

研究论文

结合边缘增强的全局自注意力遥感建筑物提取网络

李振¹, 张振鑫¹, 王涛², 彭雪丽³, 岳贵杰⁴, 张德宇², 刘先林^{2,5*}, 李建华⁶

摘要 设计了一种结合边缘增强的全局自注意力深度学习网络(global self-attention network with edge-enhancement, E-GSANet), 用于遥感影像建筑物提取, 在编码主干构建并融入边缘增强模块, 为网络赋予边界先验知识信息; 通过全局自注意力特征表达模块构建影像的长距离依赖关系, 实现显著特征与边缘增强特征的表达融合; 使用逐级上采样解码模块, 将空间细节信息丰富的浅层特征与具有高阶语义信息的深层特征相融合, 得到建筑物的精确提取结果。基于 2 个开源的遥感建筑物数据集, 将 E-GSANet 与当前主流方法进行对比, 定量和定性分析表明, E-GSANet 在各项指标中都取得了最优的结果, 提取出的建筑物更为完整, 边缘更加精确, 精度更高。此外, 网络结构的消融实验分析证明了各模块的有效性。

关键词 遥感影像; 深度学习; 边缘增强; 全局自注意力; 建筑物提取

随着航空航天遥感技术的发展, 高分辨率对地观测系统逐渐形成, 人们可以方便快速地获取大面积高分辨率、高光谱分辨率和高时间分辨率的对地观测数据^[1]。遥感影像的建筑物自动提取作为遥感智能解译领域的一个重要研究课题^[2], 如何快速、准确、稳定地从影像中提取建筑物, 并基于此信息进行相关分析, 是城市精细化管理^[3]、高精度制图^[4]、土地资源调查^[5]等应用的基础。然而, 遥感影像在采集中易受光照、阴影等因素的影响, 且地球表面建筑物实体的尺度不

一、周围环境多变, 容易出现同谱异物、同物异谱的现象。这些情况造成了地物区分难、边缘模糊等问题, 对遥感影像的精确建筑物提取提出了一系列的要求和挑战。

研究人员通过人工设计特征来实现建筑物提取的传统算法, 大致可以分为面向像素的方法和面向对象的方法。在面向像素的算法中, 通过几何、纹理、阴影等方式逐像素提取建筑物特征^[6]; 在面向对象的方法中, 通过使用支持向量机(SVM)^[7]、AdaBoost^[8]和条件随机场(CRF)^[9]等算法, 以对象为处理单元来标记影像。这些方法虽然可以有效地提取特定特征信息, 但很容易受到图像细节、噪声和质量变化的干扰和影响, 且模型泛化能力不强, 影响模型后续的推广应用。

伴随遥感影像大数据时代的到来^[10]和计算机视觉的发展, 深度学习方法被迅速引入到遥感影像建筑物提取应用中, 许多成熟的算法, 如 Unet^[11]、ResNet^[12]、

1. 首都师范大学资源环境与旅游学院, 北京 100048
2. 北京四维远见信息技术有限公司, 北京 100070
3. 中国科学院空天信息创新研究院, 北京 100094
4. 北京工业职业技术学院, 北京 100144
5. 华北水利水电大学, 郑州 450045
6. 中关村科学城城市大脑股份有限公司, 北京 100081

收稿日期: 2024-01-03; 修回日期: 2024-09-26

基金项目: 国家重点研发计划项目(2022YFB3903602); 北京工业职业技术学院校立课题(BGY2022KY-06QT)

作者简介: 李振, 博士研究生, 研究方向为遥感影像智能解译, 电子信箱: sdlz123@126.com; 刘先林(通信作者), 研究员, 中国工程院院士, 研究方向为测绘、摄影测量、遥感的理论与设备, 电子信箱: liuxl@cae.cn

引用格式: 李振, 张振鑫, 王涛, 等. 结合边缘增强的全局自注意力遥感建筑物提取网络[J]. 科技导报, 2025, 43(13): 69-77; doi:10.3981/j.issn.1000-7857.2024.01.00025

HRNet^[13]等都取得了较好的结果。在此基础上,一些研究人员还在后处理过程中结合多源数据^[14-15]、多层训练样本^[16]等方式来提高建筑物提取的效果。但是,这些模型受卷积核大小与感受野范围的限制,缺乏对整体特征的提取能力,且上下文空间关系建模不足。目前,计算机视觉领域中的 Transformer 结构^[17]因其具有强大的全局建模能力,可以建立长距离的特征依赖,受到了遥感领域广大研究者的关注,但纯粹的 Transformer 结构消耗计算资源巨大,且容易忽略空间细节丰富的浅层特征信息及局部信息,造成建筑物边缘像素和小目标对象丢失。

