

文章编号 1004-924X(2026)18-2907-15

融合多头自注意力与双重特征重标定的 高分辨率遥感建筑物提取

韩轶龙^{1,2,3}, 巩丹超^{4,5}, 郑 敏^{1*}, 侯嘉雯¹

(1. 山东科技大学 测绘与空间信息学院, 山东 青岛 266590;

2. 空间基准全国重点实验室, 陕西 西安 710054;

3. 中国测绘科学研究院, 北京 100036;

4. 智能空间信息国家级重点实验室, 北京 100081;

5. 西安测绘研究所, 陕西 西安 710054)

摘要:针对高分辨率遥感影像建筑物提取中目标尺度跨度大、顶层纹理复杂及背景环境干扰,导致传统模型提取大型建筑物时易产生内部空洞、小型目标丢失边界几何特征的问题,本文提出一种融合多尺度局部特征与全局空间表征的双流网络架构。构建以全对称 U-Net 为拓扑基础的混合网络:在编码阶段,设计层级式特征金字塔提取多分辨率局部纹理信息,并于空间分辨率为原始尺寸 1/16 的最深层嵌入 Vision Transformer 模块,利用多头自注意力机制完成长距离依赖建模;在解码阶段,引入自适应双重注意力重标定机制,通过级联的卷积块注意力模块(CBAM)从空间与通道维度对跳跃连接传递的特征进行噪声抑制,并融合交叉注意力模块以对齐多尺度特征;在网络末端构建分割与边界预测并行的多任务输出分支,利用边界损失函数进行联合优化。在 WHU 航空影像数据集中,本文模型的交并比(IoU)达到 89.48%, F1 分数达到 94.45%,精确率为 90.12%,召回率达到 99.21%;对比 U-Net 架构,交并比与 F1 分数分别提升 13.95% 与 14.05%。在背景复杂的 ISPRS Potsdam 数据集中,模型召回率稳定在 90.48%。消融实验表明,全局自注意力模块的引入使交并比提升了 6.67%,有效改善了目标内部特征丢失现象。该网络架构在可控的计算开销下实现了对多尺度建筑物空间拓扑与局部轮廓的有效提取,弥补了二维卷积模型的局部感受野局限,在城市空间演变监测与地理基础数据库更新等工程应用中具有实际部署价值。

关键词:建筑物提取;语义分割;多头自注意力机制;特征重标定;多任务学习

中图分类号: TP394.1; TH691.9 **文献标识码:** A

doi: 10.37188/OPE.20263418.2907 **CSTR:** 32169.14.OPE.20263418.2907

Building extraction from high-resolution remote sensing imagery fusing multi-head self-attention and dual feature recalibration

HAN Yilong^{1,2,3}, GONG Danchao^{4,5}, ZHENG Min^{1*}, HOU Jiawen¹

(1. College of Geodesy and Geomatics, Shandong University of Science and Technology,
Qingdao 266590, China;

2. State Key Laboratory of Spatial Datum, Xi'an 710054, China;

收稿日期: 2026-06-03; **修订日期:** 2026-08-03.

基金项目: 空间基准全国重点实验室开放基金(No. SKLSD2026-KF-10); 山东省自然科学基金青年基金(No. ZR2022QD141); 国家自然科学基金(No. 42571528, 42271273); 国家重点研发计划(No. 2022YFC3002704)

3. *Chinese Academy of Surveying and Mapping, Beijing 100036, China;*

4. *National Key Laboratory of Intelligent Spatial Information, Beijing 100081, China;*

5. *Xi'an Research Institute of Surveying and Mapping, Xi'an 710054, China)*

* *Corresponding author, E-mail: a924208181@163.com*

Abstract: Aiming at the challenges of large scale spans, complex roof textures, and background interference in high-resolution remote sensing imagery, which caused traditional models to suffer from internal holes when extracting large buildings and boundary geometric feature loss for small targets, a dual-stream network architecture integrating multi-scale local features and global spatial representations was proposed. A hybrid network based on the fully symmetric U-Net topology was constructed. In the encoding stage, a hierarchical feature pyramid was designed to extract multi-resolution local texture information, and a Vision Transformer module was embedded at the deepest layer (1/16 of the original spatial resolution) to accomplish long-range dependency modeling using a multi-head self-attention mechanism. In the decoding stage, an adaptive dual attention recalibration mechanism was introduced, utilizing a cascaded Convolutional Block Attention Module (CBAM) to suppress noise in features transferred via skip connections from both spatial and channel dimensions, while a cross-attention module was fused to align multi-scale features. At the network's end, a multi-task output branch with parallel segmentation and boundary prediction was constructed, which was jointly optimized by using a boundary loss function. On the WHU aerial imagery dataset, the proposed model achieves an intersection over union (IoU) of 89.48%, an F1 score of 94.45%, a precision of 90.12%, and a recall of 99.21%. Compared with the baseline U-Net architecture, the IoU and F1 score are improved by 13.95% and 14.05%, respectively. On the ISPRS Potsdam dataset with complex backgrounds, the model's recall rate remains stable at 90.48%. Ablation experiments demonstrate that the introduction of global self-attention module increases IoU by 6.67%, effectively mitigating internal feature loss of the targets. With controllable computational overhead, the proposed network architecture effectively extracts the spatial topology and local contours of multi-scale buildings, overcoming the limitation of the local receptive field in two-dimensional convolutional models. It exhibits significant practical deployment value in engineering applications such as urban spatial evolution monitoring and geographic basic database updating.

Key words: building extraction; semantic segmentation; multi-head self-attention mechanism; feature recalibration; multi-task learning

1 引言

在智慧城市建设与数字孪生底座构建的进程中,城市空间格局演变监测、灾害损毁评估以及地理信息系统(Geographic Information System, GIS)基础数据库的高频更新,均依赖于从高分辨率遥感影像中自动、精确地提取建筑物目标^[1-3]。随着对地观测技术的发展,高分辨率遥感影像提供了前所未有的几何与纹理细节^[4]。然而,由于同一屋顶在不同光照下反射不均引起的类内差异,以及周边行道树遮挡、高层建筑投影

等背景干扰^[5-7],高精度、自动化的建筑物精细提取算法,始终是遥感图像处理与测绘工程交叉领域亟待突破的难点^[8-9]。

目前,以全卷积网络(Fully Convolutional Network, FCN)^[10], U-Net^[11]以及DeepLab系列^[12]为代表的深度语义分割模型在遥感图像处理中得到了广泛应用。Chen等^[13]提出了改进的FCN框架SDFCNv2,通过混合卷积块与注意力机制有效提升了遥感图像语义分割性能。Özbay等^[14]进一步将区间Type-2模糊门控机制引入U-Net跳跃连接,通过估计像素级不确定性权重来

增强边界感知特征学习,有效缓解了遥感图像中混合像素与类间光谱相似性导致的分割碎片化问题。任凤雷等^[15]融合超像素分割算法对语义分割结果进行边缘优化,有效提升了分割精度,弥补了DeepLabV3+在编码阶段细节信息丢失导致边缘分割不佳的不足。然而,现有主流模型在应对复杂建筑群时存在固有的理论局限:卷积神经网络(Convolutional Neural Network, CNN)受限于局部感受野与空间平移等变性偏置,在处理大规模连片建筑或大型单体建筑时,难以有效建立跨越长距离的全局空间依赖^[16]。这种物理感受野的限制易导致特征在逐层传递中衰减,并在预测结果中表现为建筑物内部的响应缺失(即空洞现象)^[17]。此外,为扩大感受野而依赖深层池化下采样的机制,易破坏底层高频几何信息,导致小型建筑物漏检以及边界模糊^[18]。尽管近年来引入的纯Transformer架构通过全局自注意力机制解决了长程建模问题^[19],但其计算复杂度呈输入图像尺寸的二次方指数增长,难以满足工程应用中对实时性与算力开销的严苛要求^[20]。例如, Pan 等人利用生成对抗网络^[21]整合空间与通道注意力,虽增强了对不规则建筑的提取能力,却显著加剧模型训练的难度与计算资源的消耗^[22]。

为缓解局部细节感知与全局语义表征难以兼顾的工程应用瓶颈,本文基于现有先进网络模块,提出了一种融合CNN-Transformer与双重特征重标定的建筑物提取网络适配架构。研究在编码器最深层嵌入多头自注意力模块,旨在以较小的计算代价重构全局上下文依赖关系;在解码器特征恢复路径中引入双重空间-通道特征重标定机制,自适应抑制多尺度跳跃连接中混入的高频噪声。同时,通过融合边界联合监督的多任务学习策略,强化了模型对建筑物几何轮廓的特征表达。该方法主要聚焦于为复杂城市场景下的遥感高精度解译提供一套在算力与精度间具备较优平衡的深度学习实践方案。

