DOI:10.26974/j.cnki.XBGC.2026.03.009

面向遥感影像建筑物提取的边界感知动态原型网络

董 畅[1], 郭中华[1,2*], 周 末[1], 马全聪[1]
(1.宁夏大学 电子与电气工程学院,宁夏 银川 750021;
2.能源电子测控技术宁夏重点实验室,宁夏 银川 750021)

摘 要:针对高分辨率遥感影像建筑物提取中边界模糊、背景干扰和轮廓恢复不完整等问题,提出了一种边界感知动态原型网络(DPB-Net)。该网络首先利用平行交叉条带注意力提取水平和垂直方向的空间信息,以增强建筑物边界区域的表达能力;然后,通过双模式动态原型解码器对建筑物区域特征进行建模,降低复杂背景对分割结果的干扰;最后,结合上下文感知重组解码器和边界感知门控机制,逐步恢复建筑物边界与局部轮廓细节,提高预测结果的边界完整性。在WHU和Inria建筑物数据集上的实验结果表明,该方法在WHU数据集上的交并比达到93.94%,与现有方法相比,95%Hausdorff距离降低近20%,并在区域分割精度与边界恢复质量方面表现出优势;在Inria数据集上,交并比达到84.80%,95%Hausdorff距离为18.720 4 px。该方法能够改善复杂城市场景下建筑物区域分割和边界保持效果。

关键词:建筑物提取;高分辨率遥感影像;边界感知;动态原型;深度学习
中图分类号:TP751.1 **文献标志码**:A

建筑物作为城市空间最核心的基础设施,其精准的轮廓提取在城市制图、数字孪生、灾害评估及土地资源管理等应用中具有不可替代的支撑价值[1]。相较于常规的遥感地物分割,建筑物提取的独特性与挑战性在于:不仅要求对主体区域进行准确的语义判定,更需要精细刻画屋顶边界、角点几何结构,并严格保持相邻建筑间的空间分离。然而,在真实的复杂城市场景中,阴影遮挡、植被覆盖以及类屋顶材质的干扰无处不在;加之建筑物自身尺度跨度大、排列紧密且形态迥异,常规方法极易出现误检、漏检、边界平滑及相邻建筑物粘连等现象。

伴随深度神经网络的演进,高分辨率遥感影像建筑物提取的研究范式在不断突破。已有研究从多任务学习、注意力编码-解码结构和多尺度特征融合等方面开展改进,在建筑物区域识别中取得了较好的效果[2-4]。随后,注意力机制和Transformer结构被引入遥感影像分割任务,用于增强长距离上下文建模能力和复杂场景下的特征表达。面向建筑物提取任务,BuildFormer采用全局上下文路径和空间细节路径相结合的方式,在建模长距离依赖的同时保留局部空间细节[5]。纵观近期的主流研究,技术路线正进一步向多源特征的精细化协同演进:UA-Net引入不确定性预测来锚定易混淆区域[6];MARS-Net和3DJA-UNet3+利用多维注意力机制强化局部与全局特征的耦合[7-8];TDFNet将V-Mamba与CNN(convolutional neural network)特征进行融合,并采用两次解码方式改善建筑物提取结果[9];SECT-Net构建了一种稀疏边缘感知的CNN-Transformer混合架构,以增强建筑物轮廓的提取精度[10];LMVMamba将CNN、Transformer与Mamba结构相结合,通过多尺度特征融合增强遥感影像的全局上下文建模和局部细节表达能力[11]。这些研究表明,建筑物提取方法正由单一卷积解码逐渐向全局上下文建模、多尺度结构保持和边界细节增强的方向发展。

尽管现有方法在特征聚合与上下文建模层面取

收稿日期:2026-06-05
基金项目:国家自然科学基金项目(62365016)
作者简介:董畅(2000—),男,硕士研究生,主要从事遥感图像处理研究(12024130768@stu.nxu.edu.cn)。
* **通信作者**:郭中华(1973—),男,教授,博士,主要从事遥感图像处理研究(guozhh@nxu.edu.cn)。
引用格式:董畅,郭中华,周末,等.面向遥感影像建筑物提取的边界感知动态原型网络[J].西北工程技术学报(中英文),2026,25(3):256-265.

得了长足进展,但在极致的高分辨边界提取要求下,仍面临三大难以规避的技术瓶颈。首先是浅层空间细节的退化,主干网络中连续的下采样操作不可避免地会削弱建筑物边缘、尖角和细长轮廓等几何先验,直接导致预测结果的边界模糊[12]。其次是复杂背景引发的语义混淆,在复杂城市环境中,道路、裸地或特定植被往往表现出与建筑物高度相似的纹理特征,极易在语义解码阶段引发误导性响应。最后是高分辨率轮廓重建能力不足,常规的上采样或简单的跳跃连接融合难以提供足够的结构信息来充分恢复高频轮廓,从而造成相邻建筑物的粘连与局部边界断裂。因此,如何在维持深层语义强大判别力的同时,有效引入浅层空间先验并全面强化高分辨率边界的恢复能力,成为该领域破局的关键。

针对上述问题,本文提出一种边界感知动态原型网络(boundary-aware dynamic prototype network,DPB-Net)。不同于单纯依赖深层语义特征恢复建筑物掩膜的方法,DPB-Net 围绕空间先验保持—原型语义重建—高分辨率边界细化的逐级建模思路展开设计:首先,从原始影像中提取浅层空间先验,并通过方向结构建模增强建筑物边界相关信息;随后,利用动态原型对深层语义特征进行重建,降低复杂背景响应在解码过程中的传播;最终,通过两级上下文感知重组和边界感知门控机制逐步恢复高分辨率轮廓细节。本文旨在提高复杂城市场景下建筑物区域提取精度和边界恢复质量,增强模型对背景干扰及复杂轮廓的适应能力。