结合遥感影像建筑物类间差异小,尺度不一、复杂性高等特点,并针对当前大部分深度学习算法对影像全局特征建模不足、边缘细节丢失等问题,提出了一种结合边缘增强的全局自注意力遥感建筑物提取网络 E-GSNet,在增强建筑物边缘的同时,获取特征的全局上下文语义,同时采用逐级上采样解码,兼顾不同尺度建筑物的完整性和准确性,提升遥感影像的建筑物提取精度。

1 研究方法

1.1 网络整体架构

在基于遥感影像的深度学习建筑物提取网络中,浅层特征包含丰富的建筑物空间细节,深层特征包含高阶的语义信息。卷积神经网络可以实现从浅层到

深层逐步提取特征,但在线性传递过程中,容易丢失边缘细粒信息;且网络感受野受卷积核大小固定的限制,缺乏提取全局上下文信息的能力,长距离空间关系建模不足;此外,多数网络直接将深层特征解码为预测图,容易丢失大量空间细节。针对这些问题,提出了嵌入遥感边缘先验特征,顾及全局上下文语义,高、低层次信息融合互补解码的 E-GSNet。

网络整体架构如图 1 所示,在改进的 ResNet50^[12]主干上,添加边缘增强模块,全局自注意力特征表达模块和逐级上采样解码模块。首先,将公开数据集的训练子集影像采样为 256×256 大小,在编码前,首先,使用边缘增强模块对建筑物边界信息显式提取,捕获边缘先验信息与原始影像叠加融合,输入主干网络进行特征编码,线性地获取并传递浅层空间细节与深层高阶语义;其次,将富含高阶语义信息的深层特征映射为一组序列向量,使用全局自注意力特征表达模块实现长距离特征依赖关系与多层次边缘增强特征的融合表达,学习并建模全局上下文语义信息;再次,使用逐级上采样的方式对增强后的特征进行解码,通过跃层连接的方式与富含空间细节的低阶编码特征融合并输出阶段性预测结果;最后,聚集所有的阶段性预测结果,实现特征对象间信息的传递与继承,生成精细边缘的遥感影像建筑物提取结果。

1.2 边缘增强模块

在主干网络前端,构造了一种边缘增强模块,其

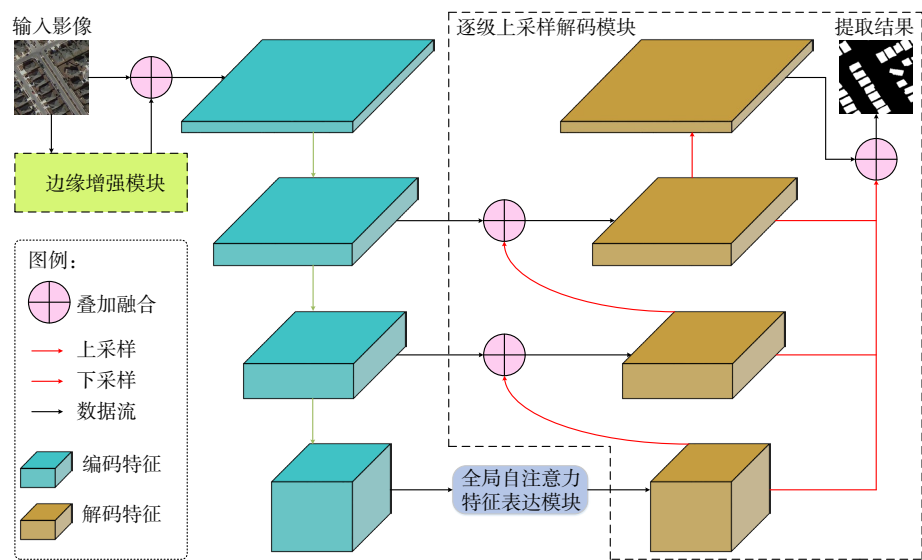


图 1 E-GSNet 网络整体框架

结构如图 2 所示,由卷积数据流与边缘提取数据流 2 个部分并行组成。卷积数据流通过执行不同大小的卷积核运算初步获取影像空间细节,捕获不同尺度感受野的特征;边缘提取数据流使用 Canny 算法^[18],细化并凸出最优边界显著信息;边缘增强模块将边界显著信息与卷积特征融合,使得特征在提取的初步阶段可以有效集成多尺度建筑物边界,挖掘遥感影像内在知识先验,引导深度学习网络实现精细的特征编码。