2 遥感影像建筑物提取理论与方法

高分辨率遥感影像中的建筑物提取本质上属于像素级的语义分割任务。近年来,深度学习技术的演进为该领域提供了丰富的理论基础与

网络范式,主要可归纳为以下三大技术分支:

2.1 基于卷积神经网络的局部特征建模

在深度学习广泛应用之前,遥感影像建筑物提取主要依赖手工设计特征。随着全卷积网络(FCN)提出端到端的像素级密集预测范式,基于卷积神经网络(CNN)的方法成为主流^[23]。经典如U-Net构建了完全对称的编码器-解码器架构,通过跳跃连接有效缓解了深层网络中空间信息的不可逆丢失,在建筑物边缘提取上表现出较好的几何保真度。DeepLab系列则通过引入空洞空间金字塔池化(Atrous Spatial Pyramid Pooling, ASPP)模块,在不降低特征图分辨率的前提下显著扩大了感受野,增强了模型的多尺度目标建模能力^[24]。然而,CNN的底层机制依赖于局部感受野的滑动窗口计算,难以高效建立像素间的长距离依赖关系。在处理大尺度或呈连片分布的建筑物群时,仅依靠堆叠卷积层极易发生内部特征信息的衰减,从而导致目标内部出现空洞现象^[25]。

2.2 注意力机制的特征重标定

为优化CNN在复杂背景环境下的特征表达能力,注意力机制被广泛引入建筑物提取任务中,其本质是一种特征空间的自适应资源分配策略^[26-27]。空间注意力机制专注于定位图像中的关键几何区域,引导网络强化目标物理位置的特征响应;而通道注意力机制则侧重于利用全局池化与感知机评估各特征通道对特定语义的贡献度,从而动态抑制冗余或与复杂背景强相关的干扰通道。卷积块注意力模块(Convolutional Block Attention Module, CBAM)等先进架构将上述二者级联组合,实现了特征图在空间与通道双维度的深度重标定,有效增强了模型过滤植被、阴影等高频背景噪声的能力,显著提升了建筑物轮廓的精细化感知度^[28]。

2.3 基于Transformer的全局上下文建模

为彻底突破CNN物理感受野的理论边界,源于自然语言处理的Transformer架构被跨模态引入遥感图像处理领域^[29]。Vision Transformer (ViT)^[30]等模型通过将输入图像转化为固定的序列图像块(Patch),并利用多头自注意力机制(Multi-Head Self-Attention)并行计算序列中任意两个位置之间的长距离相关性,从根本上克服

了传统网络在全局依赖建模上的局限。在建筑物提取场景中,它能跨越障碍物(如行道树遮挡)直接连通建筑物的整体拓扑^[31]。然而,纯 Transformer 模型应用于高分辨率遥感影像时,其计算复杂度会随输入序列长度呈二次方爆炸式增长;且由于缺乏 CNN 固有的局部连接与权值共享偏置,在提取边缘走向、角点位置等底层高频几何细节时极易发生信息丢失,导致小型建筑物漏检及边界模糊^[32]。因此,将 CNN 的局部细节刻画优势与 Transformer 的全局表征能力相融合,构建混合式网络架构,已成为当前解决建筑物提取

痛点的高效研究范式。

3 融合多头自注意力与双重特征重标定的高分辨率遥感建筑物提取

图 1 为融合多头自注意力与双重特征重标定的高分辨率遥感建筑物提取模型整体架构图。本文模型整体沿用 U-Net 编码器-解码器基础拓扑,在此框架内进行深度结构优化,形成了前端局部特征金字塔、底层全局关系映射、后端多重标定解码的处理架构。

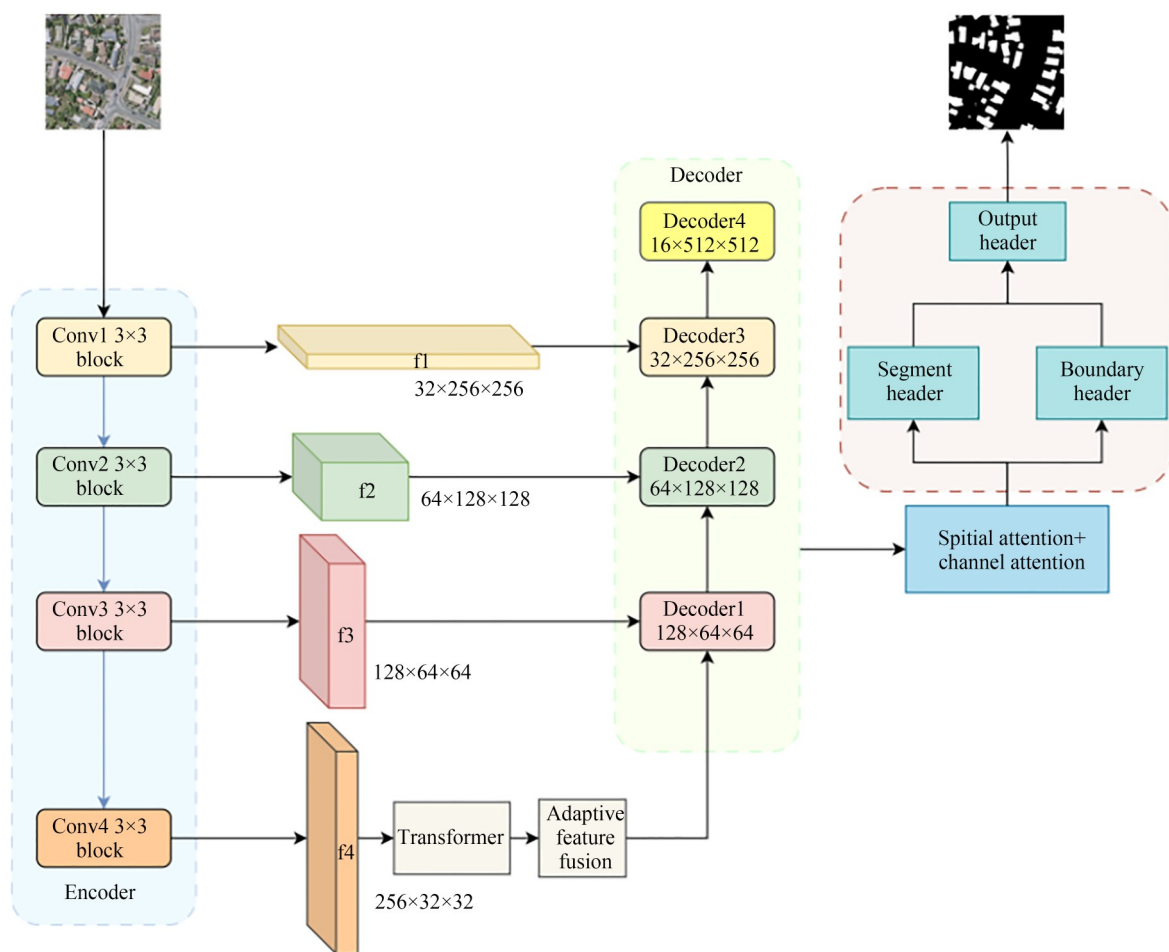


图 1 融合多头自注意力与双重特征重标定的高分辨率遥感建筑物提取模型整体架构图

Fig. 1 Overall architecture diagram of a high-resolution remote sensing building extraction model integrating multi-head self-attention and dual feature recalibration

3.1 编码器前馈与多尺度特征金字塔提取

编码器的主要功能是从地表背景中逐级抽象建筑物的多尺度特征。设输入的高分辨率光

学遥感影像为三维张量 $I \in R^{C \times H \times W}$, 其中空间分辨率 $H = W = 512$, 通道数 $C = 3$ 。编码器的前向传播切分为四个阶段, 表示为:

$$F_k = \phi_k(F_{k-1}), k \in \{1, 2, 3, 4\}, \quad (1)$$

其中: $F_0 = I$, 算子 ϕ_k 表示第 k 阶段的下采样卷积模块, 每个模块内部包含两组 Conv-BN-ReLU 基础计算单元与一个步长为 2 的最大池化层。针对输入特征图 X , 卷积前馈逻辑为:

$$Y = \text{ReLU}(\text{BN}(\text{Conv}(X))). \quad (2)$$

对于给定的二维卷积运算, 其数学定义展开为:

$$Y_{c,i,j} = \sum_{c'=1}^{C_{in}} \sum_{u=1}^K \sum_{v=1}^K W_{c,c',u,v} \times X_{c',i+u-1,j+v-1} + b_c, \quad (3)$$