1　边界感知动态原型网络

1.1　网络总体结构

图 1 呈现了边界感知动态原型网络(DPB-Net)的整体架构。该模型以经典的编码器-解码器范式为骨架,构建了四大核心模块:语义特征提取、空间结构信息保持、动态原型解码以及边界细化重建。在特征提取阶段,引入 ConvNeXt V2 作为语义编码器,以捕获多尺度深层语义表征[13];同时,图像引导模块(image-guided stem,IGS)从原始输入影像中提取浅层空间先验特征,并生成不同分辨率的结构引导信息。随后,平行交叉条带注意力(parallel cross-strip attention,PCSA)模块对上述先验特征进行定向强化,通过水平与垂直维度的拓扑建模,为建筑物的刚性边界注入丰富的结构补充。进入解码阶段,双模式动态原型解码器(dual-mode dynamic prototype decoder,DPD)通过实施全局与局部原型的差异化交互,实现深层特征的逐级重构。之后,上下文感知重组解码器(context-aware reassembly decoder,CARD)将解码特征与空间结构信息进行融合,恢复高分辨率细节。最后,边界感知门控(boundary-aware gating,BAG)机制对边界区域响应进行调节,进一步增强建筑物轮廓的一致性。整体上,该网络通过语义编码—原型解码—边界重建的逐级处理方式,实现建筑物区域分割与边界细化的协同优化。

1.2　图像引导与条带注意力模块

高分辨率遥感影像中的建筑物通常具有较明显的边缘、角点和局部规则轮廓,但这些细节信息在主干网络多次下采样过程中容易被削弱。为在早期阶段保留与建筑物边界相关的空间信息,本文构建了由图像引导模块和平行交叉条带注意力模块组成的几何先验分支。图像引导模块(IGS)用于从原始输入影像中生成多尺度空间先验特征,平行交叉条带注意力模块则进一步增强水平和垂直方向的结构响应,二者结合能够在深层语义特征之外补充与边界相关的几何引导信息,从而缓解主干网络下采样造成的边界细节损失,并为后续 CARD 和 BAG 的高分辨率边界恢复提供结构依据。

如图 2 所示,图像引导模块首先采用 3×3 卷积和 1×1 卷积对输入影像进行初始特征映射,然后,通过不同膨胀率的空洞卷积分支提取多尺度空间信息,并经特征拼接与 1×1 卷积压缩得到共享空间先验特征 F_{IGS}^{512}。随后,该特征被转换为两个层级的结构引导特征:高分辨率分支 F_{IGS}^{512} 用于最终边界恢复,降采样分支 F_{IGS}^{256} 用于中间尺度特征重建。通过上述结构,IGS 能够为后续解码过程提供浅层几何先验信息。

平行交叉条带注意力(PCSA)模块的结构如图 3 所示。PCSA 模块首先采用 5×5 深度卷积聚合局部上下文信息,得到局部增强特征。随后,分别将局部增强特征输入 $1\times K$ 和 $K\times1$ 两个条带卷积分支中,分别用于建模水平与垂直方向的长距离空间关系。两个方向的响应相加后,经 1×1 卷积和 sigmoid 函数生成条带注意力权重。最后,将条带注意力权重作用于输入先验特征,得到输出特征。

1.3　双模式动态原型解码器

在高分辨率遥感影像建筑物提取中,复杂背景区域与建筑物区域常存在相似纹理,若在解码阶段直接进行密集特征交互,容易引入道路、阴影、植被和类屋顶区域等背景响应,导致建筑物边界附近出