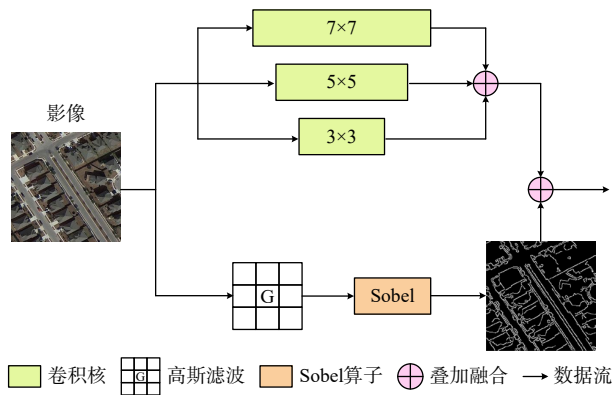


图 2 边缘增强模块结构

卷积数据流由 3 个大小为 7×7 、 5×5 和 3×3 的并行卷积组成,对于输入的原始影像,首先执行 3 次卷积操作获取不同感受野的多尺度特征,并通过设置相应的零填充与采样步长,保持特征图大小一致,然后使用通道叠加实现特征通道间的信息聚合。该过程可以表示为

$$F_i = \text{Conv}_{2i+1}(\text{Image}) \quad (1)$$

$$F_{\text{final}} = \text{Concat}(F_1, F_2, F_3) \quad (2)$$

式中, F_i 为卷积后得到的特征图, $i=1, 2, 3$, Conv_{2i+1} 为核大小为 $2i+1$ 的卷积操作, Image 为输入影像, Concat 表示特征通道叠加操作, F_{final} 表示卷积数据流最后的输出特征。

边缘提取数据流的主要设计思路是通过捕获建筑物和背景之间的灰度值突变以凸出边缘显著信息,使用 Canny 算法实现,具体步骤包括高斯滤波去噪和 Sobel 算子找寻梯度,高斯滤波可以对图像进行平滑去噪处理, Sobel 算子求每个像素点梯度的幅值和方向,经过非极大值抑制和弱边缘舍弃后,得到最终的边缘提取结果。高斯滤波器和像素梯度计算可以用以下公式表示

$$G(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (3)$$

$$\begin{cases} g_{(x,y)} = \sqrt{g_x^2 + g_y^2} \\ \theta_{(x,y)} = \arctan\left(\frac{g_y}{g_x}\right) \end{cases} \quad (4)$$

式中, G 为二维高斯函数经过采样以及归一化之后得到的高斯滤波器, x, y 为像素位置, g 为梯度的模, θ 表示梯度的方向。

1.3 全局自注意力特征表达模块

全局自注意力特征表达模块主要依靠 Transformer 机制实现,它打破了卷积神经网络中局部感受野的限制,通过自注意力建模边缘增强特征的长距离依赖关系并为其赋予全局上下文信息。

模块结构如图 3 所示,首先,将特征输入展平并线性映射为一组序列向量,然后对向量增添一个可学习的位置编码向量以保留其空间位置信息;其次,将序列向量输入 Transformer 模块编码,交替使用多头自注意力与多层感知机对向量进行全局信息交互与非线性变换,在多头自注意力与多层感知机之间分别使用 LayerNorm^[19]对输入向量进行归一化,并使用残差连接保证信息高效传递继承;重复 N 次编码(图 3 中 $N=2$),最后,将 Transformer 模块编码的向量输出,映射恢复为原始大小的特征输出。具体过程可以表示为