其中: W 表示尺寸为 $C_{out} \times C_{in} \times K \times K$ 的权重滤波器, 卷积核边长 $K = 3$ 。为缓解深层网络可能出现的梯度弥散现象, 在第 3 与第 4 阶段引入了残差直连映射, 即输出 $y = F(x) + x$ 。通过四级下采样, 生成了包含丰富多尺度信息的特征金字塔序列。最深层特征张量 $F_4 \in R^{256 \times 32 \times 32}$ 的空间分辨率被压缩至原始尺寸的 1/16, 该低分辨率空间极大地降低了序列长度, 为后续的全局自注意力计算提供了适宜的算力平台。

表 1 和图 2 给出了编码器四级下采样过程中各阶段特征图的空间分辨率、通道数变化。浅层特征 (Stage 1, Stage 2) 保留了丰富的空间细节, 对建筑物边缘定位与小型目标敏感; 深层特征

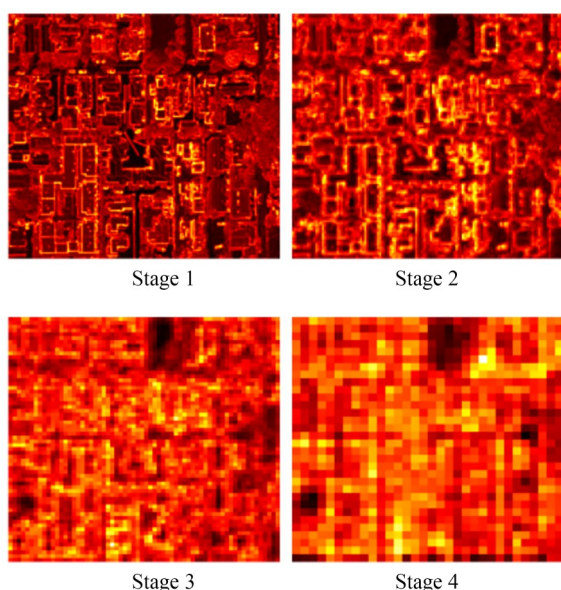


图 2 编码器各阶段特征图

Fig. 2 Visualization of feature maps at different encoder stages

表 1 编码器各阶段特征金字塔参数与语义特性

Tab. 1 Parameters and semantic characteristics of encoder feature pyramid at each stage

阶段	输出尺寸	输出通道数	提取信息类型
Stage 1	256×256	32	边缘、角点、纹理
Stage 2	128×128	64	局部形状、结构
Stage 3	64×64	128	建筑轮廓、部件
Stage 4	32×32	256	整体拓扑、语义类别

(Stage 3, Stage 4) 语义抽象程度高, 对大型建筑区域响应完整稳定, 但分辨率降低导致小型目标细节丢失。浅层细节与深层语义通过跳跃连接实现互补, 支撑多尺度建筑物的提取。

3.2 基于多头自注意力的全局上下文建模

在遥感物理场景中, 高分辨率光学影像受限于传感器的侧视成像几何与太阳高度角, 建筑物立面投影与伴生阴影常导致二维成像平面上同一建筑物的连续光谱特征发生物理断裂。传统 CNN 的局部马尔可夫假设依赖固定尺寸的感受野, 在遭遇强光照畸变或遮挡时, 其卷积核易陷入低反照率阴影区, 难以跨越此类光谱断层建立显式连接, 导致模型推断出内部响应缺失的破碎斑块。

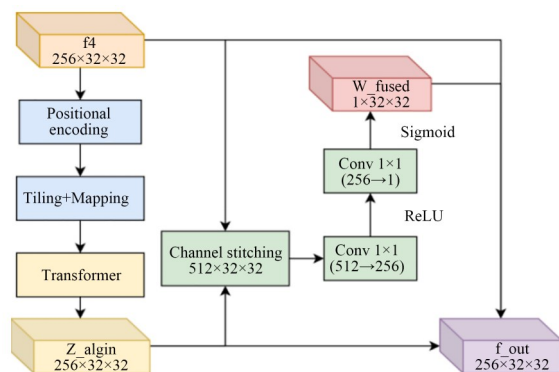


图 3 Transformer 特征融合原理图

Fig. 3 Schematic diagram of Transformer feature fusion

为缓解上述工程瓶颈, 本研究采用在编码器路径末端嵌入成熟 Vision Transformer (ViT) 模块的部署策略。多头自注意力 (MHSA) 机制通过计算全局亲和度矩阵, 从物理空间上建立无视空间距离的非局部特征关联。由于自注意力机制是置换不变的, 为感知二维像素格网的坐标信

息,对深层特征 F_4 注入显式的位置编码:

$$F_4' = F_4 + P_h + P_w, \quad (4)$$

其中: $P_h \in \mathbb{R}^{256 \times 32 \times 1}$ 和 $P_w \in \mathbb{R}^{256 \times 1 \times 32}$ 分别表示行与列方向上的可学习一维位置向量。随后,将特征图 F_4' 划分为互不重叠的 4×4 空间图块,展开并经线性映射至维度 $D = 384$ 的嵌入空间,生成序列表示 $Z \in \mathbb{R}^{N \times D}$, 序列长度 $N = 64$ 。

多头自注意力(MHSA)用于建立建筑实体间及建筑物与环境背景间的长程依赖。通过将输入序列投影至三个并行的子空间,获取查询(Q)、键(K)和值(V)矩阵:

$$Q = ZW_Q, \quad (5)$$

$$K = ZW_K, \quad (6)$$

$$V = ZW_V, \quad (7)$$

其中, $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ 为动态更新的权重矩阵。全局注意力的计算通过缩放点积实现:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V. \quad (8)$$

除以缩放因子 $\sqrt{D}$ 旨在防止内积数值过大导致 softmax 函数梯度饱和。在经过 $L = 4$ 层的 Transformer 块(内含多头注意力与前馈网络)推

断后,输出序列 $Z^{(l)}$ 充分捕获了地物的全局拓扑规律。

为将序列化特征平滑回注至卷积主干,网络采用双线性插值对重塑后的三维张量进行上采样,并通过 1×1 卷积进行通道对齐,得到 $Z_{align} \in \mathbb{R}^{256 \times 32 \times 32}$ 。为实现局部细粒度特征与全局特征的自适应融合,设计了一个动态门控权重矩阵:

$$W_f = \sigma(Conv_{1 \times 1}(ReLU(Conv_{1 \times 1}([F_4; Z_{align}]))))). \quad (9)$$

最终融合输出以 W_f 为加权系数,在空间坐标系下自适应调配 CNN 特征与 Transformer 特征的响应强度。

3.3 跨层特征重标定与解码器重建

解码阶段旨在将融合了全局先验的抽象语义表征,逐步上采样恢复至高分辨率。针对同级跨层跳跃连接易将浅层非目标高频噪声(如阴影、植被纹理)直接引入解码路径的固有缺陷,本文在特征拼接前串行接入了双重注意力重标定模块(CBAM)。

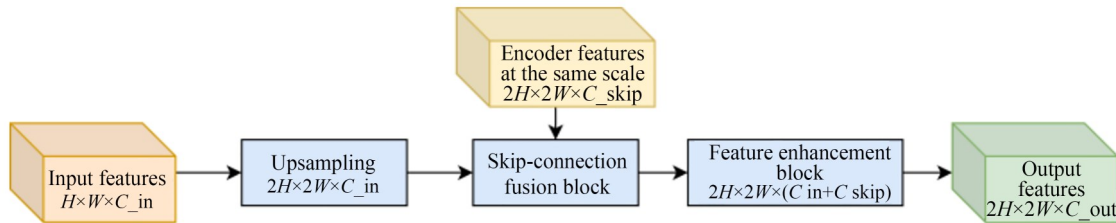


图4 单解码器块结构图

Fig. 4 Single decoder block architecture

首先,空间注意力子模块通过对通道域执行最大池化与平均池化聚合空间分布统计量,生成空间权重掩膜并重新赋权至原始跳跃特征 X_{skip} :

$$X_{spatial} = X_{skip} \odot SpatialAttention(X_{skip}). \quad (10)$$

随后,通道注意力子模块利用全局池化与多层感知机(MLP)评估各特征通道的鉴别贡献度,动态压制冗余特征:

$$X_{enhanced} = X_{spatial} \odot ChannelAttention(X_{spatial}). \quad (11)$$

这种级联重标定有效滤除了注入解码器的特征流噪声。在上采样特征与 $X_{enhanced}$ 拼接后,为增强二者的一致性,在解码块末端叠加了交叉注

意力机制。设当前解码特征为 Q , 编码特征为 K 和 V , 通过注意力寻址后与原始特征进行残差连接:

$$Q_{out} = Q + \tau \cdot Conv_{1 \times 1}(Attention(Q, K, V)), \quad (12)$$

其中, τ 为可学习的自适应标量参数,确保模型在训练初期维持梯度稳定,逐步过渡至跨层特征深度交互阶段。

3.4 边界联合优化的多任务监督头

针对建筑物边缘在预测时易出现模糊与扩张的现象,模型在最高分辨率解码层($D_1 \in \mathbb{R}^{16 \times 512 \times 512}$)后设置为双重预测头:主分支