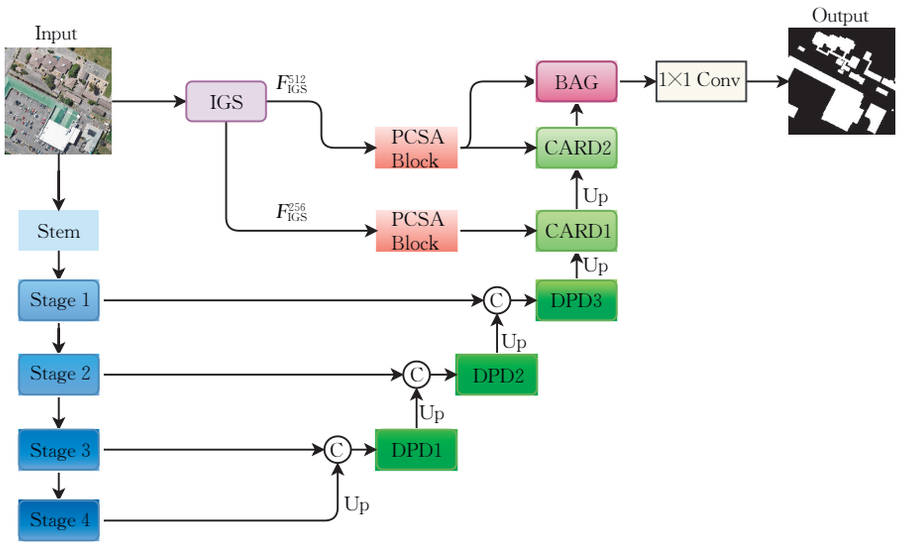


图 1　DPB-Net 总体结构

Fig. 1　Overall architecture of DPB-Net

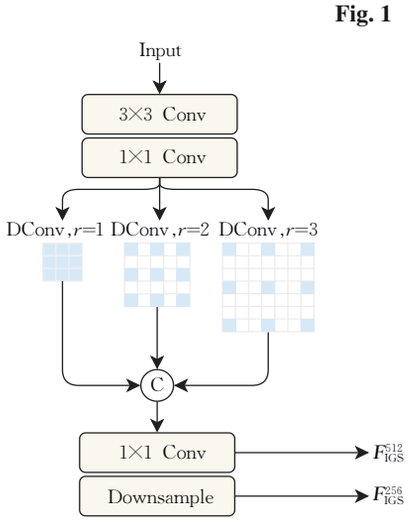


图 2　图像引导模块结构

Fig. 2　Structure of the image-guidance module

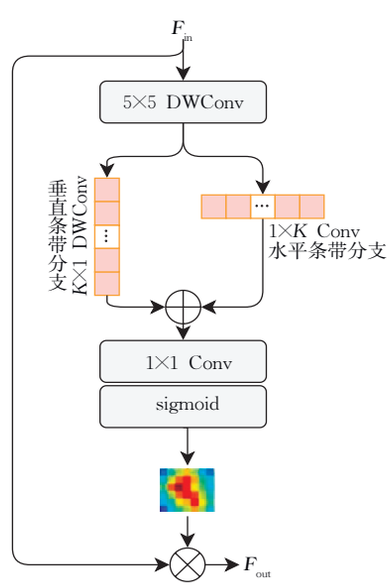


图 3　平行交叉条带注意力模块结构

Fig. 3　Structure of the parallel cross-strip attention module

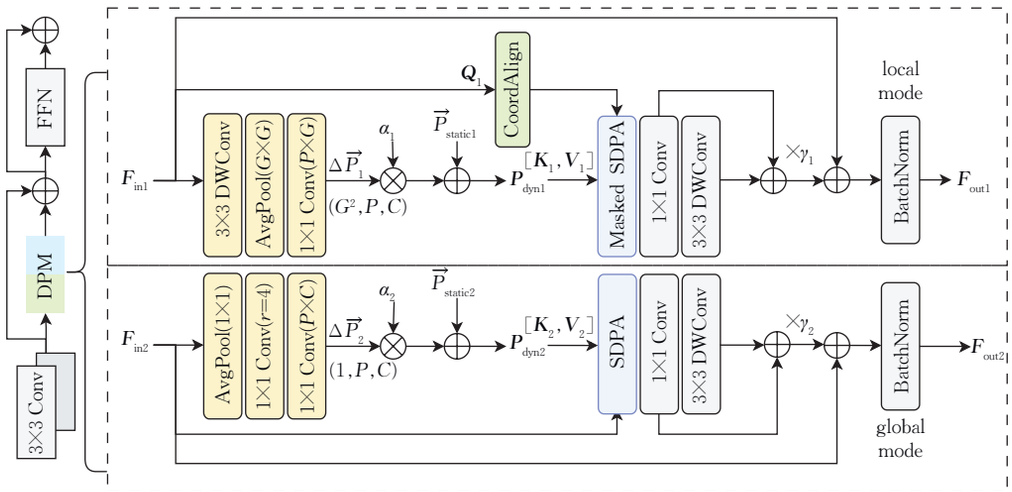
现语义混淆。为降低冗余特征交互并增强建筑物区域判别能力，设计了双模式动态原型解码器(DPD)。

DPD 的结构如图 4 所示，作为中间环节，其在整体网络中承担着语义重建与背景抑制的功能。与直接在所有像素位置之间建立密集交互不同，双模式动态原型解码器以动态原型模块(dynamic prototype module，DPM)为核心单元，根据当前输入特征生成少量与图像内容相关的动态原型，并将其作为注意力机制中的键和值参与特征重建，使解码特征主要与代表性语义原型进行交互。对于建筑物提取任务，动态原型能够聚合屋顶主体和轮廓附近的有效语义模式，同时抑制与建筑物相关性较弱的背景响应，从而减少复杂背景在解码过程中的传播。由此，DPD 不仅用于深层语义特征的聚合与重建，也为后续 CARD 和 BAG 的边界细化提供背景干扰较少的解码特征。

给定共享静态原型矩阵 $\boldsymbol{P}_{\text{static}}$，DPM 根据输入特征生成原型偏移量 $\Delta\boldsymbol{P}$，并与静态原型相加得到最终动态原型 $\boldsymbol{P}_{\text{dyn}}$。即

$$\boldsymbol{P}_{\text{dyn}}=\boldsymbol{P}_{\text{static}}+\alpha\cdot\Delta\boldsymbol{P}。\tag{1}$$

式中，α 为可学习缩放因子，用于控制原型偏移量对动态原型的影响。根据解码特征的空间分辨率不同，DPD 采用全局模式和局部模式两种交互方式。对于低分辨率深层特征，DPD1 采用全局模式，通过全局上下文生成动态原型，用于确定建筑物区域的大致位置，并为后续边界细化提供语义引导。对于



注:FFN—feed-forward network;SDPA—scaled dot-product attention。

图 4　双模式动态原型解码器结构

Fig. 4　Structure of the dual-mode dynamic prototype decoder

较高分辨率特征,DPD2 和 DPD3 采用局部模式,将特征划分为 $G\times G$ 个网格,并在局部网格内生成动态原型,从而限制交互范围,减少远距离背景区域对建筑物细节恢复的干扰。局部模式下的掩膜原型交互过程如图 5 所示。

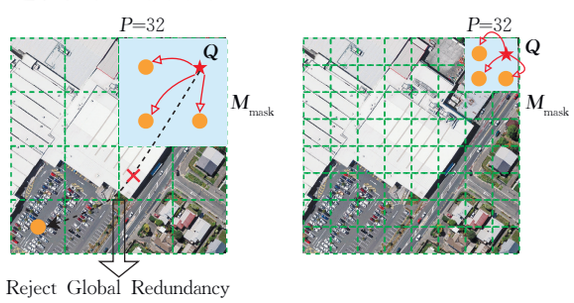


图 5　局部掩膜原型交互机制

Fig. 5　Local mask-prototype interaction mechanism

在局部模式中,首先,将查询位置映射到原型网格。对于特征图中坐标为 (y,x) 的查询位置,其在 $G\times G$ 网格中的相对坐标为

$$y_{\text{grid}}=\frac{(G-1)y}{H-1},\tag{2}$$

$$x_{\text{grid}}=\frac{(G-1)x}{W-1}。\tag{3}$$

式中:H 和 W 分别表示特征图的高度和宽度;G 表示每个空间维度上原型网格数量。随后,根据相对坐标确定左上角锚定网格索引,即

$$i=\text{Clamp}\big(\text{Floor}(y_{\text{grid}}),0,G-2\big),\tag{4}$$

$$j=\text{Clamp}\big(\text{Floor}(x_{\text{grid}}),0,G-2\big)。\tag{5}$$

式中,Clamp(•) 表示索引截断函数,用于避免边界位置产生无效网格索引。根据锚定索引,查询特征仅与相邻 2×2 网格内的原型进行交互,即对应 (i,j)、$(i,j+1)$、$(i+1,j)$ 和 $(i+1,j+1)$ 4 个邻域位置,从而在保持局部连续性的同时减少无关区域参与计算。