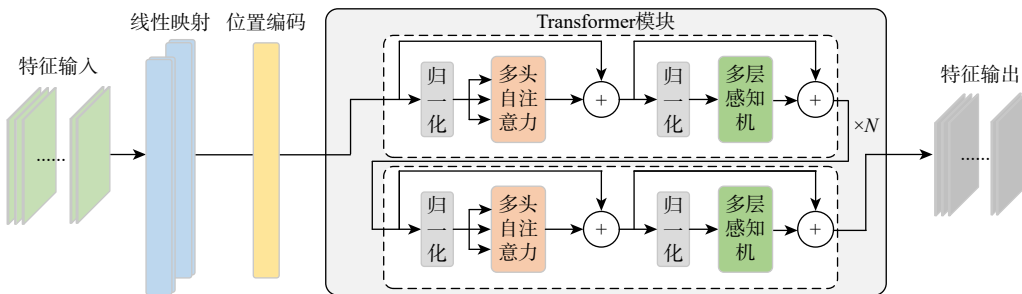


图 3 全局自注意力特征表达模块结构

$$V_{in} = Flat(F_{in}) + V_{pos} \quad (5)$$

$$V_{mid} = MSA(LN(V_{in})) + V_{in} \quad (6)$$

$$V_{out} = MLP(LN(V_{mid})) + V_{mid} \quad (7)$$

式中, V_{in} 、 V_{mid} 和 V_{out} 分别为 Transformer 编码模块的输入向量、中间向量和输出向量, F_{in} 、 V_{pos} 是全局自注意力特征表达模块中的输入特征和位置编码向量, $Flat(\cdot)$ 、 $MSA(\cdot)$ 、 $MLP(\cdot)$ 、 $LN(\cdot)$ 为展平、多头注意力、多层感知机和层归一化操作。

1.4 逐级上采样解码模块

在 ResNet 等网络中, 直接将最深层特征解码的方式极易丢失空间信息细节。为此, 受 UNet^[11] 结构启发, 设计了一种改进的逐级上采样解码模块, 架构如图 1 所示。在解码过程中, 将被赋予了全局上下文交互信息的特征通过逐层级的双线性插值上采样方法, 在保证计算速度的同时获取较好的插值效果, 渐进地恢复为与主干网络对应阶段编码特征相同的大小, 与富含空间细节的同阶编码特征通过跃层连接的方式叠加融合, 同时输出阶段性预测结果; 逐层重复此操作, 最后, 叠加融合所有的阶段性预测结果并输出。通过这种逐级上采样解码的方式, 将深层特征的高阶语义信息与浅层特征的丰富空间细节相结合, 实现特征对象间信息的高效融合与恢复, 准确提取建筑物边界, 得到高精度的预测结果。

2 实验过程

2.1 数据集介绍

本文采用 2 个公开的遥感影像建筑物数据集对各提取方法进行对比与测试。WHU 建筑物数据集^[20]对各个建筑物提取方法进行对比与测试(https://study.rsgis.whu.edu.cn/pages/download/building_dataset.html)。该数据集是目前国际上范围最大、标注精度最高的遥感影像建筑物公开数据集, 能够很好地验证不同建筑物提取方法的性能和鲁棒性。数据来自新西兰土地信息服务网站, 涵盖了新西兰基督城(Christchurch, New Zealand)中村庄、住宅、商业区、工业区等不同场景下 18.7 万座建筑物, 共计 8188 张影像, 大小为 512×512 像素, 空间分辨率为 0.3 m, 包含红、绿、蓝 3 个波段。在数据集中, 官方划分 4736 张影像为训练集, 1036 张影像为验证集, 2416 张

影像为测试集。顾及当前计算机硬件图形处理器的性能, 将数据集中每张影像无重叠地裁剪成 256×256 大小的切片。INRIA 建筑物数据集^[21](<https://project.inria.fr/aerialimagelabeling>)由法国国家信息与自动化研究所提供, 涵盖不同国家的 5 个城市地区, 建筑密度、风格、排列各不相同, 数据类型具有多样性, 可以验证各方法的泛化能力。数据集空间分辨率为 0.3 m, 包括红、绿、蓝 3 个波段。本文将每张影像不重叠分割, 得到 18000 张 500×500 大小的标签影像, 设置训练子集和验证子集的比例为 8:2。

2.2 实验环境与训练细节

模型代码基于 Pytorch 框架实现。所有的实验都是在 1 台配备 Intel(R) Xeon(R) CPU E5-2686 v4 CPU(2.30 GHz), 操作系统为 Ubuntu18.04 的服务器上运行, 使用一块 NVIDIA GeForce RTX 3090 图形处理器(24 G 内存)进行训练和测试。