输出像素级类别概率,辅助分支输出边界轮廓概率。分割主干采用加权二元交叉熵损失(BCE)与Dice损失协同驱动,以缓解背景与建筑物像素分布极度不平衡的问题:

$$L_{\text{bce}} = - \frac{\sum_{i=1}^N [\mathbf{w} \cdot g_i \log(p_i) + (1 - g_i) \log(1 - p_i)]}{N}, \quad (13)$$

$$L_{\text{dice}} = 1 - \frac{2 \sum_i p_i g_i}{\sum_i p_i + \sum_i g_i + c}, \quad (14)$$

其中: $\mathbf{w}$ 为常数权重(设为20以强化建筑物类别响应), c 为防止除零的平滑项。BCE损失的偏导数与预测概率成反比:当网络对某个背景像素给出高置信度的误判时,BCE能够对其施加极强的惩罚梯度,确保像素级分类的局部稳定性。然而,BCE对类别极度不平衡十分敏感。相反,Dice损失的偏导数是对全局区域响应的直接优化,其梯度幅度不受背景像素总量的支配。两者联合使用,在梯度优化空间中形成互补,兼顾微观像素的强制收敛与宏观轮廓的几何对齐。边界辅助分支采用无加权标准交叉熵损失 L_{boundary} ,促使顶层网络学习边缘梯度特征。总体目标函数定义为:

$$L_{\text{total}} = L_{\text{bce}} + L_{\text{dice}} + \lambda \cdot L_{\text{boundary}}, \quad (15)$$

其中, λ 为边界损失权重系数。为确定 λ 的最优取值,本文在Potsdam验证集上进行了网格搜索($\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$),结果如表3所示。当 $\lambda=0.7$ 时模型取得最优综合性能,因此本文采用 $\lambda=0.7$ 作为边界损失权重,在提供拓扑边缘感知力的同时,避免了辅助任务对分割主任务的梯度侵占。

4 实验与分析

4.1 实验环境与评价指标

实验均在搭载 Intel i7-14650HX CPU 与 NVIDIA GeForce RTX 4060 GPU 的平台上进行,算法基于 PyTorch 1.13.0 开发。模型初始学习率设定为 5×10^{-5} ,采用 AdamW 优化器并结合 ReduceLROnPlateau 动态调度策略,权重衰减系数为 1×10^{-4} ,批处理量(Batch Size)设定为8,共执行80个训练轮次。以验证集损失为监控指标,若连续15个轮次内验证损失未出现下降,则终止

训练并恢复至验证损失最低点的模型权重。实际训练中模型约在第60~70轮触发早停。

数据预处理阶段,输入图像被统一裁剪为 512×512 pixels 的标准切片,在输入网络前进行归一化处理。训练阶段在线施加了以下数据增强策略:随机水平翻转与垂直翻转(概率各为 $p=0.5$);随机旋转,旋转角在 $\{90^\circ, 180^\circ, 270^\circ\}$ 、中等概率选取,施加概率 $p=0.75$;随机缩放与裁剪,缩放系数在 $[0.8, 1.2]$ 内均匀采样,施加概率 $p=0.5$;色彩抖动,亮度、对比度、饱和度调整幅度均为 $\pm 20\%$,施加概率 $p=0.5$;高斯噪声采用标准差 $\sigma=0.01$,施加概率 $p=0.25$ 。

评价指标选取语义分割领域通用的四个定量标准进行综合评估:精确率(Precision)、召回率(Recall)、F1分数(Dice系数)及交并比(IoU)。其中,精确率衡量模型抗误检控制能力,召回率反映对真实目标的覆盖完整性,交并比则量化预测结果与真实标签的全局几何重叠度。

设 X 为预测结果, Y 为真实标注。

F1分数(Dice系数)的计算为:

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|}, \quad (16)$$

其中: $2|X \cap Y|$ 为交集像素数, $|X|$ 和 $|Y|$ 分别为预测和真实区域的像素总数。

交并比的计算为:

$$\text{交并比} = \frac{|X \cap Y|}{|X \cup Y|}. \quad (17)$$

精确率的计算为:

$$\text{精确率} = \frac{TP}{TP + FP}, \quad (18)$$

其中:TP(True Positive)为真正例,模型正确预测为正类的样本数量。FP(False Positive)为假正例,模型错误预测为正类的样本数量

召回率的计算为:

$$\text{召回率} = \frac{TN}{TN + FN}, \quad (19)$$

其中:TP代表真正例,即正确预测为正类的样本数量,FN代表假负例,即实际为正类但被预测为负类的样本数量。

4.2 数据集划分与对比实验

为全面验证算法在不同空间分辨率与城市场景下的泛化性能,选取 WHU 航空影像(包含 4 736 张训练集、1 036 张验证集和 2 416 张测试

集)、ISPRS Potsdam 极高分辨率城区数据集(4 000 张训练集、712 张验证集和 1 301 张测试集)和 DFC2019 数据融合竞赛库(4 422 张训练集、1 189 张验证集和 1 404 张测试集)三个公开基准数据集进行评估。WHU 数据集源自新西兰基督城 0.3 m 分辨率航空影像,覆盖面积约 450 km²; ISPRS Potsdam 数据集为 5 cm 极高分辨率航空正射影像,涵盖德国波茨坦复杂城市场景,原始影像尺寸 6 000×6 000 pixels; DFC2019 数据集来自 IEEE GRSS 数据融合竞赛,包含美国杰克逊维尔和奥马哈两个城区,WorldView-3 卫星影像全色波段地面采样距离 0.35 m。三个数据集均采用 RGB 三通道影像,统一裁剪为 512×512 pixels 标准切片。正负样本分布方面,WHU 测试集中建筑物像素占比为 11.13%(背景 88.87%), ISPRS Potsdam 为 29.40%(背景 70.60%), DFC2019 为 29.23%(背景 70.77%), 均呈现典型的类别不平衡特征。

表 2 不同模型在 WHU 测试集上的精度与效率对比
Tab. 2 Comparison of accuracy and efficiency of different models on the WHU test dataset

方法	参数量	FLOPs	FPS	IoU
U-Net	4.37	31.02	104.0	75.53
TransUNet	89.29	120.10	25.5	77.78
Swin-UNet	30.55	32.85	71.4	82.64
纯 CNN	3.62	35.35	44.80	79.54
本文模型	12.63	36.21	46.15	89.48

为进一步评估模型的计算复杂度与推理效率,本文统计了不同方法在相同硬件环境下的参数量、FLOPs 及 FPS(表 2)。U-Net 结构简单,推理速度最快(104.0 FPS),但 IoU 仅 75.53%; TransUNet 引入完整 Transformer 后参数量激增至 89.29 M,推理速度降至 25.5 FPS; Swin-UNet 通过窗口注意力在 32.85 G FLOPs 下达到 71.4 FPS 和 82.64% IoU。纯 CNN(移除 Transformer)参数量仅 3.62 M,推理速度 44.80 FPS,但 IoU 为 79.54%。本文模型在此基础上引入低分辨率 ViT, FLOPs 仅增加 0.86 G,推理速度基本持平(46.15 FPS),而 IoU 提升至 89.48%,验证了以极小计算代价获取显著精度增益的设计目标,实现了精度与效率的较优平衡。