在原型注意力计算中,查询特征 $\boldsymbol{Q}_1$ 与动态原型映射得到的 $\boldsymbol{K}_1$、$\boldsymbol{V}_1$ 进行交互,并通过空间掩膜 $\boldsymbol{M}_{\text{mask}}$ 抑制邻域之外的响应。在具体解码过程中,DPD 采用由粗到细的网格演化策略。DPD1 对低分辨率深层特征采用全局模式,获得图像级语义定位;DPD2 在中间尺度特征上采用 $G=4$ 的局部模式,限制区域级原型交互;DPD3 进一步采用 $G=8$ 的局部模式,使原型邻域更加细化,以增强建筑物边界和局部结构恢复能力。通过由全局到局部的动态原型解码方式,DPD 能够先在低分辨率阶段获得建筑物区域的整体语义定位,再在较高分辨率阶段限制局部交互范围,逐步减少远距离背景区域对建筑物细节恢复的影响。经过多阶段原型解码后,网络获得了背景干扰较少、语义结构更稳定的建筑物特征表示,为后续 CARD 的高分辨率重组和 BAG 的边界响应增强提供了语义基础。

1.4　上下文感知重组和边界门控

经过 DPD 解码后,网络已经通过动态原型交互对深层语义特征进行了重建,并在一定程度上抑制了复杂背景响应。然而,DPD 主要侧重于语义层面的区域判别,建筑物边界、角点和细长轮廓等高分辨率空间细节仍需要进一步恢复。为此,本文构建由上下文感知重组解码器(CARD)和边界感知门控机制(BAG)组成的边界细化重建结构,具体过程如图 6 所示。

CARD 由 2 个级联阶段组成,分别对应

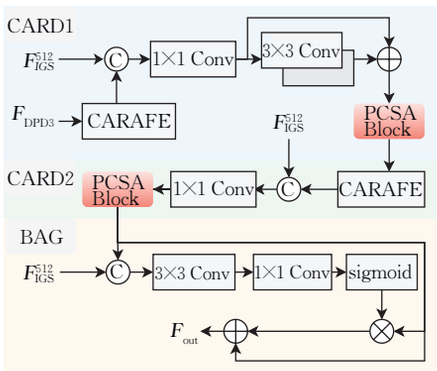


图6 上下文感知重组与边界感知门控结构

Fig. 6 Structures of the context-aware reassembly decoder and boundary-aware gating mechanism

256 px×256 px 和 512 px×512 px 的特征恢复。首先,采用内容感知特征重组(content-aware reassembly of features,CARAFE)上采样方法[14],将 DPD3 的输出特征上采样至与中间尺度空间先验特征一致的分辨率。CARAFE 根据局部内容生成重组权重,与固定插值上采样相比,其能够为建筑物边界附近的细节恢复提供更灵活的特征重建方式。随后,将上采样特征与图像引导模块(IGS)输出的中间尺度空间先验特征 F_{IGS}^{256} 进行拼接,并通过 1×1 卷积完成通道融合。该过程使深层语义特征在恢复过程中重新引入浅层边缘、角点和方向结构信息。融合后的特征进一步经过残差块和 PCSA 细化,得到 CARD1 输出特征,从而增强中间尺度下的局部轮廓表达。之后,采用更简洁的重组结构,将 CARD1 输出进一步恢复到 512 px×512 px,得到高分辨率语义特征。

在 CARD 完成高分辨率特征恢复后,BAG 进一步利用高分辨率空间先验特征调节建筑物边界区域响应。具体地,将 F_{CARD2} 与 F_{IGS}^{512} 进行拼接,并依次经过 3×3 卷积和 1×1 卷积映射,得到边界响应特征。接着,通过 sigmoid 函数生成边界调制权重 $\boldsymbol{W}_{edge}$。将生成的 $\boldsymbol{W}_{edge}$ 用于增强与建筑物轮廓相关的特征响应。为避免门控操作削弱原有语义信息,BAG 采用残差式乘性调制方式,将边界增强结果与原始高分辨率特征相加,得到细化后的输出特征。即

$$F_{out}=F_{CARD2}+F_{CARD2}\odot\boldsymbol{W}_{edge}。\tag{6}$$

CARD 通过内容感知上采样和空间先验融合逐级恢复建筑物细节,BAG 在最终预测前进一步突出边界相关响应。二者共同作用,使网络在复杂背景下能够获得边界更连续、轮廓更完整的建筑物提取结果。

2 数据集及实验方法

2.1 数据集

选用 WHU Building Dataset[15](简称 WHU 数据集)和 Inria Aerial Image Labeling Dataset[16](简称 Inria 数据集)2 个公开建筑物提取数据集进行验证。WHU 数据集主要包含新西兰 Christchurch 地区的航空影像,Inria 数据集包含不同城市区域的高分辨率航空影像,两者均被广泛用于遥感影像建筑物提取方法的性能评估。

2.2 模型评估

为评价模型的区域分割性能,采用精确率(precision,P)、召回率(recall,R)、F_1 值(F_1-score)和交并比(intersection over union,IoU)作为评价指标[17]。具体计算公式为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\tag{7}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\tag{8}$$

$$F_1=\frac{2PR}{P+R},\tag{9}$$

$$\mathrm{IoU}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}。\tag{10}$$

式中:TP 表示正确预测为建筑物的像素数;FP 表示背景像素被误判为建筑物的像素数;FN 表示建筑物像素被误判为背景的像素数。此外,为进一步评价建筑物边界保持能力,采用平均表面距离(average surface distance,ASD)和 95%Hausdorff 距离(HD_{95})衡量预测边界与真实边界之间的差异[17],ASD 用于反映两者边界之间的平均距离,HD_{95}用于反映较大边界偏差情况下的轮廓误差。

2.3 实验环境与参数设置

实验基于 PyTorch 深度学习框架实现,在单张 NVIDIA RTX A6000 GPU 上完成训练与测试。网络采用经 ImageNet 预训练的 ConvNeXt V2 作为语义编码器,输入影像尺寸统一设置为 512 px×512 px。训练过程中采用随机翻转、随机旋转等方式进行数据增强。模型训练轮数设置为 150,批量大小设置为 8,优化器采用 AdamW,初始学习率和权重衰减系数均设置为 1×10^{-4},并采用余弦退火策略调整学习率,最小学习率设置为 1×10^{-6}。

