al. Scene extraction methods incorporating multi-scale convolution and attention mechanisms[J]. Electronic Measurement Technology, 2023, 46(16): 172-178.

[7] 任远锐,陈朋弟,高小龙.基于增强注意力门控 U-Net 的建筑物提取研究[J].全球定位系统, 2024, 49(2): 43-53.

REN Y R, CHEN P D, GAO X L. Building extraction based on advanced attention gate U-Net[J]. GNSS World of China, 2024, 49(2): 43-53.

[8] CHEN J, ZHANG D, WU Y, et al. A context feature enhancement network for building extraction from high-resolution remote sensing imagery[J]. Remote Sensing, 2022, 14(9): 2276.

[9] YE H, ZHOU R, WANG J, et al. FMAM-Net: Fusion multi-scale attention mechanism network for building segmentation in remote sensing images[J]. IEEE Access, 2022, 10: 134241-134251.

[10] QIU W Y, GU L J, GAO F, et al. Building extraction from very high-resolution remote sensing images using refine-UNet[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.

[11] WANG H Y, MIAO F. Building extraction from remote sensing images using deep residual U-Net[J]. European Journal of Remote Sensing, 2022, 55(1): 71-85.

[12] YANG D, GAO X, YANG Y, et al. CSA-Net: Complex scenarios adaptive network for building extraction for remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 938-953.

[13] LIU W, LU H, FU H, et al. Learning to upsample by learning to sample[C]. IEEE/CVF International Conference on Computer Vision, 2023: 6027-6037.

[14] 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.

[15] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[17] HOU Q B, ZHOU D Q, FENG J SH. Coordinate attention for efficient mobile network design[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713-13722.

[18] NI Z, CHEN X, ZHAI Y, et al. Context-guided spatial feature reconstruction for efficient semantic segmentation[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 239-255.

[19] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision(ECCV), 2018: 3-19.

[20] 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, 2019: 5693-5703.

[21] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[22] CAO H, WANG Y, CHEN J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 205-218.

[23] ZHANG H, ZHENG X, ZHENG N, et al. A multiscale and multipath network with boundary enhancement for building footprint extraction from remotely sensed imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 8856-8869.

[24] SONG H, WANG Y, ZENG S, et al. OAU-net: Outlined attention U-net for biomedical image segmentation[J]. Biomedical Signal Processing and Control, 2023, 79: 104038.

作者简介

张剑飞(通信作者),博士,教授,主要研究方向为图像处理与模式识别。

E-mail:zjfnefu2008@163.com

王友为,硕士研究生,主要研究方向为图像处理。

E-mail:3542429161@qq.com