在上述数据集上,将本文模型与 U-Net, FCN, DeepLabV3+, YOLOv13-seg 等网络开展严格的对比测试,从交并比、F1 分数、精确率与召回率四个维度综合评估各模型的建筑物提取性能。

4.2.1 WHU 数据集性能验证

在 WHU 数据集中,样本类间分布存在极端的失衡现象,这对模型抗过分割能力提出了较高要求。

表 3 WHU 对比实验定量结果
Tab. 3 Quantitative comparison results on WHU

方法	Dice	交并比	精确率	召回率
U-Net	0.804 0	0.755 3	0.756 2	0.997 1
DeepLabV3	0.911 9	0.843 0	0.869 8	0.964 5
FCN	0.920 2	0.852 2	0.881 4	0.962 6
YOLOv13-seg	0.902 6	0.829 1	0.879 3	0.963 3
本文模型	0.944 5	0.894 8	0.901 2	0.992 1

在 WHU 数据集中,样本类间分布存在极端的失衡现象,这对模型抗过分割能力提出了较高要求如表 3 所示,基础 U-Net 精确率仅为 75.62%,由于缺乏全局约束极易将背景错分为建筑物。本文模型通过引入自适应特征重标定模块过滤背景噪声,使精确率提升 14.50%,交并比提升 13.95%。与 YOLOv13-seg 相比,本文模型交并比提升 6.57%;与依赖空洞卷积扩大感受野的 DeepLabV3+ 相比,本文模型在交并比上仍有 5.18% 的提升,证明了全局自注意力机制在保障建筑物空间拓扑完整性方面具有显著优势。

4.2.2 ISPRS Potsdam 场景验证

ISPRS Potsdam 数据集场景复杂,欧式建筑密度高且频繁受树木遮挡,主要检验模型的边界

表 4 ISPRS Potsdam 对比实验定量结果
Tab. 4 Quantitative comparison results on ISPRS Potsdam

方法	Dice	交并比	精确率	召回率
U-Net	0.837 6	0.780 1	0.836 0	0.879 6
DeepLabV3	0.841 9	0.776 3	0.843 1	0.882 1
FCN	0.840 5	0.771 4	0.845 6	0.873 3
YOLOv13-seg	0.864 1	0.806 4	0.850 3	0.908 9
本文模型	0.871 1	0.814 2	0.873 3	0.904 8

刻画与抗背景干扰能力。

在该组测试中,三个对比基准模型的 IoU 分别为 78.01%,77.63% 和 77.14%,均在 78% 以下,反映出密集场景下传统卷积模型难以有效区分密集排列的建筑个体。本文模型通过引入 Transformer 模块构建跨层空间依赖,IoU 提升至 81.42%。与 YOLOv13-seg 相比,本文模型的 IoU 提升 0.78%,精确率提升 2.30%,验证了本文模型在密集城市场景下对建筑群边界刻画与精细提取方面的优势。

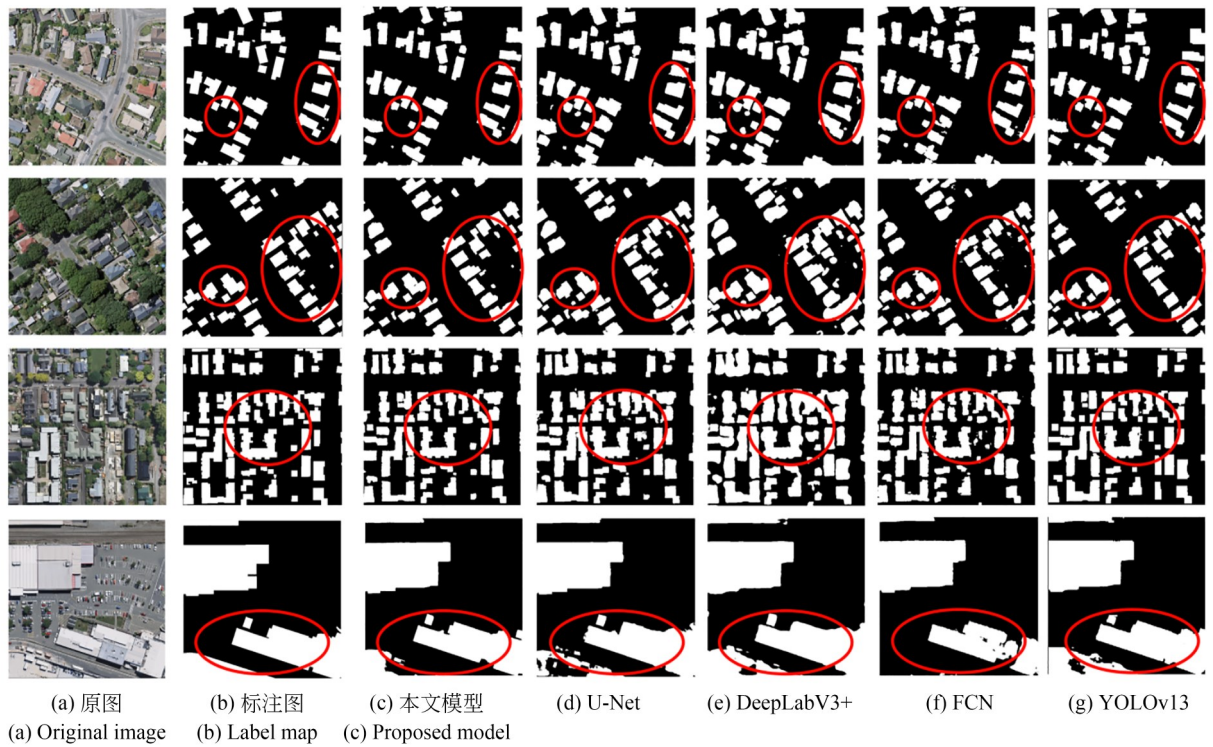


图 5 WHU 对比试验可视化结果图

Fig. 5 Visualized comparison results on the WHU dataset

4.3 误差机理与消融实验探讨

为解构模型性能跃升的驱动因素,研究通过剥离核心组件开展消融实验,并从特征提取机制角度深度剖析了典型误差来源。

表 5 的消融结果表明,边界损失权重 $\lambda=0.7$ 时模型在 Potsdam 数据集上取得最优综合性能 (IoU=81.42%,Dice=87.11%),后续实验均采用该权重设置。

表 5 ISPRS Potsdam 损失函数权重消融对比

Tab. 5 Ablation study on the loss weight for ISPRS potsdam datasett

损失函数权重 (以 ISPRS Potsdam 数据集为基准)	Dice	交并比	精确率	召回率
0.3	0.866 4	0.803 9	0.842 8	0.928 5
0.4	0.868 6	0.808 0	0.851 6	0.923 3
0.5	0.870 1	0.811 0	0.858 5	0.918 1
0.6	0.870 9	0.813 0	0.865 8	0.914 8
0.7	0.871 1	0.814 2	0.873 3	0.904 8
0.8	0.870 0	0.813 7	0.881 4	0.893 4
0.9	0.865 7	0.808 6	0.890 8	0.874 0

