

doi: 10.6046/zrzyyg.2023146

引用格式: 曲海成, 梁旭. 融合混合注意力机制与多尺度特征增强的高分辨率影像建筑物提取[J]. 自然资源遥感, 2024, 36(4): 107-116. (Qu H C, Liang X. Building extraction from high-resolution images using a hybrid attention mechanism combined with multi-scale feature enhancement[J]. Remote Sensing for Natural Resources, 2024, 36(4): 107-116.)

# 融合混合注意力机制与多尺度特征增强的高分辨率影像建筑物提取

曲海成, 梁旭

(辽宁工程技术大学软件学院, 葫芦岛 125105)

**摘要:** 由于复杂背景变换和建筑物形状多样化等因素影响, 从高分辨率遥感图像中准确提取建筑物信息面临着挑战。该文提出了一种融合混合注意力机制与多尺度特征增强的高分辨率建筑物语义分割网络 (building mining net, BMNet)。首先, 编码器部分使用 VGG-16 作为主干网络来提取特征, 得到 4 层特征表示; 然后设计解码器用于解决多尺度信息中高层特征的细节信息丢失问题, 引入了混合通道注意力和空间注意力的串联注意力机制 (series attention module, SAM), 增强高层特征的代表能力; 同时, 设计了一种渐进式特征增强的建筑物信息挖掘模块 (building mining module, BMM), 进一步提高建筑物分割的准确性。BMM 把上采样后的特征映射, 经过 SAM 处理的特征映射以及初始预测结果作为输入, 获取背景噪声信息, 并利用所设计的上下文信息探索模块滤除背景信息, 在经过多次 BMM 处理后得到最佳预测结果。对比实验结果表明: BMNet 在武汉大学建筑数据集上精度和交并比分别优于 U-net 4.6% 和 4.8%, 在马萨诸塞州建筑数据集和 Inria 航空图像标注数据集上精度和交并比分别优于 U-net 7.9%, 8.9% 和 6.7%, 11.0%, 验证了所提模型的有效性以及实用性。

**关键词:** 语义分割; 高分辨率遥感影像; 建筑物提取; U-net; 注意力机制; 空洞卷积

**中图法分类号:** TP 751 **文献标志码:** A **文章编号:** 2097-034X(2024)04-0107-10

## 0 引言

遥感影像建筑物识别和分割在城市规划、城市变化监测、土地资源管理、地质灾害监控和国防安全等方面具有重要意义<sup>[1-4]</sup>。通过识别和提取高分辨率遥感图像中的建筑物, 可以为城乡规划和管理、土地资源勘查与评估、地质灾害监控、军事情报分析等提供精确的空间信息支持, 从而满足决策制定、资源管理和国防安全等领域的需要。

遥感影像建筑物分割的本质是建立起遥感影像与待分割的建筑物目标之间的映射模型, 在此过程中需要对输入遥感影像进行特征提取, 构建高效的特征空间。目前, 用于遥感影像建筑物特征提取的传统方法可以分为基于形状<sup>[5]</sup>、基于支持向量机<sup>[6]</sup> (support vector machine, SVM) 和基于随机森林分类器<sup>[7]</sup>的方法。基于形状的方法在处理复杂建筑物

形状时可能存在困难; 基于 SVM 的方法在处理建筑物和非建筑物样本数量不平衡的数据时, 容易倾向于预测样本数量较多的类别, 从而降低对少数类别的预测精度, 进而影响整体精度; 当特征数量很高时, 随机森林分类器可能会出现过拟合的问题。此外, 这些传统方法对于中低分辨率、信息量少且简单的遥感影像虽然有较好的检测效果, 但随着遥感影像分辨率的提高, 所包含的信息量增大, 这些传统方法的误检率和漏检率都会增大。当前在高分辨率遥感影像目标识别和分割领域中, 深度神经网络的研究和应用也愈发广泛。Wang 等<sup>[8]</sup>利用全卷积神经网络 (fully convolutional network, FCN) 对遥感影像进行处理, 从而实现了对水域、建筑物以及植被的分类识别, 虽然 FCN 可以通过端到端的训练和推断, 避免了手工操作, 从而简化了建筑物提取的流程, 但是仅利用卷积操作会导致边缘模糊和细节信息丢失, 导致语义信息不够准确, 难以准确提取建筑

收稿日期: 2023-05-23; 修订日期: 2023-09-02

基金项目: 国家自然科学基金面上项目“面向数据特性保持的高光谱影像高效压缩方法研究”(编号: 42271409) 和辽宁省高等学校基本科研项目“基于脉冲混合神经网络的高效能目标检测”(编号: LJKMZ20220699) 共同资助。

第一作者: 曲海成 (1981-), 男, 博士, 副教授, 主要研究方向为遥感图像高性能计算、智能大数据处理等。Email: quhaicheng@lntu.edu.cn。

通信作者: 梁旭 (1998-), 男, 硕士研究生, 主要研究方向为数字图像处理与模式识别。Email: 15047728650@163.com。

物边缘和细节;在该领域中,具有代表性的模型有 U-net<sup>[9]</sup> 和 Deeplab<sup>[10]</sup> 系列等。其中,U-net 模型由 Ronneberger 等<sup>[9]</sup> 在 2015 年提出,由对称的编码器和解码器结构组成,U-net 中的跳跃连接可以捕捉更多的上下文信息,从而提高图像分割的准确性,但是对于对于遥感图像中存在的光照变化、背景遮挡等情况的适应性较弱,会导致分割效果不佳;Chen 等<sup>[10]</sup> 提出的 DeepLab 系列采用空白金字塔池化(atrous spatial pyramid pooling,ASPP)模块对目标进行多尺度的特征提取,利用解码器融合不同尺度的特征进行精细分割,但在高分辨率遥感图像的分割任务中仍存在细节丢失严重等问题;Fu 等<sup>[11]</sup> 提出了一个名为双重注意力网络(dual attention network,DANet)的模型,用于解决场景分割任务,该模型基于双重注意力机制,能够捕捉丰富的上下文依赖关系;2022 年,金澍等<sup>[12]</sup> 提出了一种引入卷积块注意力模块和 ASPP 的 MA-Unet,该网络通过加强建筑物特征和获取建筑物更精细的边缘特征提高了建筑物的分割精度;Yu 等<sup>[13]</sup> 将一系列注意力门控模块与单个 U-net 网络集成在一起,抑制输入图像中无关和嘈杂的特征响应,提高了提取精度,但是仅引入注意力机制并不能更好地关注小目标建筑物;He 等<sup>[14]</sup> 提出了一种用于遥感图像语义分割的新型框架,名为 ST-UNet,该框架将 Swin Transformer 嵌入传统的基于 CNN 的 U-net 模型中,以获得全局上下文信息,同时克服了传统 CNN 在直接获取全局上下文方面的困难。此外,遥感影像建筑物分割是一种像素级的语义分割任务,现有基于深度学习的遥感影像分割方法在对目标进行分割时,存在像素关联度低、特征细节信息丢失等问题,从而出现误分割、分割边界不连续的结果<sup>[15-18]</sup>。在实际应用中,建筑