在网络训练的每一个批次过程中, 实验将每个影像随机进行水平翻转与正负 15° 之间的旋转, 提高模型对数据的泛化能力, 避免网络的过拟合。其次, 选择混合的交叉熵损失和 Dice 损失^[22]作为目标函数, 使用 AdamW 方法进行模型优化, 设置初始学习率为 0.002, 权值衰减为 1×10^{-4} 。训练数据的批次大小设置为 8。

2.3 评价指标

采用当前遥感建筑物提取任务中最常见的评价指标来定量评估本文的网络性能, 包括准确率(P)、召回率(R)、 F_1 -score(F_1)、交并比(IoU)和总体像素精度(OA)。具体计算如下

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (10)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (11)$$

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \quad (12)$$

其中, TP (true positive)为正确预测的建筑物像素数; TN (true negative)为正确预测的非建筑物像素数; FP (false positive)为错误预测为建筑物的像素数, 即错检; FN (false negative)表示错误预测为非建筑物的像素数, 即漏检。

P 为正确预测的建筑物像素数与所有预测为建筑物的像素数比值; R 为正确预测的建筑物像素数与真实建筑物的像素数比值; F_1 是对 P 和 R 的调和参数平均; IoU 是真实建筑物区域和预测建筑物区域交集和并集的比值, OA 代表了所有正确预测的整体像素精度。

3 结果与分析

3.1 对比实验分析

为了验证所提方法的有效性, 实验在 2 个建筑物数据集的测试集上, 分别将 E-GSANet 与一系列的深度学习方法(UNet^[11]、PSPNet^[23]、DANet^[24]、PAN^[25]、MA-FCN^[26]、SRI-Net^[27]、FPCRF^[28]、HCRB-MSAN^[29])进行对比, 从定量指标和定性可视化 2 个方面验证该方法在遥感影像建筑物提取方面的性能。在使用的数据集上, 对于提供结果的比较算法, 直接引用相应论文的指标; 对于没有提供精度结果的算法, 使用官方或开源代码复现算法并得到了相应的测试结果, 在表格方法中使用“*”标注。

表 1 为不同方法的定量指标对比结果, E-GSANet 在各项指标上都有更好的表现, 获得了最佳的 P 、 R 、 F_1 、 IoU 和 OA , 分别为 96.82%、95.00%、95.90%、92.13% 和 99.11%, 与综合指标近似的 HCRB-MSAN 相比, E-GSANet 在获得较高指标的同时, 可以保持相对一致的 P 和 R , 有效平衡错检和漏检区域, 在建筑物对象的边缘保持方面可以获得非常好的效果和提升, 这在图 4 的可视化结果对比中也可以获得佐证。

表 1 不同方法的对比结果

方法	$P/\%$	$R/\%$	$F_1/\%$	$IoU/\%$	$OA/\%$
UNet*	92.62	93.50	93.06	87.02	98.47
PSPNet*	93.41	91.95	92.68	86.35	98.40
DANet*	94.01	89.71	91.81	84.86	98.24
PAN*	92.80	91.49	92.14	85.43	98.28
MA-FCN	94.50	94.20	94.30	89.50	—
SRI-Net	95.21	93.28	94.23	89.09	—
HCRB-MSAN	96.78	94.68	95.72	91.79	99.07
E-GSANet	96.82	95.00	95.90	92.13	99.11