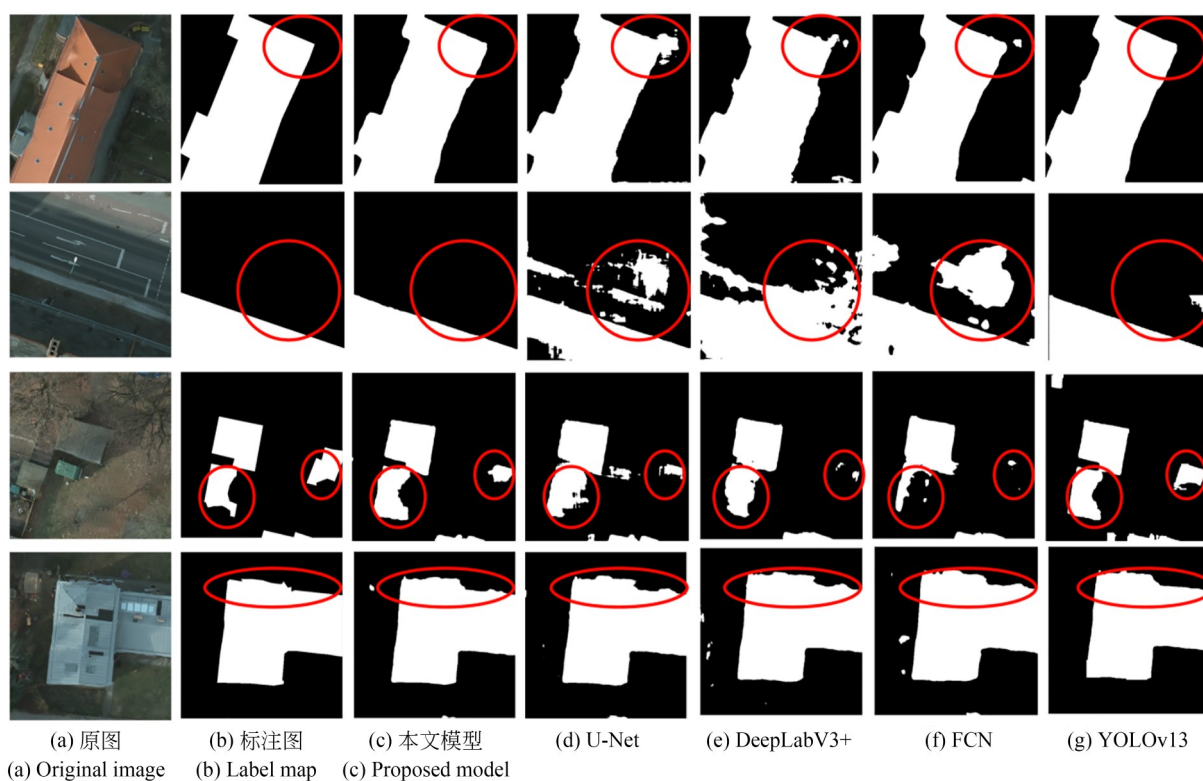


图 6 ISPRS Potsdam 对比试验可视化结果图

Fig. 6 Visualized comparison results on the ISPRS Potsdam dataset

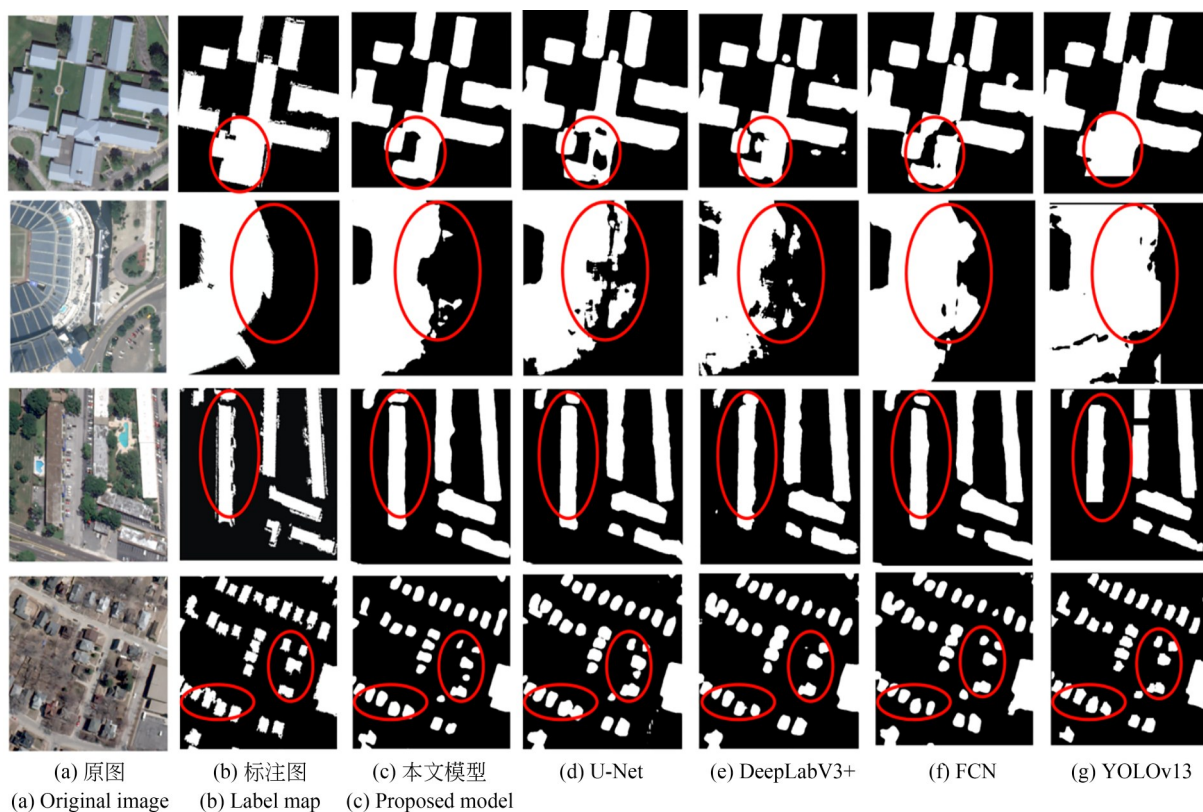


图 7 DFC2019 对比试验可视化结果图

Fig. 7 Visualized comparison results on the DFC2019 dataset

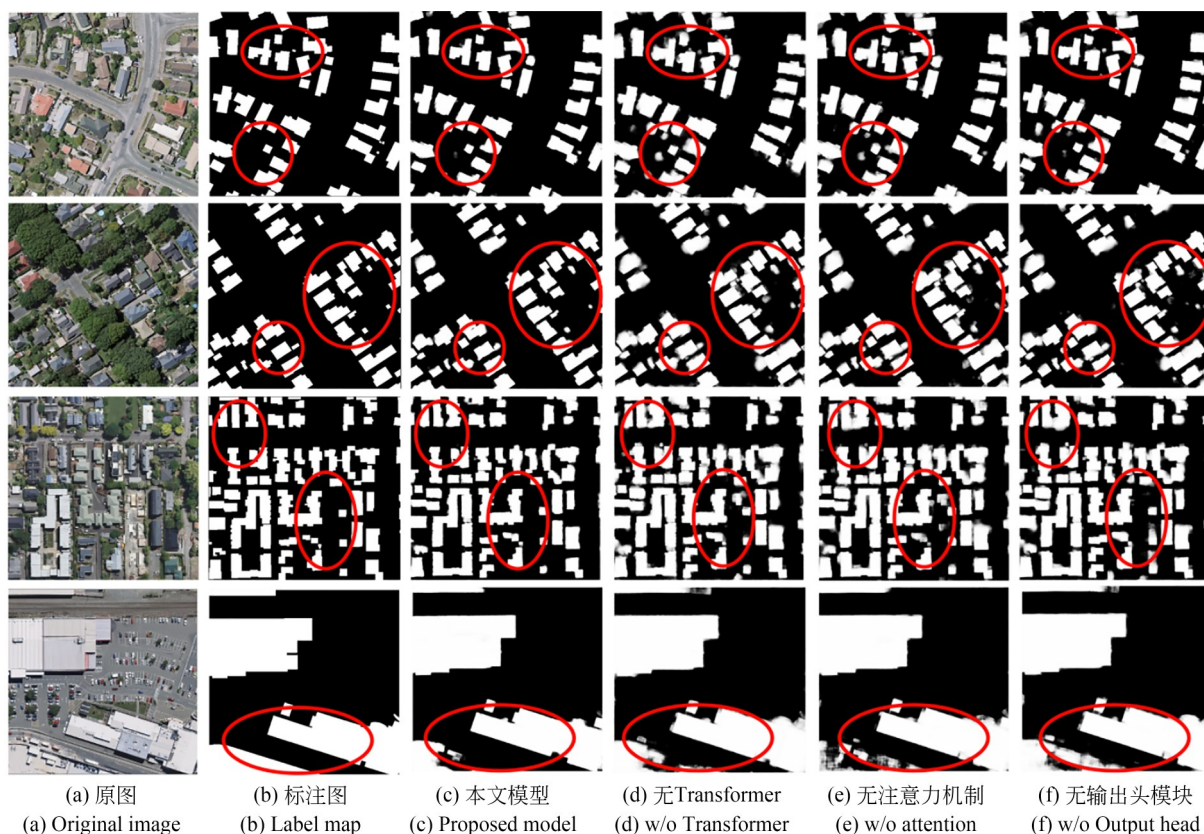


图 8 WHU 消融实验热力特征可视化对比图

Fig. 8 Comparison of feature heatmap visualizations for ablation experiments on the WHU dataset