物提取存在以下难点:首先,遥感图像的复杂性是建筑物提取的难点之一,遥感图像中不仅包含建筑物本身的信息,还包含了各种地物和环境的干扰信息;其次,建筑物的多样性也是建筑物提取的难点之一;另外,建筑物的遮挡和重叠也是建筑物提取的难点。

本文基于 U-net 网络提出融合混合注意力机制与多尺度特征增强的高分辨率遥感影像建筑物分割模型(building mining net,BMNet)。为使模型具备一定的鲁棒性和适应性,能够处理各种类型的建筑物,适应不同的背景复杂度,在网络中加入串联型注意力机制。然后,在网络中增设建筑物信息挖掘模块(building mining module,BMM),对背景噪声进行滤除,降低建筑物周围复杂背景信息对前景建筑物特征信息的干扰,以优化建筑物边缘结构;提出上下文信息探索模块(context exploration block,CEB),引入不同尺度的空洞卷积以丰富模型感受野,提高对小型建筑物的特征提取能力。

## 1 建筑物提取网络模型

### 1.1 BMNet 整体结构设计

将传统的 U-net 网络直接应用于遥感影像建筑物分割任务中会存在一些不足:小建筑目标受像素点数量影响分割效果不佳;建筑物前景与复杂背景之间通常存在粘连关系,这会导致建筑物前景特征提取存在干扰和噪声,进而影响模型对遥感影像中建筑物分割的最终效果。因此,本文构建了一个用于高分遥感影像的建筑物分割网络模型 BMNet,其整体结构如图 1 所示。首先,本文使用预训练的 VGG16-BN 模型,将其分为 5 个部分。其中,第一

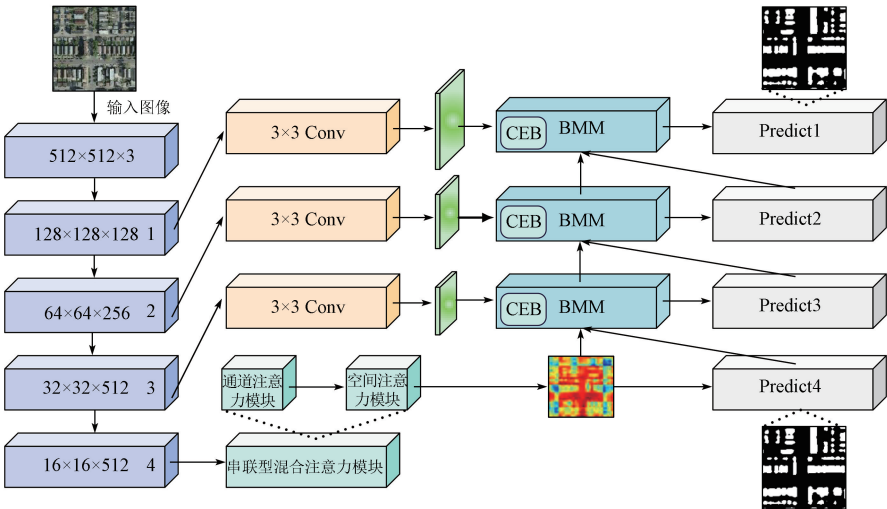


图 1 BMNet 网络结构图  
Fig.1 BMNet network structure

部分为输入层,接下来是 4 个下采样层,每个层的输出通道数逐渐增加。在最后一层下采样层引入串联型注意力模块,以增强第 4 个下采样层的输出。接着,使用 3 个 BMM 将相邻 2 个下采样层间的特征进行融合和处理。每个 BMM 模块将当前下采样层的输入特征、下一层模块的输出以及对应的下采样层特征进行融合,生成新的特征图和提取结果。最后,使用双线性插值将每个下采样层的输出进行上采样,得到与原始图像相同大小的预测结果。模型的输出包括 Predict4, Predict3, Predict2 和 Predict1,分别表示下采样层 4,3,2 和 1 的预测结果。其中,通道注意力模块用于增强特征通道之间的相关性,空间注意力模块则用于增强特征之间的空间位置相关性,二者均有助于提高模型性能。建筑物挖掘模块允许模型从多个层次获取不同抽象级别的特征,并将它们融合在一起,以生成更有意义的特征表示。

### 1.2 串联型混合注意力模块

为了提高建筑物提取模型的精度、效率和鲁棒性,本文引入了串联型混合注意力模块<sup>[19-21]</sup>,先应用通道注意力,再进行空间注意力处理,从而在网络中引入了多层次的注意力机制,增强了网络对输入图像的建模能力。混合注意力模块设计如图 2 所示。

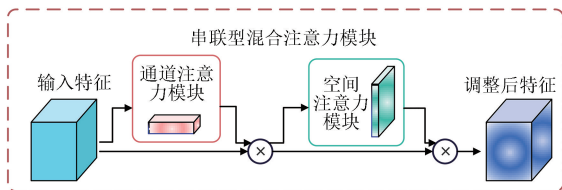


图 2 混合注意力模块

Fig.2 Mixed attention module

通道注意力模块旨在通过计算输入图像各个通道的权重来提高特征表示能力。借助通道注意力,可以提高深度学习网络提取特征的性能,使其能够更好地理解输入图像并提取更有意义的特征。