训练过程分为 2 个阶段:前 100 轮采用 Focal 损失、Dice 损失和动态原型正交约束损失共同优化

网络,其中,正交约束损失权重设置为 0.01;第 101 轮至第 150 轮在 CARD1 和 CARD2 后增加轻量化边界辅助监督分支,并引入 Boundary IoU 损失进行边界约束,其损失权重设置为 0.5。辅助监督分支仅在训练阶段使用,推理阶段予以移除,因此不会增加模型测试时的参数量和计算量。

3　结果与分析

3.1　对比实验

为验证所提方法的建筑物提取性能,选取 10 种具有代表性的语义分割和建筑物提取模型进行对比。对比方法覆盖了不同技术路线:U-Net[18]、PSPNet[19]和 DeepLabV3+[20]代表典型卷积语义分割方法;HRNet[21]代表保持高分辨率表示的分割网络;U-Net(Swin-T)[22]、SegFormer[23]和 UPerNet[24]代表基于 Transformer 或层次化特征建模的通用分割框架;BuildFormer[5]代表面向高分辨率遥感建筑物提取的 Transformer 类方法;LMVMamba[11]则代表近期引入状态空间建模思想的遥感建筑物提取方法;TDFNet[9]采用 V-Mamba 与 CNN 双编码器结构,结合双重解码策略进一步增强了建筑物特征表达,是目前较新的建筑物提取方法;SECT-Net[10]也是近期提出的一种结合多尺度边缘监督与稀疏注意力机制的混合网络,在保持高分辨率结构特征方面具有代表性。上述方法既包含了经典分割模型,也包含了近年来在遥感建筑物提取任务中具有代表性的 Transformer 和状态空间模型,能够较全面地评价本文方法在区域识别、边界保持和模型复杂度方面的性能。

考虑到部分最新建筑物提取方法的公开实现、数据划分、训练轮数、数据增强策略及指标计算方式与本文设置并不完全一致,若直接引用其公开论文中的结果,容易造成非统一实验条件下的比较偏差。因此,本文未将不同实验设置下的公开结果直接纳入主对比表,而是在统一实验条件下复现代表性方法,并在相关工作部分补充讨论近年来同类方法的发展趋势与局限。WHU 和 Inria 数据集上的定量对比结果分别如表 1 和表 2 所示。

由表 1 可知,在 WHU 数据集上,DPB-Net 的精确率、召回率、F_1 值和交并比均达到最高值,分别达到 97.52%、96.24%、96.87% 和 93.94%。与表现较好的状态空间模型 LMVMamba 相比,DPB-Net 的交并比提高了 1.19%,F_1 值提高了 0.63%。在边界评价指标方面,本文方法的优势更为明显,其平均表面距离和 95%Hausdorff 距离分别为 2.012 1 px 和 7.927 6 px,较 LMVMamba 分别降低了 0.339 1 px 和 1.918 7 px,其中,95%Hausdorff 距离下降约 19.5%。此外,近期提出的双编码器网络 TDFNet 和混合架构 SECT-Net 在各项指标上也展现了较好的性能(如 SECT-Net 的交并比达到了 91.76%),但 DPB-Net 依然在交并比上比表现最好的同类较新方法高出 2.18%,同时进一步缩小了 95%Hausdorff 误差。结果表明,DPB-Net 不仅能够提高建筑物区域的提取精度,而且能够减小预测边界与真实

表 1　WHU 数据集上的对比实验结果

Tab. 1　Comparative results on the WHU dataset

方法	骨干网络	参数量/10^6个	计算量/10^9次	P/%	R/%	F_1/%	IoU/%	ASD/px	HD_{95}/px
PSPNet	ResNet-50	24.30	81.57	94.40	94.17	94.28	89.18	2.747 1	11.514 2
DeepLabV3+	ResNet-50	26.68	73.81	95.50	94.63	95.06	90.59	3.075 3	11.939 8
U-Net(ResNet)	ResNet-50	32.52	85.66	95.42	94.88	95.15	90.75	2.986 4	11.521 8
HRNet	HRNet-W32	35.99	375.88	95.69	94.51	95.09	90.64	2.598 8	11.008 7
U-Net(Swin)	Swin-T	31.63	79.25	95.62	95.08	95.35	91.11	2.831 7	11.042 3
SegFormer	MiT-B2	27.32	113.24	95.22	95.54	95.38	91.17	2.800 6	10.969 5
UPerNet	ConvNeXt V2-T	37.02	152.90	95.34	95.06	95.20	90.84	2.894 5	11.302 1
BuildFormer	Custom ViT	40.52	233.18	95.75	95.25	95.50	91.38	2.764 1	10.561 8
LMVMamba	ResT	38.95	109.28	96.40	96.08	96.24	92.75	2.351 2	9.846 3
TDFNet	VMamba+VGG16	78.50	245.70	96.06	95.22	95.64	91.64	2.437 8	11.084 5
SECT-Net	ResNet-50	48.46	120.48	95.93	95.48	95.70	91.76	2.543 8	11.512 2
DPB-Net	ConvNeXt V2-T	45.01	126.25	97.52	96.24	96.87	93.94	2.012 1	7.927 6

边界之间的偏差。

从模型复杂度来看,DPB-Net 的参数量和计算量分别为 45.01×10⁶个和 126.25×10⁹次,其计算量低于 HRNet、UPerNet 和 BuildFormer,但高于 LM-VMamba、SegFormer 等部分对比模型。因此,DPB-Net 的主要优势体现在区域分割精度和边界保持能力的提升。

表 2 给出了 Inria 数据集上的对比结果。由于该数据集包含更加多样的建筑物外观和复杂城市场景,不同方法的提取性能均有所下降。DPB-Net 在该数据集上的精确率、召回率、F_1值和交并比分别达到 91.86%、91.69%、91.77% 和 84.80%。与 LM-VMamba 相比,DPB-Net 的交并比提高了 1.15%,F_1值提高了 0.67%,平均表面距离降低了 0.194 5 px,95%Hausdorff 距离降低了 0.812 4 px。上述结果说明,所提方法在不同建筑物数据集上均能够保持较好的区域提取精度和边界恢复效果。