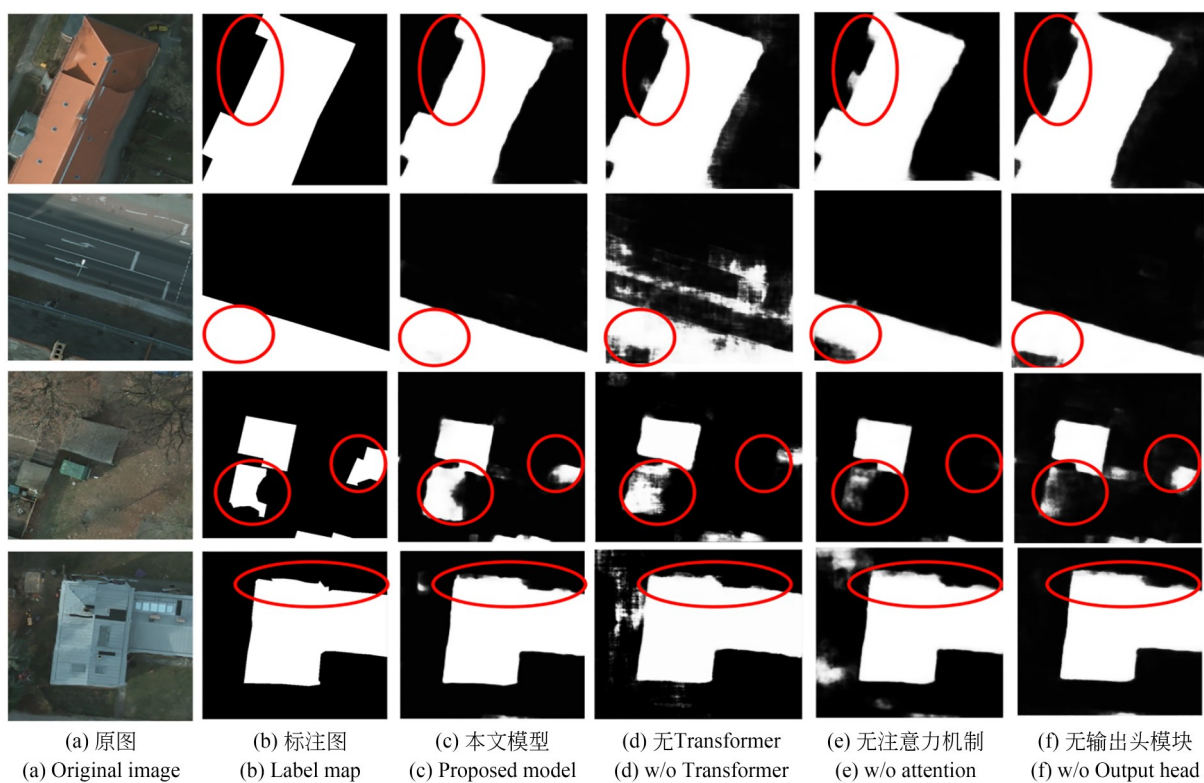


图 9 ISPRS Potsdam 消融实验热力特征可视化对比图

Fig. 9 Comparison of feature heatmap visualizations for ablation experiments on the ISPRS Potsdam dataset

表 6 WHU 消融实验定量结果
Tab. 6 Ablation experiments on the WHU dataset

模块设计状态(以 WHU 数据集为基准)	Dice	交并比	精确率	召回率
移除全局 Transformer 映射机制	0.906 0	0.828 1	0.844 8	0.976 7
移除双重注意力重标定(CBAM)	0.893 5	0.807 4	0.818 3	0.983 9
移除空间注意力	0.886 3	0.795 8	0.814 3	0.972 1
移除通道注意力	0.888 8	0.799 8	0.815 1	0.979 1
移除多任务边界感知输出头	0.902 6	0.822 6	0.831 7	0.986 8
原生完整融合模型	0.944 5	0.894 8	0.901 2	0.992 1

表 7 不同解码层级跳跃连接特征的信噪比(dB)对比
Tab. 7 Comparison of signal-to-noise ratio (SNR) of skip connection features at different decoder levels

解码 层级	特征 分辨率	CBAM 前 SNR	CBAM 后 SNR	增益
Skip1	256×256	2.53	2.44	−3.4%
Skip2	128×128	1.50	1.61	+7.8%
Skip3	64×64	1.54	1.81	+17.4%

(1)局部感受野受限引发的中心区域断裂机理:

消融数据明确揭示,当剔除 Transformer 模块后,交并比(IoU)下降了 6.67%。从网络层级分析,纯 CNN 网络受限于最大感知半径,在面临大型建筑平面时,极易依据局部单一平滑纹理做出孤立推断,从而导致目标中心区域预测响应下降。Transformer 算子通过连通结构边缘与中心特征,利用全连通空间注意力修补了此类缺陷。

(2)多尺度特征混叠诱发的误检机理:

移除双重注意力重标定模块(CBAM)后,模型精确率由 90.12% 下降至 81.83%,而召回率仅由 99.21% 降低至 98.39%,说明该模块主要作用于抑制背景误激活,而非提升目标覆盖能力。进一步的分量消融实验表明,移除空间注意力和通道注意力后,IoU 分别下降至 79.58% 和 79.98%,验证了二者在特征筛选中的互补作用。由于 CBAM 属于空间注意力和通道注意力的级联组合结构,完整移除 CBAM 会同时去除两种互补机制,而单独移除某一支时,另一支仍能够提供部分特征校正能力。因此,单分支消融结果与整体 CBAM 移除结果并非简单线性关系,其差异反映了两类注意力之间的协同调节作用。其中,空间注意力通过增强建筑物区域响

应、削弱非目标位置激活,提高了模型对空间分布差异的感知能力;通道注意力通过调整不同语义通道权重,降低了植被、道路及阴影等冗余纹理特征的干扰。因此,缺少 CBAM 时,跳跃连接传递的浅层特征中包含的光谱相似背景信息更易参与预测,导致假阳性增加,最终表现为精确率下降。

为进一步验证双重注意力重标定对跳跃连接特征噪声抑制的作用,本文对 CBAM 处理前后的多尺度跳跃特征进行信噪比(SNR)分析。结果表明,CBAM 对不同语义层级特征具有差异化调节作用。对于包含较强低级纹理和边缘信息的浅层 Skip1 特征,由于目标细节与背景噪声高度耦合,简单抑制冗余响应可能导致局部信噪比轻微下降;而随着网络深度增加,Skip2 和 Skip3 中的语义信息逐渐增强,CBAM 能够更有效强化建筑相关响应,使信噪比分别提升 7.8% 和 17.4%。这说明随着网络深度增加,CBAM 能够更有效地增强具有建筑语义的特征响应,并抑制背景纹理干扰。需要指出的是,Skip1 层由于包含大量边缘、纹理等低级视觉信息,背景噪声与目标细节高度混合,经过注意力重标定后信噪比出现轻微下降(2.53 dB 降至 2.44 dB),但该层信息仍通过后续特征融合参与建筑边界恢复,因此未造成整体性能下降。

(3)三维场景下的二维投影局限分析:

在空间分辨率较低的 DFC2019 卫星数据集中,本模型最高 IoU 回落至 70.67%。该现象暴露了二维深度网络在处理大尺度复杂真实世界场景时的物理短板。卫星的侧视成像物理特性使得大量建筑物立面与阴影在二维像素空间产生极度扭曲纠缠。在缺乏数字地表模型(DSM)等高程维度先验的情况下,仅依靠 RGB 可见光

的平面二维投影特征难以彻底消除类别间的概率混叠,这一深度误差溯源为后续立体多模态融合研究指明了方向。

5 结 论

本研究针对传统语义分割模型在宏观拓扑建模与微观几何特征保留上的局限,提出了一种严密融合 CNN 与 Transformer 的双流建筑物提取架构。经过多源高分辨率影像数据集的系统对比验证,提炼出以下主要结论:

(1)确立了宏微观异构特征耦合的空洞应对机制:验证了在极低分辨率特征空间部署全局自注意力模块的有效性。该机制成功突破了单一卷积网络感受野受限导致的大型建筑物内部空洞问题,在 WHU 数据集上的测试表明该设计实现了目标高召回率(99.21%)与计算效率的有效平衡。