通道注意力机制的具体实现步骤如下:首先,分别通过全局平均池化(global average pooling, GAP)和全局最大池化(global max pooling, GMP)将  $H \times W \times C$  的输入图像压缩为  $1 \times 1 \times C$  大小的通道响应值。该过程将全局空间信息转化为通道响应值,减少网络参数量,防止过拟合。接下来,通过 2 个全连接网络获得注意力权重矩阵,这些权重矩阵用于表示不同通道的重要性,反映了包含关键信息的通道,这些通道会得到更多的关注。最后,将得到的注意力权重矩阵与原始特征图进行逐元素乘法运算,得到校准后的注意力特征图。假设通道注意力模块的输入为  $F$ ,输出为  $F_c$ ,实现过程如图 3 所示。每一步输出特征的计算方式为:

$$F_{\max} = \text{Conv}_{1 \times 1}(\text{GMP}(F)), \quad (1)$$

$$F_{\text{mean}} = \text{Conv}_{1 \times 1}(\text{GAP}(F)), \quad (2)$$

$$F_{\text{fusion}} = F_{\max} + F_{\text{mean}}, \quad (3)$$

$$F_c = \text{Sigmoid}(\text{Conv}_{1 \times 1}(F_{\text{fusion}}) \times F), \quad (4)$$

式中:  $\text{Conv}_{1 \times 1}$  为卷积核大小为 1 的卷积操作;  $\text{Sigmoid}$  为非线性激活函数,将输出映射到 0~1 之间。

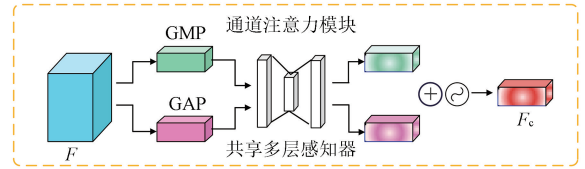


图 3 通道注意力模块

Fig.3 Channel attention module

空间注意力机制的目的关注输入特征图中的空间位置,以找到关键信息在图像上的位置。通道注意力机制主要关注不同通道上的特征权重,而空间注意力机制则通过对通道注意力权重的引导,进一步关注特定位置上的信息聚集情况。

空间注意力的具体实现步骤如下:首先沿着通道轴应用平均池化和最大池化操作,将它们连接起来生成一个有效的特征描述符。这一步骤旨在通过对不同通道之间的数值进行比较,而非同一通道不同区域的数值,得到一个能够描述通道之间关联性的特征描述符。接下来,将这个特征描述符送入一个卷积网络进行卷积操作,并通过激活函数得到最终的空间注意力特征图。这一过程通过对输入特征图中不同位置的信息进行加权融合,从而确定关键信息在图像上的位置分布,以提高网络的特征表示能力。假设空间注意力模块的输入为  $F_c$ ,输出为  $F_s$ ,实现过程如图 4 所示。每一步输出特征的计算方式为:

$$F_{\max} = \text{LMP}(F_c, \text{dim} = 1), \quad (5)$$

$$F_{\text{mean}} = \text{LAP}(F_c, \text{dim} = 1), \quad (6)$$

$$F_{\text{fusion}} = \text{Concat}(F_{\max}, F_{\text{mean}}), \quad (7)$$

$$F_s = \text{Sigmoid}(\text{Conv}_{1 \times 1}(F_{\text{fusion}}) \times F), \quad (8)$$

式中:  $\text{LMP}$  为在通道维度上进行最大池化操作;  $\text{LAP}$  为在通道维度上进行平均池化操作;  $\text{dim}$  为在数据指定维度上执行操作;  $\text{Concat}$  为拼接操作。

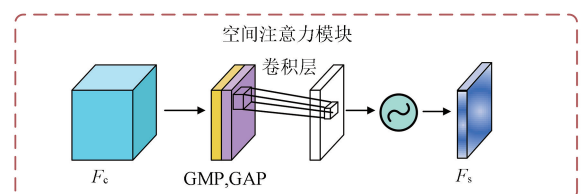


图 4 空间注意力模块

Fig.4 Spatial attention module



### 1.3 建筑物信息挖掘模块

由于建筑物周围经常存在大量的背景干扰,如树木、道路遮挡等,经典模型对于建筑物前景和背景的划分不够清晰,难以准确地分离出建筑物,导致误检率提升,因此本文进行了如下操作:对高层预测结果中的前景(接近 1)与背景(接近 0)进行 *Sigmoid* 操作,用 1 减去 *Sigmoid* 后的结果,并与上采样后的高层特征  $F_h^{\text{up}}$  相乘,得到背景噪声特征  $F_1^{\text{b}}$ ;接着引入 CEB 模块丰富背景噪声特征的感受野和上下文信息,通过可学习参数  $\alpha$  自适应地滤除低层特征中的背景信息,得到鲁棒和清晰的特征表示,从而获得更精细的建筑物分割图。假设输入低层特征  $F_l$ ,高层特征  $F_h$ ,高层预测结果  $P_h$ ,输出融合特征  $F_{\text{refine}}$  和更精细的分割结果  $P_1$ ,则各层特征及

分割结果的计算方式如下:

$$F_h^{\text{up}} = \text{Upsample}(\text{Conv}_{7 \times 7}(F_h)), \quad (9)$$

$$P_h^{\text{up}} = \text{Upsample}(\text{Sigmoid}(P_h)), \quad (10)$$

$$F_1^{\text{b}} = \text{CEB}((1 - P_h^{\text{up}}) \times F_l), \quad (11)$$

$$F_{\text{refine}} = \text{Relu}(\text{BN}(F_h^{\text{up}} - \alpha \times F_1^{\text{b}})), \quad (12)$$

$$P_1 = \text{Conv}_{3 \times 3}(F_{\text{refine}}), \quad (13)$$

式中:角标 up 为上采样操作;*BN* 为批归一化(batch normalization)操作;*Relu* 为修正线性单元操作。

BMM 模块的具体框架如图 5 所示,图中  $d$  表示卷积操作中的膨胀率,用于控制卷积核的感受野大小。

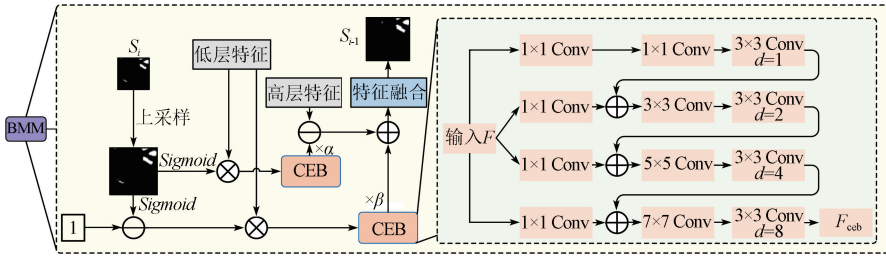


图 5 建筑物挖掘模块结构图

Fig.5 Building mining module structure

为了使模型更加关注小建筑物目标,CEB 模块采用多个尺度分支并应用了空洞卷积,包含 4 个不同采样率的空洞卷积层,分别为 1,2,4 和 8,以获得不同大小的感受野。这些分支通过通道维度实现多尺度的特征融合。通过 CEB 模块,模型能够获取对多尺度上下文信息感知的特征表示能力。假设输入特征  $F$ ,则输出  $F_{\text{coh}}$ ,如图 5 所示;每一个分支(branch)  $b_1, b_2, b_3, b_4$  的输出计算公式为:

$$F_{b1} = \text{Conv}_{3 \times 3}^{d=1}(\text{Conv}_{1 \times 1}(\text{Conv}_{1 \times 1}(F))), \quad (14)$$

$$F_{b2} = \text{Conv}_{3 \times 3}^{d=2}(\text{Conv}_{1 \times 1}(F) + F_{b1}), \quad (15)$$

$$F_{b3} = \text{Conv}_{3 \times 3}^{d=4}(\text{Conv}_{5 \times 5}(\text{Conv}_{1 \times 1}(F) + F_{b2})), \quad (16)$$

$$F_{b4} = \text{Conv}_{3 \times 3}^{d=8}(\text{Conv}_{7 \times 7}(\text{Conv}_{1 \times 1}(F) + F_{b3})), \quad (17)$$

$$F_{\text{coh}} = \text{Conv}_{1 \times 1}(\text{Concat}(F_{b1}, F_{b2}, F_{b3}, F_{b4})). \quad (18)$$

### 1.4 损失函数

本文采用交叉熵损失作为损失函数。一般损失函数对自变量求偏导的结果中带有 *Sigmoid* 导数因子项,而当自变量值较大时,*Sigmoid* 导数值会趋向 0,这样整个梯度会变得很小,从而在使用梯度下降法更新自变量时,迭代速度变慢,导致难以收敛或者发散,进而导致模型无法在早期阶段达到最优值。而交叉熵损失函数求导后不带 *Sigmoid* 函数导数

项,梯度下降时更新很快,从而能够较早达到最优结果,加速模型的收敛速度。

为保证实验的公平性和可行性,损失函数计算公式为:

$$\text{Loss} = -\frac{1}{M} \left[ \sum_{i=1}^M \sum_{j=1}^N p_{ij} \ln(q_{ij}) \right], \quad (19)$$

式中: $M$  为样本总数; $N$  为类别数量; $p_{ij}$  为第  $i$  个样本属于第  $j$  个类别的概率; $q_{ij}$  为模型预测的第  $i$  个样本属于第  $j$  个类别的概率。

## 2 实验结果与分析

### 2.1 数据集及评价指标

为了评估提出的 BMNet 的性能,综合考虑了 3 个公开的建筑物数据集进行实验,包括武汉大学建筑数据集([https://study.rsgis.whu.edu.cn/pages/download/building\\_dataset.html](https://study.rsgis.whu.edu.cn/pages/download/building_dataset.html))、马萨诸塞州建筑数据集(<https://www.cs.toronto.edu/~vmnih/data/>)和 Inria 航空图像标注数据集(<https://project.inria.fr/aerialimagelabeling/>)。

1) 武汉大学建筑数据集<sup>[22]</sup>由航空影像数据集和卫星数据集 2 个子集组成,本实验中仅使用航空



影像数据集,航空影像取自海拔 0.075 m 空间分辨率和 450 km<sup>2</sup>覆盖土地,来自新西兰的新基督城。该数据集被裁剪为 8 188 个图块,训练集数量为 4 736(包含 130 500 座建筑物),验证集数量为 1 036(包含 14 500 座建筑物),测试集为 2 416(包含 42 000座建筑物)。每个图块的大小为 512 像素×512 像素。

2) 马萨诸塞州建筑数据集由波士顿地区的 151 张航空影像组成,每张影像均为 1 500 像素×1 500 像素,面积为 2.25 km<sup>2</sup>,覆盖面积约 340 km<sup>2</sup>,地面官方数据被拆分为包含 137 张图像的训练集、包含 10 张图像的测试集和包含 4 张图像的验证集,为了使模型在此数据集有更好的泛化性能,采用垂直翻转和水平翻转等数据增强技术扩展了训练集,扩展后的训练集含 400 张图像。数据集涉及城市和郊区场景,建筑物的大小、形状、纹理和颜色都各不相同,因此这个数据集非常具有挑战性,适合用来验证模型的有效性。

3) Inria 航空图像标注数据集<sup>[23]</sup>包含 5 个不同的城市定居点的不同建筑物和非建筑物图像,每个区域包含 180 张高分辨率航空图像,覆盖区域从人口密集的地区(旧金山的金融区)到高山城镇(奥地利蒂罗尔的利恩茨)。原始图像大小为 5 000 像素×5 000 像素,为了满足计算机计算需求,首先将每张原始图像转换为 RGB 格式,将图片填充为 5 120 像素×5 120 像素大小,然后将它们分割成尺寸为 512 像素×512 像素的小块,并将符合条件(即建筑物类别像素占比不低于 5%)的小块作为实验数据集,最终将 9 695 张作为训练集,1 945 张作为验证集,1 545张作为测试集。

本文选择通用的交并比(intersection over union, IOU)、正样品分类精度(Precision)、正样品召回率(Recall)和 F1 分数(F1)来评估网络模型的性能,公式为:

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}, \quad (20)$$

$$Precision = \frac{TP}{TP + FP}, \quad (21)$$

$$Recall = \frac{TP}{TP + FN}, \quad (22)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Recall + Precision}, \quad (23)$$

式中:TP为判断正样本正确的数量;FP为将正确

的负样本判断为正样本;FN为将正确的正样本判断为负样本;IoU为预测结果与真实标签之间交集与并集之间的比值,IoU 值越大,目标分割精度越大;F1 为 Precision 和 Recall 的调和平均数,F1 值越大,分类模型性能越好。