注: *表示没有提供精度结果的算法, 使用官方或开源代码复现算法并得到了相应的测试结果。

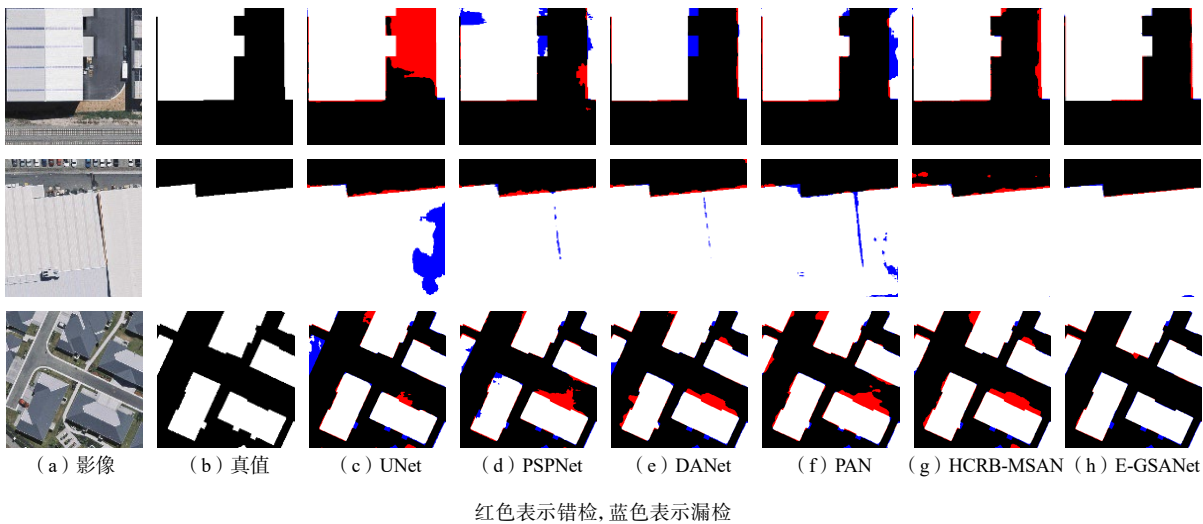


图 4 WHU 建筑物数据集不同方法可视化对比结果

从测试集中选取了 3 组典型类型的建筑物来对比其他方法与 E-GSANet 定性可视化结果的差异, 包括易混淆地物(第 1 行), 大尺度建筑物(第 2 行)和密集小尺度建筑物(第 3 行)。如图 4 所示, 为了直观显示, 使用红色表示错检, 蓝色表示漏检。在第 1 行中, UNet 没有正确区分建筑物与相似路面的差异, 其他

方法在建筑物不规则突起和右侧边缘存在一定漏检和错检; 在第 2 行大尺度建筑物影像中, 与 E-GSANet 相比, 其他方法出现了内部缺失和外部轮廓不清晰的现象; 同样, 在第 3 行密集小尺度建筑物中, 本文方法可以获得较为精确的提取结果, 很少有差异区域。从视觉表现来看, 得益于边缘增强、全局自注意力特征

表达和逐级上采样解码,E-GSANet 较好地保持了各类建筑物的边缘准确性、整体一致性和空间细节。

表 2 展示了各方法在 INRIA 测试子集上的定量评估结果,E-GSANet 在 R、F1、IoU 和 OA 这 4 项指标上得到了最优的结果,分别为 89.87%、89.06%、80.28%、97.02%。同时,网络在取得具有竞争力的 P 和 R 指标之外,也保持了这两个指标的高度一致性。取得最优 P 值的 PAN 方法,与对应的 R 相差 13.09%,这种 P 与 R 之间的巨大差异,一定程度上反映了网络性能的失衡,无法鲁棒地对建筑物进行准确提取。

如图 5 所示,选取 INRIA 测试子集中的 3 组典型建筑物,展示各方法的可视化结果差异,包括类内差异大的建筑物(第 1 行),被遮挡的建筑物(第 2 行)和不规则建筑物(第 3 行)。从对比图 5 中可以看出,E-GSANet 可以提取出各类建筑物较为准确的边界和完整的内部。在第 1 行中,大部分模型不能准确区分同一栋建筑的不同颜色差异,有不同程度的漏检现

象;在第 2 行中,E-GSANet 可以识别出大部分的被遮挡建筑,在所有模型中保持了最少的漏检区域;对于第 3 行的不规则建筑物,与其他模型相比,E-GSANet 保持了最为精准的建筑物边界。

表 2 INRIA 建筑物数据集上不同方法的对比结果

方法	$P/\%$	$R/\%$	$F_1/\%$	$IoU/\%$	$OA/\%$
UNet [*]	82.73	80.09	81.39	68.61	95.06
PSPNet [*]	81.63	86.55	84.02	72.45	95.56
DANet [*]	82.66	84.37	83.51	71.68	95.51
PAN [*]	90.11	77.02	83.05	71.02	95.76
FPCRF	—	—	87.65	74.79	95.81
SRI-Net	85.77	81.46	83.56	71.76	—
HCRB-MSAN	89.56	88.13	88.84	79.92	97.01
E-GSANet	88.26	89.87	89.06	80.28	97.02