3.2 消融实验

为分析各组成模块对建筑物提取性能的影响,在 WHU 和 Inria 数据集上进行逐步消融实验。基础模型采用以 ConvNeXt V2-Tiny 为编码器、标准双卷积结构为解码器的 U-Net 类网络。在此基础上,依次加入 IGS、DPD、PCSA、CARD 和 BAG,实验结果如表 3 所示。

由表 3 可知,在 WHU 数据集上基础模型的 F_1 值和 IoU 分别为 95.45% 和 91.29%。引入 IGS 后,IoU 提高至 92.16%,初步验证了浅层空间先验对改善区域提取的积极作用;但 ASD 和 HD_{95}未得到改善,表明仅依靠空间先验特征还不足以准确恢复建

表 2 Inria 数据集上的对比实验结果

Tab. 2 Comparative results on the Inria dataset

方法	骨干网络	参数量/10^6个	计算量/10^9次	P/%	R/%	F_1/%	IoU/%	ASD/px	HD_{95}/px
PSPNet	ResNet-50	24.30	81.57	89.47	86.60	88.01	78.59	6.954 0	26.123 8
DeepLabV3+	ResNet-50	26.68	73.81	91.37	87.84	89.57	81.12	6.008 7	22.365 4
U-Net(ResNet)	ResNet-50	32.52	85.66	90.23	88.30	89.25	80.59	6.237 9	23.677 6
HRNet	HRNet-W32	35.99	375.88	90.10	91.57	90.10	81.99	6.295 3	23.705 1
U-Net(Swin)	Swin-T	31.63	79.25	91.61	88.54	90.05	81.90	5.804 8	21.529 5
SegFormer	MiT-B2	27.32	113.24	90.41	91.07	90.41	82.50	5.771 1	20.842 7
UPerNet	ConvNeXt V2-T	37.02	152.90	92.08	88.74	90.38	82.45	5.364 4	20.464 4
BuildFormer	Custom ViT	40.52	233.18	92.12	88.90	90.48	82.62	5.892 4	22.649 1
LMVMamba	ResT	38.95	109.28	91.35	90.85	91.10	83.65	5.201 4	19.532 8
TDFNet	VMamba+VGG16	78.50	245.70	91.24	89.25	90.23	82.21	6.443 7	22.861 7
SECT-Net	ResNet-50	48.46	120.48	91.11	89.26	90.17	82.11	6.329 6	22.775 9
DPB-Net	ConvNeXt V2-T	45.01	126.25	91.86	91.69	91.77	84.80	5.006 9	18.720 4

表 3 DPB-Net 各组成模块的消融实验结果

Tab. 3 Ablation results for the components of DPB-Net

配置	参数量/10^6个	计算量/10^9次	WHU				Inria			
			F_1/%	IoU/%	ASD/px	HD_{95}/px	F_1/%	IoU/%	ASD/px	HD_{95}/px
Base	34.88	87.98	95.45	91.29	2.688 6	11.650 5	90.50	82.65	5.833 4	22.601 3
+IGS	34.90	91.16	95.92	92.16	2.690 2	11.720 9	90.58	82.78	5.927 1	22.792 1
+DPD	41.86	95.60	96.33	92.93	2.569 7	10.809 2	91.06	83.59	5.855 2	21.984 3
+PCSA	41.90	108.08	96.54	93.31	2.349 0	9.721 1	91.27	83.94	5.592 0	20.021 5
+CARD	45.01	122.58	96.63	93.48	2.097 1	8.012 4	91.69	84.66	5.013 8	18.946 7
+BAG	45.01	126.25	96.87	93.94	2.012 1	7.927 6	91.77	84.80	5.006 9	18.720 4

注:各配置均在前一行基础上继续加入对应模块。

筑物边界。在此基础上加入 DPD 模块,IoU 提高至 92.93%,HD_{95}由 11.720 9 px 降低至 10.809 2 px,证实了动态原型交互不仅能提升区域判别力,亦可有效减弱背景噪声对边界预测的干扰。

引入 PCSA 后,WHU 数据集上的 IoU 和 HD_{95} 分别达到 93.31% 和 9.721 1 px,表明水平和垂直方向结构信息的增强有利于改善建筑物轮廓表达。CARD 的加入使边界指标得到进一步提升,WHU 数据集上的 HD_{95}由 9.721 1 px 降低至 8.012 4 px,Inria 数据集上的 HD_{95} 由 20.021 5 px 降低至 18.946 7 px,说明结合空间先验的高分辨率特征重组能够有效恢复建筑物边界细节。最后,引入 BAG 后,完整模型在 WHU 数据集上的 IoU 和 HD_{95}分别达到 93.94% 和 7.927 6 px,在 Inria 数据集上的 IoU 和 HD_{95} 分别达到 84.80% 和 18.720 4 px。

从复杂度变化来看,完整模型的参数量和计算量分别为 45.01×10^6 个和 126.25×10^9 次。与加入 CARD 后的模型相比,BAG 未增加参数量,仅使计算量由 122.58×10^9 次增加至 126.25×10^9 次,同时进一步改善了区域精度和边界指标。综合 2 组数据集上的实验结果可知,IGS、DPD、PCSA、CARD 和 BAG 在建筑物区域提取与边界恢复方面具有互补作用,其中,CARD 对边界质量改善较为明显,BAG 则进一步提升了最终预测结果的轮廓一致性。

为进一步验证上下文感知重组解码器中两级解码块的作用,本文在 WHU 数据集上对 CARD1 和 CARD2 进行消融实验,结果如表 4 所示。在未引入 CARD 的情况下,模型的 IoU 为 93.31%,HD_{95}为 9.721 1 px。单独引入 CARD1 后,IoU 提升至 93.42%,HD_{95}降至 8.521 7 px,说明中间尺度空间先验融合与结构细化能够有效改善建筑物边界恢复质量。单独引入 CARD2 后,IoU 提升至 93.36%,HD_{95}降至 9.128 6 px,也能在一定程度上改善高分辨率恢复效果,但提升幅度小于 CARD1,表明仅进行最终尺度恢复难以充分补充边界细节。当 CARD1 和 CARD2 同时引入时,模型取得最优结果,F_1、IoU 分别达到 96.63% 和 93.48%,ASD 和 HD_{95}分别降低至 2.097 1 px 和 8.012 4 px。结果表明,CARD1 与 CARD2 具有互补作用,前者侧重于中间尺度空间结构融合,后者侧重于最终尺度细节恢复,两级解码块级联能够进一步提升建筑物轮廓连续性和边界恢复精度。