(2)验证了双重特征重标定对复杂背景的抑噪作用:级联的空间与通道注意力阀门结合边界辅助优化策略,有效剥离了建筑物阴影与伴生植

被的混杂干扰。在建筑高度致密的 Potsdam 场景下,该机制显著降低了伪正类错判率,大幅改善了建筑物拓扑边界的平滑度与提取精确性。

(3)局限性与未来展望:尽管本文模型在 WHU 和 Potsdam 数据集上取得了较好的提取效果,但仍存在以下局限性:模型对高分辨率航空影像表现优异,但在卫星影像上性能下降明显(DFC2019 IoU 为 70.67%),说明其对影像分辨率和成像条件的敏感性较高;模型参数量为 12.63 M,推理速度为 46.15 FPS,相较于轻量级模型仍有优化空间;当前仅依赖 RGB 影像,未来拟通过融入数字地表模型(DSM)和 LiDAR 高程先验,并结合深度网络剪枝技术,进一步拓展该方案的应用。

作者贡献声明:

韩铁龙:论文构思,项目管理,初稿写作;

巩丹超:提供资源与指导;

郑敏:方法论,数据整理与分析,可视化呈现,审核与编辑写作;

侯嘉雯:调查研究,数据管理。

参考文献:

- [1] 李国燕,武海苗,董春华,等. 面向遥感建筑物提取的轻型多尺度差异网络[J]. 光学精密工程, 2023, 31(22): 3371-3382.
- [2] WANG H F, HE W, LI Z H, *et al.* Light-weight multi-scale difference network for remote sensing building extraction [J]. *Opt. Precision Eng.*, 2023, 31(22): 3371-3382. (in Chinese)
- [3] LI Q Y, MOU L C, SHI Y L, *et al.* BANet: a bilateral attention network for extracting changed buildings between remote sensing imagery and cadastral maps [J]. *International Journal of Applied Earth Observation and Geoinformation*, 2025, 139: 104486.
- [4] YE P, LIU G W, HUANG Y. Geographic scene understanding of high-spatial-resolution remote sensing images: methodological trends and current challenges[J]. *Applied Sciences*, 2022, 12(12): 6000.
- [5] 杨坚华,张浩,花海洋. 并行路径与强注意力机制遥感图像建筑物分割[J]. 光学精密工程, 2023, 31(2): 234-245.
- [6] YANG J H, ZHANG H, HUA H Y. Parallel path and strong attention mechanism for building segmentation in remote sensing images[J]. *Opt. Precision Eng.*, 2023, 31(2): 234-245. (in Chinese)
- [7] WU B, MENG X X, WANG H, *et al.* GASNet: geometric robust adaptive spatial-enhanced network for building extraction [C]. *Advanced Intelligent Computing Technology and Applications*. Singapore: Springer, 2025: 437-449.
- [8] LUO H, FENG X B, DU B, *et al.* A multimodal feature fusion network for building extraction with very high-resolution remote sensing image and LiDAR data [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 5621819.
- [9] YUAN Y, SHI X F, GAO J Y. Building extraction from remote sensing images with deep learning: a survey on vision techniques[J]. *Computer Vision and Image Understanding*, 2025, 251: 104253.

- [9] 王舒洋, 慕晓冬, 杨东方, 等. 融合高阶信息的遥感影像建筑物自动提取[J]. 光学精密工程, 2019, 27(11): 2474-2483.
WANG S Y, MU X D, YANG D F, *et al.* High-order statistics integration method for automatic building extraction of remote sensing images [J]. *Opt. Precision Eng.*, 2019, 27(11): 2474-2483. (in Chinese)
- [10] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 640-651.
- [11] RONNEBERGER O, FISCHER P, BROX T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* [M]. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.
- [12] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al.* Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs [EB/OL]. 2014: *arXiv*: 1412.7062. <https://arxiv.org/abs/1412.7062>
- [13] CHEN G Z, TAN X L, GUO B B, *et al.* SDFC-Nv2: an improved FCN framework for remote sensing images semantic segmentation [J]. *Remote Sensing*, 2021, 13(23): 4902.
- [14] ÖZBAY F, ÇETIN M. Interval Type-2 Fuzzy uncertainty modeling for U-Net based remote sensing semantic segmentation [J]. *Journal of Computational and Applied Mathematics*, 2026, 487: 117751.
- [15] 任风雷, 何昕, 魏仲慧, 等. 基于DeepLabV³⁺与超像素优化的语义分割[J]. 光学精密工程, 2019, 27(12): 2722-2729.
REN F L, HE X, WEI Z H, *et al.* Semantic segmentation based on DeepLabV³⁺ and superpixel optimization [J]. *Opt. Precision Eng.*, 2019, 27(12): 2722-2729. (in Chinese)
- [16] ZHANG Y N, YANG L, ZHU L C, *et al.* A deep supervised change detection network based on context-rich information [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025, 18: 21564-21577.
- [17] CHANG J H, CEN Y F, CEN G. Asymmetric network combining CNN and transformer for building extraction from remote sensing images [J]. *Sensors*, 2024, 24(19): 6198.
- [18] 李智杰, 惠爱婷, 李昌华, 等. 面向遥感图像道路提取的多尺度上下文感知网络[J]. 光学精密工程, 2025, 33(4): 610-623.
LI Z J, HUI A T, LI C H, *et al.* Multi-scale context-aware network for road extraction in remote sensing images [J]. *Opt. Precision Eng.*, 2025, 33(4): 610-623. (in Chinese)
- [19] YU M X, LI J P, HE W. Multi-scale guided context-aware transformer for remote sensing building extraction [J]. *Sensors*, 2025, 25(17): 5356.
- [20] YE H, CAI J J, DENG J T, *et al.* Efficient machine learning-based semantic segmentation algorithm for consumer-grade AAV remote sensing [J]. *IEEE Transactions on Consumer Electronics*, 2025, 71(4): 10812-10825.
- [21] ZHU L, CHEN Y S, GHAMISI P, *et al.* Generative adversarial networks for hyperspectral image classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, 56(9): 5046-5063.
- [22] PAN X R, YANG F, GAO L R, *et al.* Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms [J]. *Remote Sensing*, 2019, 11(8): 917.
- [23] 杨明旺, 赵丽科, 叶林峰, 等. 基于卷积神经网络的遥感影像建筑物提取方法综述[J]. 地球信息科学学报, 2024, 26(6): 1500-1516.
YANG M W, ZHAO L K, YE L F, *et al.* A review of convolutional neural networks related methods for building extraction from remote sensing images [J]. *Journal of Geo-Information Science*, 2024, 26(6): 1500-1516. (in Chinese)
- [24] 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.
- [25] 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.
- [26] LEI Y J, ZHOU W J, CHAN S X, *et al.* UNet reimagined: advanced building extraction through dual attention and global-local synthesis [J]. *Systems Science & Control Engineering*, 2025, 13:

- 2503238.
- [27] 陶从辉,高青山,赵梦琳.改进U-Net网络的高分辨率遥感影像建筑物提取方法[J].地理空间信息,2025,23(1):9-12.
- TAO C H, GAO Q S, ZHAO M L. Building extraction method in high-resolution remote sensing images based on improved U-Net network [J]. *Geospatial Information*, 2025, 23(1): 9-12. (in Chinese)
- [28] HE S, JIANG W S. Boundary-assisted learning for building extraction from optical remote sensing imagery[J]. *Remote Sensing*, 2021, 13(4): 760.
- [29] 李健康,张桂欣,祝善友,等.融合多尺度特征Transformer的高分辨率遥感图像变化检测[J].遥感学报,2025,29(1):266-278.
- LI J K, ZHANG G X, ZHU S Y, *et al.* Change detection for high-resolution remote sensing images with multi-scale feature transformer[J]. *National Remote Sensing Bulletin*, 2025, 29(1): 266-278. (in Chinese)
- [30] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. 2020; *arXiv*: 2010.11929. <https://arxiv.org/abs/2010.11929>
- [31] ZHOU K, LIU C, ZHANG Y N, *et al.* Enhanced building extraction via STMC-UNet: integrating super token transformer and multi-scale convolution [J]. *Signal, Image and Video Processing*, 2025, 19(12): 1019.
- [32] YANG F, JIANG F L, LI J Z, *et al.* MSTrans: multi-scale transformer for building extraction from HR remote sensing images[J]. *Electronics*, 2024, 13(23): 4610.

作者简介:



韩轶龙(1988—),男,河北深泽人,博士,讲师,硕士研究生导师,2020年毕业于武汉大学,获摄影测量与遥感专业博士学位,主要从事空天信息与人工智能等方面的研究。E-mail: hanyl@sdust.edu.cn

通讯作者:



郑敏(2004—),女,山东济南人,本科生,2022年于山东科技大学开始攻读遥感科学与技术专业学士学位,主要从事航空航天测绘遥感与计算机视觉研究。E-mail: a924208181@163.com