注:^{*}表示没有提供精度结果的算法,使用官方或开源代码复现算法并得到了相应的测试结果。

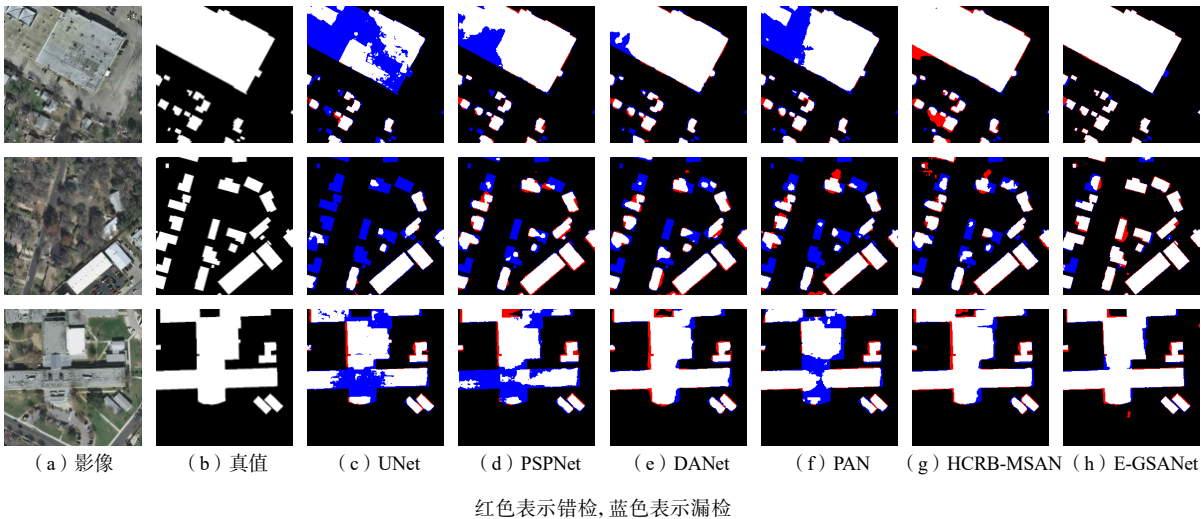


图 5 INRIA 建筑物数据集不同方法可视化对比结果

3.2 消融实验分析

为了验证 E-GSANet 中 3 个模块的效果,进行了一系列的消融实验。结果如表 3 所示,当不添加任何

模块,仅使用主干网络时,可以获得 92.25% 的 F_1 和 85.62% 的 IoU 指标值;以此作为基线,当增加逐级上采样解码模块时,效果增益显著, F_1 和 IoU 分别提升

表 3 E-GSANet 各模块的消融实验结果

边缘增强模块	全局自注意力特征表达模块	逐级上采样解码模块	$P/\%$	$R/\%$	$F_1/\%$	$IoU/\%$	$OA/\%$
—	—	—	91.43	93.09	92.25	85.62	98.28
—	—	√	94.98	94.93	94.95	90.39	98.89
—	√	√	95.81	94.88	95.34	91.10	98.98
√	√	√	96.82	95.00	95.90	92.13	99.11

了 2.7% 和 4.77%;在此基础上添加全局自注意力特征表达模块,网络性能得到进一步的增强,相比仅使用逐级上采样解码模块, F_1 增加了 0.39%, IoU 增加了 0.71%;最后,网络集成 3 个模块,各项指标都达到了最优的结果,其精度结果相较于基线方法有明显的改善。实验结果表明,边缘增强模块、全局自注意力特征表达模块和逐级上采样解码模块可以有效提高建筑物特征的提取效果,增益网络性能。

4 结论

针对现有遥感影像建筑物提取深度学习算法的全局特征建模不足、边缘细节丢失等问题,提出了嵌入遥感边缘先验特征,顾及全局上下文语义、高、低层次信息融合互补解码的遥感建筑物提取网络 E-GSNet。网络包含 3 个模块,边缘增强模块可以对建筑物边界信息显式提取,挖掘遥感影像内在知识先验,引导网络充分利用边缘显著特征;全局自注意力特征表达模块使用 Transformer 机制,建模边缘增强特征的长距离依赖关系,捕获全局上下文信息;逐级上采样解码模块可以实现特征对象间信息的高效融合与恢复,准确提取建筑物边界。在 WHU 建筑物数据集上进行了大量且详尽的实验与分析,实验分析证明,与列举的其他方法相比,E-GSNet 能够取得最佳的遥感影像建筑物提取结果,这可以为城市精细化管理,土地资源调查等应用提供一定的支持。