2.2 实验设置

本实验训练测试环境基于 Pytorch 框架,版本为 Pytorch1.8.1,CUDA11.1,采用 AdamW 优化器,采用余弦退火和热重启学习率优化策略训练所有模型,采用随机垂直或水平旋转进行数据增强。首先对武汉大学建筑数据集进行训练,初始学习率为0.001,batch size 设置为 8,epoch 设置为 150。对于马萨诸塞州的建筑数据集和 Inria 航空图像标记数据集,使用在武汉大学建筑数据集上训练得到的权重进行微调,学习速率调整为 0.000 5。在测试阶段,采用了测试时间增强策略,对图片进行翻转操作。

硬件环境为:操作系统 Ubuntu18.04,处理器为 NTEL Intel Xeon Gold 5218R,显卡为英伟达 RTX3090,内存为 32 G。

2.3 消融实验及分析

为了验证所提出的 BMNet 模型各模块的有效性,在 Inria 航空图像标注数据集上开展了消融实验。消融实验主要从分割结果的 IoU, Precision, Recall, F1 等指标在不同组合模型上进行对比,各组合模型的相关参数设置和训练策略均保持一致,本文选择带有 VGG-NET 编码器的 U-net 作为基线,直接用图像和建筑物标签训练。所有对比模型输入图像分辨率为 512 像素×512 像素。表 1 列出网络各组合模型的配置与说明,消融实验对比结果见表 2。

表 1 BMNet 模型组合结构  
Tab.1 Combination structure of BMNet

| 网络组合<br>模块 | 说明                            |
|------------|-------------------------------|
| BM-0       | 基线:为仅使用 VGG-16 作为主干的 U-net 网络 |
| BM-1       | 基线+空间注意力模块(SA)                |
| BM-2       | 基线+空间注意力模块+通道注意力模块(SA+CA)     |
| BM-3       | 基线+ SA+CA +BMM(不添加 CEB 模块)    |
| BM-4       | 基线+ SA+CA +BMM(添加 CEB 模块)     |
| BM-5       | 基线+通道注意力模块(CA)                |
| BM-6       | 基线+ CA+BMM(不添加 CEB)           |
| BM-7       | 基线+ CA+BMM(添加 CEB)            |
| BM-8       | 基线+ SA+BMM(不添加 CEB)           |
| BM-9       | 基线+ SA+BMM(添加 CEB)            |
| BM-10      | 基线+BMM(不添加 CEB)               |
| BM-11      | 基线+BMM(添加 CEB)                |

表 2 消融实验结果  
Tab.2 Results of ablation experiment (%)

| 网络组合模块 | <i>IoU</i> | <i>F1</i> | <i>Precision</i> | <i>Recall</i> |
|--------|------------|-----------|------------------|---------------|
| BM-0   | 71.40      | 83.33     | 84.60            | 82.10         |
| BM-1   | 72.67      | 86.96     | 89.36            | 84.68         |
| BM-2   | 74.51      | 88.60     | 90.68            | 86.62         |
| BM-3   | 79.78      | 89.35     | 91.35            | 87.44         |
| BM-4   | 81.93      | 90.07     | 91.78            | 88.41         |
| BM-5   | 73.28      | 85.06     | 85.96            | 84.17         |
| BM-6   | 78.51      | 85.72     | 86.54            | 84.92         |
| BM-7   | 80.66      | 86.45     | 87.02            | 85.88         |
| BM-8   | 77.94      | 87.66     | 90.10            | 85.35         |
| BM-9   | 80.09      | 87.30     | 88.92            | 85.74         |
| BM-10  | 76.67      | 84.91     | 86.32            | 83.55         |
| BM-11  | 78.82      | 85.69     | 86.79            | 84.62         |

由表 2 数据可知,在加入了空间、通道混合注意力机制以及 BMM 后,基线网络的性能得到了显著提升。具体而言,BM-1 在 *IoU*, *F1*, *Precision*, *Recall* 这 4 个指标上比 BM-0 提高了 1.78%~5.63%; BM-2 在 4 个指标上比 BM-0 提高了 4.36%~7.19%; BM-3 在 *IoU* 指标上比 BM-0 提高了 11.74%; BM-4 比 BM-0 在 4 个指标上平均提高了 9.76%。此外,从 BM-6/7, BM-8/9, BM-10/11 的对比结果中

验证了 CEB 模块的有效性。BM-1 与 BM-5 的实验对比结果可以看出,空间注意力和通道注意力对模型的影响不相上下,但与 BM-2 的对比结果可以看出,空间注意力与通道注意力结合时,模型能够获得更好的性能。表 2 的结果进一步验证了本文所提出的方法的有效性和优越性。结果表明,混合注意力机制可以集中网络的注意力于重要区域,同时强化有用特征,减少无用信息,从而提高网络对于图像特征的感知和理解能力。BMM 通过调整空洞卷积中的空洞数目,可以实现多尺度特征提取,从而使得网络可以同时关注不同尺度的特征,使模型的特征提取能力进一步增强。

图 6 为消融实验生成的中间结果图,即特征热图。从图(e)中可以看出前景(建筑物)和背景特征并不明显,较为粗略;而图(h)中可以明显地看出,相比较基线的分割预测结果(e)更为精细。此结果得益于 BMM 对前景与背景特征的加强,CEB 模块中空洞卷积可以增加特征图中的采样点数目,从而可以更好地提取图像的细节特征。

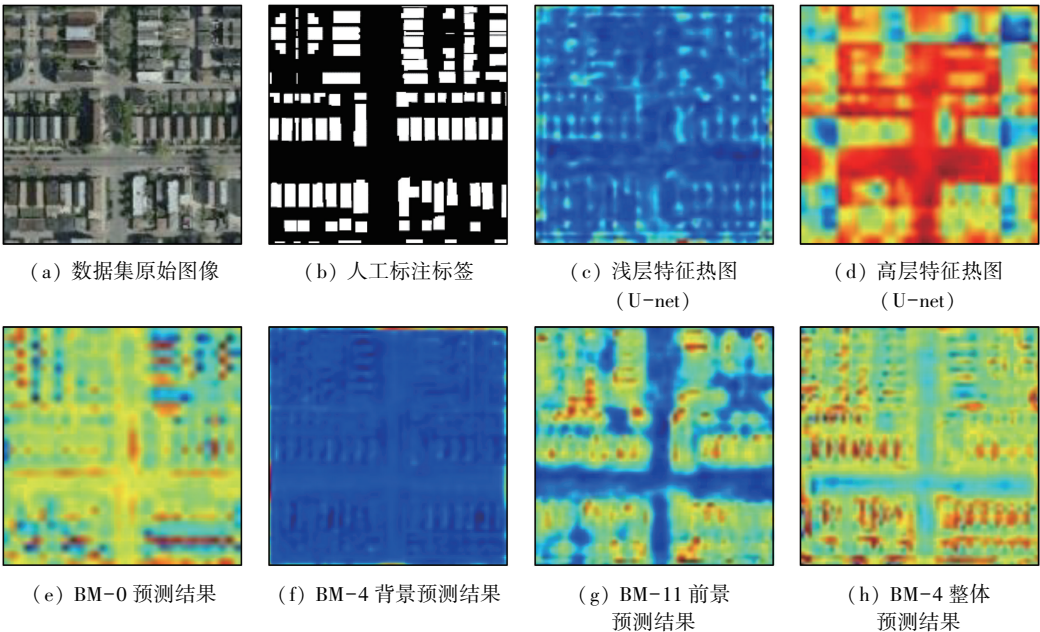


图 6 消融实验中间结果图(特征图)