表 4　WHU 数据集上 CARD 消融实验结果

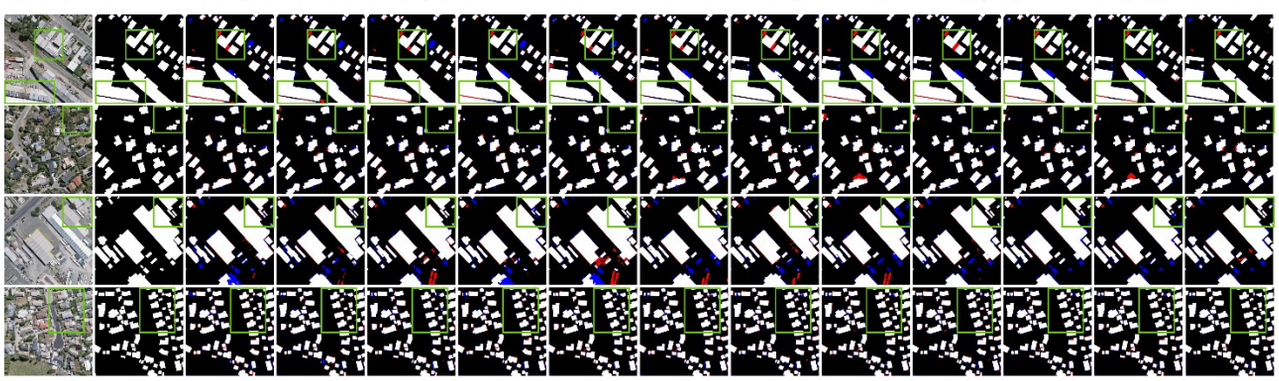
Tab. 4　Ablation results for CARD on the WHU dataset

CARD1	CARD2	F_1/%	IoU/%	ASD/px	HD_{95}/px
×	×	96.54	93.31	2.349 0	9.721 1
√	×	96.60	93.42	2.162 9	8.521 7
×	√	96.57	93.36	2.250 7	9.128 6
√	√	96.63	93.48	2.097 1	8.012 4

3.3　可视化分析

为进一步分析不同方法在复杂场景下的建筑物提取效果,本文选取了 WHU 和 Inria 数据集中的典型样例进行可视化比较,结果如图 7 和图 8 所示,图中白色区域表示正确预测的建筑物像素,黑色区域表示正确预测的背景像素,红色区域表示误检像素,蓝色区域表示漏检像素。

在 WHU 数据集的密集建筑场景中,部分对比方法在相邻建筑物之间容易出现粘连现象,且在屋顶边缘和狭长结构位置存在不同程度的漏检。DPB-Net 能够较完整地恢复建筑物主体区域,并减少边界附近的误检和漏检,说明几何先验分支与边

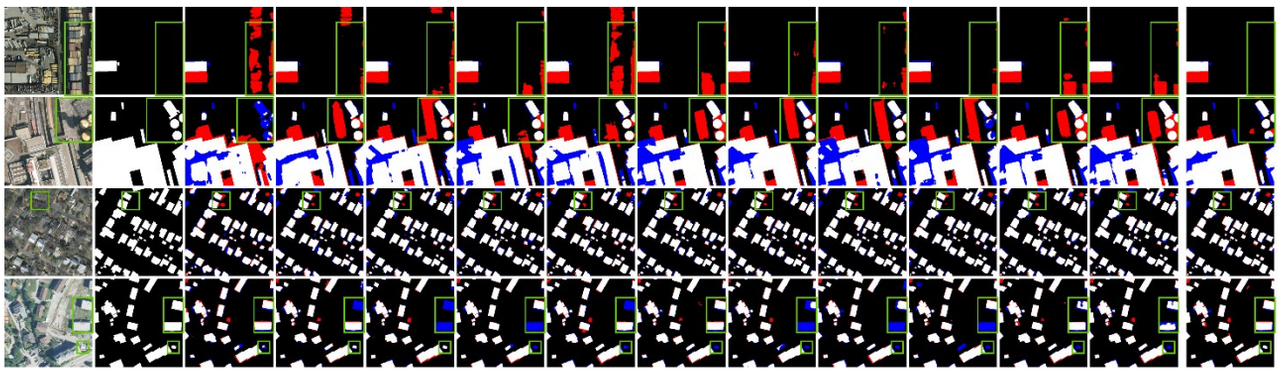


a　b　c　d　e　f　g　h　i　j　k　l　m　n

注:a—原始影像;b—真实标签;c—PSPNet;d—DeepLabV3+;e—U-Net(ResNet);f—HRNet;g—U-Net(Swin);h—SegFormer;i—UPerNet;j—BuildFormer;k—LMVMamba;l—TDFNet;m—SECT-Net;n—DPB-Net。

图 7　WHU 数据集上的建筑物提取结果对比

Fig. 7　Comparison of building extraction results on the WHU dataset



a b c d e f g h i j k l m n

注:a—原始影像;b—真实标签;c—PSPNet;d—DeepLabV3+;e—U-Net(ResNet);f—HRNet;g—U-Net(Swin);h—SegFormer;i—UPerNet;j—BuildFormer;k—LMVMamba;l—TDFNet;m—SECT-Net;n—DPB-Net。

图8 Inria数据集上的建筑物提取结果对比

Fig. 8 Comparison of building extraction results on the Inria dataset

界细化结构有利于保持建筑物轮廓的连续性。

在 Inria 数据集的复杂城市场景中,部分对比方法在背景区域出现误检,或在建筑物轮廓附近产生明显缺失。相比之下,DPB-Net 的预测结果具有更完整的区域覆盖和更清晰的边界表现。该可视化结果与表 1 和表 2 中较低的 ASD 和 HD_{95}指标相一致,进一步说明所提方法能够改善复杂场景下建筑物边界恢复效果。