Fig.6 Figures of intermediate results of ablation experiment

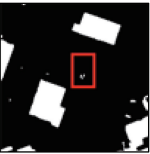
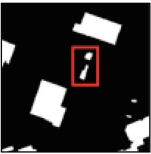
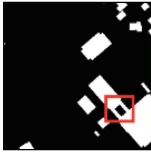
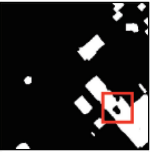
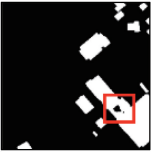
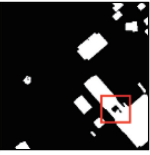
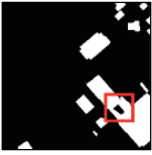
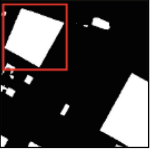
2.4 对比实验及分析

为了验证所提出的 BMNet 的有效性,本文将 BMNet 与现有方法在马萨诸塞州建筑数据集、武汉大学建筑数据集和 Inria 航空建筑数据集上进行了比较。选择经典的分割方法进行对比,并且每个方法在相应数据集上取得较为先进的指标,因此对比方法有 U-net<sup>[9]</sup>, DeepLabV3+<sup>[10]</sup>, ST-UNet<sup>[14]</sup>, SwinUpNet<sup>[15]</sup>, BANet<sup>[16]</sup> 和 STT<sup>[17]</sup>。定量分析采

用的实验评价指标为交并比(*IoU*)、F1 分数(*F1*)、精确度(*Precision*)、召回率(*Recall*)和参数量(*Param*)。表 3 为各类对比方法在武汉大学建筑数据集部分可视化分割结果。红色框表示待分割目标,从图中分割的位置准确性和目标完整性可以看出,本文所提方法在所选的 3 张图片中均取得了最优的分割效果。

表 3 武汉大学建筑数据集定对比实验结果

Tab.3 Qualitative comparison experiment results of WHU Building Dataset

| 序号 | 图像                                                                                | 地面真值                                                                              | U-net                                                                             | DeeplabV3+                                                                        | STT                                                                               | ST-UNet                                                                            | BMNet                                                                               |
|----|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| 1  |  |  |  |  |  |  |  |
|    |  |  |  |  |  |  |  |
|    |  |  |  |  |  |  |  |

上述不同算法在武汉大学建筑数据集上分割结果的各指标定量分析见表 4。由表 4 可知,提出的 BMNet 算法  $IoU$  为 91.73%, $Precision$  为 95.01%, $Recall$  为 95.27%, $Param$  为 21.34 MB。实验表明:相较于 VGG16-U-net,DeeplabV3+,STT 等对比算法,对比指标  $IoU$  分别提升了 5.54%,2.99%,1.25%;指标  $Precision$  提升了 0.17%~5.09%;指标  $Recall$  提升了 0.19%~1.00%;指标  $F_1$  分别提升了 2.94%,0.97%和 0.18%。总体参数量高于 U-net 和 STT,小于 DeeplabV3+和 ST-Unet。综上所述,与对比算法相比,本文提出的 BMNet 算法在武汉大学建筑数据集上具有较好的分割效果,且计算参数量都在可接受的范围。但是由于该数据集影像分辨率高,建筑物与其他地物区分比较明显,建筑的边缘比较清晰,所

以在可视化对比实验中的各种算法差距不大。

表 4 武汉大学建筑数据集定量对比实验结果

Tab.4 Quantitative comparison experiment results of WHU Building Dataset

| 方法          | $IoU/\%$ | $F1/\%$ | $Precision/\%$ | $Recall/\%$ | $Param/MB$ |
|-------------|----------|---------|----------------|-------------|------------|
| VGG16-U-net | 86.91    | 92.42   | 90.41          | 94.52       | 18.26      |
| DeepLab V3+ | 89.07    | 94.23   | 94.15          | 94.32       | 30.61      |
| STT         | 90.48    | 94.97   | 94.85          | 95.09       | 21.02      |
| ST-UNet     | 89.65    | —       | 95.16          | —           | 26.47      |
| BMNet       | 91.73    | 95.14   | 95.01          | 95.27       | 21.34      |

表 5 为各类对比方法在马萨诸塞州建筑数据集上的部分可视化分割结果。从表中可以看出,本文所提方法在所选的 3 张图片中均取得了最优的分割效果。

表 5 马萨诸塞州建筑数据集定性对比实验结果

Tab.5 Qualitative comparison experiment results of Massachusetts Buildings Dataset

| 序号 | 图像                                                                                  | 地面真值                                                                                | U-net                                                                               | DeeplabV3+                                                                          | BANet                                                                               | ST-UNet                                                                              | BMNet                                                                                 |
|----|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 1  |  |  |  |  |  |  |  |
|    |  |  |  |  |  |  |  |
|    |  |  |  |  |  |  |  |

不同算法在马萨诸塞州建筑数据集上的各指标结果见表 6。由表 6 可知,提出的 BMNet 算法在

萨诸塞州建筑数据集上的  $IoU$  为 75.79%, $Precision$  为 87.02%, $Recall$  为 84.19%, $Param$  为 21.34 MB。



表 6 马萨诸塞州建筑数据集定量对比实验结果

Tab.6 Quantitative comparison experiment results of Massachusetts Buildings Dataset

| 方法          | IoU/% | F1/%  | Precision/% | Recall/% | Param/MB |
|-------------|-------|-------|-------------|----------|----------|
| VGG16-U-net | 66.92 | 80.65 | 79.15       | 82.21    | 18.26    |
| DeepLab V3+ | 69.16 | 81.38 | 83.85       | 79.05    | 30.61    |
| BANet       | 72.20 | 83.86 | 83.07       | 84.66    | 38.72    |
| ST-UNet     | 71.26 | —     | 86.49       | —        | 26.47    |
| BMNet       | 75.79 | 85.58 | 87.02       | 84.19    | 21.34    |

实验表明,相较于 4 种对比算法,BMNet 在 *IoU*, *Precision* 和 *F1* 3 个指标上均有提升。尽管 BMNet 在

Recall 上略低于 BANet,但都高于其余对比算法。综上所述,与对比算法相比,本文提出的 BMNet 算法在马萨诸塞州建筑数据集上具有较好的分割效果,且计算参数量都在可接受的范围,并且相比其余对比模型有显著的优势,验证了 BMNet 模型的鲁棒性和泛化性。

表 7 为各类对比方法在 Inria 航空数据集上的可视化分割结果。由表 7 可知,本文所提方法在 Inria 航空数据集上具有更好的分割结果,从小建筑物目标到较大建筑物目标的提取结果都更为精细,分割效果更接近人工标注的地面真值。

表 7 Inria 航空图像数据集定性对比实验结果

Tab.7 Qualitative comparison experiment results of Inria Aerial Image Labeling Datasets

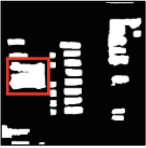
| 序号 | 图像                                                                                  | 地面真值                                                                                | U-net                                                                               | DeeplabV3+                                                                          | SwinUpNet                                                                           | ST-UNet                                                                              | BMNet                                                                                 |
|----|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 1  |    |    |    |    |    |    |    |
| 2  |   |   |   |   |   |   |   |
| 3  |  |  |  |  |  |  |  |

表 8 为 Inria 航空图像标注数据集定量对比实验结果,从中可以看出,对于 Inria 航空图像标记数据集,本文所提方法仍然保持了更先进的性能,其 *IoU* 为 81.93%,*F1* 达到 90.06%。

表 8 Inria 航空图像标注数据集定量对比实验结果

Tab.8 Quantitative comparison experiment results of Inria Aerial Image Labeling Datasets

| 方法          | IoU/% | F1/%  | Precision/% | Recall/% | Param/MB |
|-------------|-------|-------|-------------|----------|----------|
| VGG16-U-net | 70.92 | 83.12 | 85.07       | 81.26    | 18.26    |
| DeepLab V3+ | 74.52 | 83.47 | 81.82       | 85.19    | 30.61    |
| SwinUpNet   | 79.37 | 88.19 | 87.43       | 88.96    | 65.84    |
| ST-UNet     | 70.14 | —     | 91.36       | —        | 26.47    |
| BMNet       | 81.93 | 90.06 | 91.78       | 88.41    | 21.34    |

对于大规模遥感图像分割,有的模型要么结构复杂<sup>[23-24]</sup>,要么捕获的建筑物语义信息不够丰富<sup>[25-28]</sup>,因此需要建立高效的建筑物提取算法,能够在保证高精度的同时,快速处理大量的遥感图像数据,满足实际应用需求。VGG16 中的卷积层使用了非常小的卷积核,并且每个卷积层都有多个通道。这种设计使得模型参数的数量大大减少,从而减少

了过拟合的风险。相比于传统的卷积操作,空洞卷积可以在不改变特征图大小的情况下,增加卷积核中的空洞数目,使模型在计算中不增加过多的参数数量。这在大规模的神经网络中尤为重要,因为可以有效地减少模型的复杂度和训练时间。由图 7 可以看出(蓝色柱状表示参数量,红色折线表示 *IoU* 值),本文的方法在满足精度优势的同时,少量增加模型的参数量,可以运用到实际提取任务中。

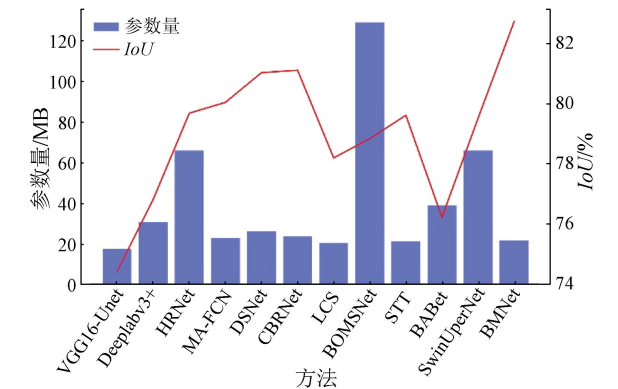


图 7 模型参数量和 IoU 数据对比

Fig.7 Comparison graph of parameter count and IoU data

3 结论

与之前的研究工作类似,本文选择经典的 U-net 网络作为基础网络提出了一个用于高分辨率遥感影像建筑物分割的模型 BMNet。

1)通过混合空间、通道注意力提高对各类型建筑物的分割能力,以提高模型的鲁棒性和适应性;引入建筑物信息挖掘模块 BMM,降低与建筑物前景粘连在一起的复杂背景信息对建筑物特征提取的干扰;结合所设计的上下文信息探索模块 CEB 对背景噪声进行滤除,通过引入不同尺度的空洞卷积以丰富模型的感受野。

2)与主流方法的对比实验结果表明,本文所提出的模型在武汉大学建筑数据集、马萨诸塞州建筑数据集和 Inria 航空图像标注数据集上都取得了相对更优的分割性能,对目标分割的准确率高。大量的消融实验也验证了本文所提模型中各个部件的有效性。

但出于对模型实际应用的考量,尤其是在移动端设备上的应用,未来将更多地关注于在保证模型精度的前提下,设计更加轻量化的模型。

参考文献 (References):

[1] 徐宗学,程涛,洪思扬,等.遥感技术在城市洪涝模拟中的应用进展[J].科学通报,2018,63(21):2156-2166.  
Xu Z X, Cheng T, Hong S Y, et al. Review on applications of remote sensing in urban flood modeling[J]. Chinese Science Bulletin, 2018, 63(21): 2156-2166.