4 结语

针对高分辨率遥感影像建筑物提取中边界模糊、复杂背景干扰和轮廓恢复不完整等问题,提出了边界感知动态原型网络(DPB-Net)。该方法借助图像引导与平行交叉条带注意力补充建筑物边界相关空间先验,利用双模式动态原型解码器降低背景区域对建筑物特征恢复的干扰,并结合上下文感知重组与边界感知门控逐级恢复高分辨率轮廓细节。在 WHU 和 Inria 数据集上的实验表明,DPB-Net 在区域分割精度和边界评价指标上均取得较好结果,尤其在复杂场景下能够减小预测边界与真实边界之间的偏差。未来的研究重心将聚焦于探索该模型在资源受限设备上的轻量化压缩策略,并进一步提升其在跨域、异构数据分布下的泛化能力。

参考文献:

[1] 范荣双,陈洋,徐启恒,等. 基于深度学习的高分辨率遥感影像建筑物提取方法[J]. 测绘学报,2019,48(1):34-41.

[2] 朱盼盼, 李帅朋, 张立强, 等. 基于多任务学习的高分辨率遥感影像建筑提取[J]. 地球信息科学学报, 2021, 23(3): 514-523.

[3] 李传林, 黄风华, 胡威, 等. 基于Res_AttentionUnet的高分辨率遥感影像建筑物提取方法[J]. 地球信息科学学报, 2021, 23(12): 2232-2243.

[4] 赵元昊, 赵莹莹, 刘东升, 等. 遥感影像建筑物提取多尺度特征深度学习网络[J]. 航天返回与遥感, 2022, 43(4): 25-35.

[5] Wang L B, Fang S H, Meng X L, et al. Building extraction with vision transformer[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5625711.

[6] Li J P, He W, Cao W N, et al. UANet: An uncertainty-aware network for building extraction from remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5608513.

[7] Pang Z J, Hu R M, Zhu W, et al. A building extraction method for high-resolution remote sensing images with multiple attentions and parallel encoders combining enhanced spectral information[J]. Sensors, 2024, 24(3): 1006.

[8] Li Y J, Li Y G, Zhu X B, et al. A method for extracting buildings from remote sensing images based on 3DJA-UNet3+[J]. Scientific Reports, 2024, 14: 19067.

[9] Wang W L, Yu P, Li M M, et al. TDFNet: Twice decoding V-Mamba-CNN fusion features for building extraction[J]. Geo-spatial Information Science, 2026, 29(1): 19-38.

[10] Zhang Y, Deng J Y, Yan W J. Urban building footprints extraction and heights estimation from high-resolution spaceborne remote sensing imagery using a CNN-transformer network [J]. Remote Sensing, 2026, 18(10): 1484.

[11] Li F, Wang X, Wang H C, et al. LMVMamba: A hybrid U-shape mamba for remote sensing segmentation with adaptation fine-tuning [J]. Remote Sensing,

2025, 17(19): 3367.
[12] 张艳, 王翔宇, 张众维, 等. 一种基于边界感知的遥感影像建筑物提取方法[J]. 西安电子科技大学学报(自然科学版), 2022, 49(1): 236-244.
[13] Woo S, Debnath S, Hu R H, et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders[C]//CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 16133-16142.
[14] Wang J Q, Chen K, Xu R, et al. CARAFE: content-aware ReAssembly of FEatures[C]//CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 3007-3016.
[15] Ji S P, Wei S Q, 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.
[16] Maggiori E, Tarabalka Y, Charpiat G, et al. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark[C]//International Geoscience and Remote Sensing Symposium (IGARSS). Piscataway, NJ: IEEE, 2017: 3226-3229.
[17] Taha A A, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool[J]. BMC Medical Imaging, 2015, 15(1): 29.
[18] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention (MICCAI). Cham: Springer International Publishing, 2015: 234-241.
[19] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 6230-6239.
[20] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Computer Vision - ECCV. Cham: Springer International Publishing, 2018: 833-851.
[21] Wang J D, Sun K, Cheng T H, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364.
[22] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 9992-10002.
[23] Xie E Z, Wang W H, Yu Z D, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[C]//Ranzato M, Beygelzimer A, Dauphin Y, et al. Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2021: 12077-12090.
[24] Xiao T T, Liu Y C, Zhou B L, et al. Unified perceptual parsing for scene understanding[C]//Computer Vision - ECCV. Cham: Springer International Publishing, 2018: 432-448.

Boundary-Aware Dynamic Prototype Network for Building Extraction from Remote Sensing Imagery

DONG Chang[1], *GUO Zhonghua*[1,2*], *ZHOU Mo*[1], *MA Quancong*[1]
(1.School of Electronics and Electrical Engineering, Ningxia University, Yinchuan 750021, China; 2. Ningxia Key Laboratory of Energy Electronic Measurement and Control Technology, Yinchuan 750021, China)

Abstract: To address blurred boundaries, background interference, and incomplete contour recovery in building extraction from high-resolution remote sensing imagery, a dynamic prototype boundary-aware network (DPB-Net) is proposed. First, a parallel cross-strip attention module is designed to capture horizontal and vertical spatial information, thereby enhancing feature representation along building boundaries. Second, a dual-mode dynamic prototype decoder is introduced to model building-region features and suppress interference from complex backgrounds in the segmentation results. Finally, a context-aware reassembly decoder, together with a boundary-aware gating mechanism, progressively recovers building boundaries and local contour details and thus improves boundary completeness in the predictions. Experiments on the WHU and Inria building datasets show that the proposed method achieves IoU scores of 93.94% and 84.80% on the two datasets, respectively. On the WHU dataset, it reduces the 95th-percentile Hausdorff distance (HD_{95}) by nearly 20% compared with existing methods, demonstrating clear advantages in both region segmentation accuracy and boundary recovery quality; on the Inria dataset, it attains an HD_{95} of 18.720 4 pixels. The proposed method effectively improves building segmentation and boundary preservation in complex scenes.

Keywords: building extraction; high-resolution remote sensing imagery; boundary awareness; dynamic prototype; deep learning

(责任编辑　李　琼)