[2] 向煜,黄志,华媛媛,等.深度学习支持下的宅基地复垦项目真实性智能审查技术研究与应用[J].测绘通报,2023(1):163-167.  
Xiang Y, Huang Z, Hua Y Y, et al. Research and application of intelligent verification technology for authenticity of homestead reclamation project based on deep learning[J]. Bulletin of Surveying and Mapping, 2023(1): 163-167.

[3] 张莹,郭红梅,尹文刚,等.基于特征提取的 SVM 图像分类技术的无人机遥感建筑物震害识别应用研究[J].灾害学,2022, 37(4):30-36,56.  
Zhang Y, Guo H M, Yin W G, et al. Application of SVM image classification technology based on feature extraction in seismic damage identification of buildings by UAV remote sensing[J]. Journal of Catastrophology, 2022, 37(4): 30-36, 56.

[4] Nielsen M M. Remote sensing for urban planning and management: The use of window-independent context segmentation to extract urban features in Stockholm[J]. Computers, Environment and Urban Systems, 2015, 52:1-9.

[5] Karantzalos K, Paragios N. Recognition-driven two-dimensional competing priors toward automatic and accurate building detection [J]. IEEE Transactions on Geoscience and Remote Sensing, 2009,

47(1):133-144.

[6] Huang X, Zhang L. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2013, 51(1): 257-272.

[7] Belgiu M, Drăgut L. Comparing supervised and unsupervised multi-resolution segmentation approaches for extracting buildings from very high resolution imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 96:67-75.

[8] Wang Y, Wang C, Zhang H. Integrating H-A- $\alpha$  with fully convolutional networks for fully PolSAR classification[C]//2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP). May 18-21, 2017, Shanghai, China. IEEE, 2017: 1-4.

[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[10] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 833-851.

[11] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 3141-3149.

[12] 金澍,关沫,边玉婵,等.基于改进 U-Net 的遥感影像建筑物提取方法[J].激光与光电子学进展, 2023, 60(4): 3788/LOP213004.  
Jin S, Guan M, Bian Y C, et al. Building extraction from remote sensing images based on improved U-Net[J]. Laser & Optoelectronics Progress, 2023, 60(4): 3788/LOP213004.

[13] Yu M, Chen X, Zhang W, et al. AGs-unet: Building extraction model for high resolution remote sensing images based on attention gates U network[J]. Sensors, 2022, 22(8): 2932.

[14] He X, Zhou Y, Zhao J, et al. Swin transformer embedding UNet for remote sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4408715.

[15] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). October 10-17, 2021, Montreal, QC, Canada. IEEE, 2021: 9992-10002.

[16] Wang L, Li R, Wang D, et al. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images[J]. Remote Sensing, 2021, 13(16): 3065.

[17] Chen K, Zou Z, Shi Z. Building extraction from remote sensing images with sparse token transformers[J]. Remote Sensing, 2021, 13(21): 4441.

[18] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 5686-5696.

[19] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 3-19.

[ 20 ] Hu J,Shen L,Sun G.Squeeze-and-excitation networks[ C ]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18–23,2018,Salt Lake City, UT, USA.IEEE,2018;7132–7141.

[ 21 ] Jaderberg M,Simonyan K,Zisserman A, et al.Spatial transformer networks[ J ].Advances in Neural Information Processing Systems, 2015,28.

[ 22 ] Ji S,Wei S, Lu M.Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[ J ].IEEE Transactions on Geoscience and Remote Sensing, 2019,57( 1 ) :574–586.

[ 23 ] Maggiori E,Tarabalka Y,Charpiat G, et al.Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark[ C ]//2017 IEEE International Geoscience and Remote Sensing Symposium ( IGARSS ). July 23 – 28, 2017, Fort Worth, TX, USA.IEEE,2017;3226–3229.

[ 24 ] Zhou Y,Chen Z,Wang B, et al.BOMSC-net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery[ J ].IEEE Transactions on Geoscience and Remote Sensing,2022,60;5618617.

[ 25 ] Fang L,Zhang L,Nie D, et alAutomatic brain labeling via multi-atlas guided fully convolutional networks[ J ].Medical Image Analysis,2019,51 :157–168.

[ 26 ] Zhang H,Liao Y, Yang H, et al.A local-global dual-stream network for building extraction from very-high-resolution remote sensing images[ J ].IEEE Transactions on Neural Networks and Learning Systems,2022,33( 3 ) :1269–1283.

[ 27 ] Guo H,Du B,Zhang L, et al.A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery[ J ].ISPRS Journal of Photogrammetry and Remote Sensing, 2022,183;240–252.

[ 28 ] Liu Z,Shi Q,Ou J.LCS: A collaborative optimization framework of vector extraction and semantic segmentation for building extraction [ J ].IEEE Transactions on Geoscience and Remote Sensing,2022, 60;5632615.

Building extraction from high-resolution images using a hybrid attention mechanism combined with multi-scale feature enhancement

QU Haicheng, LIANG Xu

( College of Software , Liaoning Technical University , Huludao 125105 , China )

**Abstract:** Accurately extracting building information from high-resolution remote sensing images faces challenges due to complex background transformations and the diversity of building shapes. This study developed a high-resolution building semantic segmentation network-building mining net ( BMNet ), which integrated a hybrid attention mechanism with multi-scale feature enhancement. First, the encoder utilized VGG-16 as the backbone network to extract features, obtaining four layers of feature representations. Then, a decoder was designed to address the issue of detail loss in high-layer features within multi-scale information. Specifically, a series attention module ( SAM ), which combined channel attention and spatial attention, was introduced to enhance the representation capabilities of high-layer features. Additionally, the building mining module( BMM ) with progressive feature enhancement was designed to further improve the accuracy of building segmentation. With the upsampled feature mapping, the feature mapping post-processed using SAM, and initial prediction results as input, the BMM output background noise information and then filtered out background information using the context information exploration module designed in this study. Optimal prediction results were achieved after multiple processing using the BMM. Comparative experiment results indicate that the BMNet outperformed U-Net, with accuracy and intersection over union ( IoU ) increasing by 4.6% and 4.8%, respectively on the WHU Building dataset, by 7.9% and 8.9%, respectively on the Massachusetts buildings dataset, and by 6.7% and 11.0%, respectively on the Inria Aerial Image Labeling Dataset. These results validate the effectiveness and practicality of the proposed model.

**Keywords:** semantic segmentation; high spatial resolution remote sensing image; building information extraction; U-net; attention mechanism; dilated convolution

( 责任编辑: 张 仙 )