在未来的工作中,计划探寻一种轻量化方法,在减少对计算硬件消耗的前提下,降低参数量的同时保持提取精度,使模型更适合于实际生产任务;此外,还将尝试迁移该模型至多类别提取任务,验证方法的泛化能力,并研究遥感影像中城市多类别地物的高效及高精度提取工作。

参考文献(References)

- [1] Li D R, Wang M, Jiang J. China's high-resolution optical remote sensing satellites and their mapping applications[J]. Geo-spatial Information Science, 2021, 24(1): 85-94.
- [2] 龚健雅, 张展, 贾浩巍, 等. 面向多源数据地物提取的遥感知识感知与多尺度特征融合网络[J]. 武汉大学学报(信息科学版), 2022, 47(10): 1546-1554.
- [3] Hu S G, Wang L. Automated urban land-use classification with remote sensing[J]. International Journal of Remote Sensing, 2013, 34(3): 790-803.
- [4] Ural S, Hussain E, Shan J. Building population mapping with aerial imagery and GIS data[J]. International Journal of Applied Earth Observation and Geoinformation, 2011, 13(6): 841-852.
- [5] Goetz S J, Smith A J, Jantz C, et al. Monitoring and predicting urban land use change[C]//Proc. IEEE International, Geoscience and Remote Sensing Symposium, IGARSS'03. Toulouse, France: IEEE, 2003, 3: 1567-1569.
- [6] Peng J, Liu Y C. Model and context-driven building extraction in dense urban aerial images[J]. International Journal of Remote Sensing, 2005, 26(7): 1289-1307.
- [7] Inglada J. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2007, 62(3): 236-248.
- [8] Cetin M, Halici U, Aytekin Ö. Building detection in satellite images by textural features and Adaboost[C]//Proceedings of IAPR Workshop on Pattern Recognition in Remote Sensing. Piscataway, NJ: IEEE, 2010: 1-4.
- [9] Li E, Femiani J, Xu S B, et al. Robust rooftop extraction from visible band images using higher order CRF[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(8): 4483-4495.
- [10] Zhang B, Chen Z C, Peng D L, et al. Remotely sensed big data: Evolution in model development for information extraction[point of view][J]. Proceedings of the IEEE, 2019, 107(12): 2294-2301.
- [11] 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.
- [12] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770-778.
- [13] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 5686-5696.
- [14] Bittner K, Adam F, Cui S Y, et al. Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks[J]. IEEE Journal of Selected Topics in Applied

- Earth Observations and Remote Sensing, 2018, 11(8): 2615–2629.
- [15] Cao Z Y, Fu K, Lu X D, et al. End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(11): 1766–1770.
- [16] Liu Y B, Zhang Z X, Zhong R F, et al. Multilevel building detection framework in remote sensing images based on convolutional neural networks[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 11(10): 3688–3700.
- [17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. [2023–12–23]. <http://arxiv.org/abs/2010.11929>.
- [18] Canny J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679–698.
- [19] Ba J L, Kiros J R, Hinton G E. Layer normalization[EB/OL]. [2023–12–23]. <https://arxiv.org/abs/1607.06450v1>.
- [20] 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.
- [21] Maggiori E, Tarabalka Y, Charpiat G, et al. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark[C]//Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Piscataway, NJ: IEEE, 2017: 3226–3229.
- [22] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[EB/OL]. [2023–12–23]. <https://arxiv.org/abs/1606.04797v1>.
- [23] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 6230–6239.
- [24] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 3141–3149.
- [25] Li H C, Xiong P F, An J, et al. Pyramid attention network for semantic segmentation[EB/OL]. [2023–12–23]. <https://arxiv.org/abs/1805.10180v3>.
- [26] Wei S Q, Ji S P, Lu M. Toward automatic building footprint delineation from aerial images using CNN and regularization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(3): 2178–2189.
- [27] Liu P H, Liu X P, Liu M X, et al. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network[J]. Remote Sensing, 2019, 11(7): 830.
- [28] Li Q Y, Shi Y L, Huang X, et al. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(11): 7502–7519.
- [29] Li Z, Zhang Z X, Chen D, et al. HCRB-MSAN: Horizontally connected residual blocks-based multiscale attention network for semantic segmentation of buildings in HSR remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 5534–5544.

Global self-attention remote sensing building extraction network combined with edge enhancement

LI Zhen¹, ZHANG Zhenxin¹, WANG Tao², PENG Xueli³, YUE Guijie⁴, ZHANG Deyu², LIU Xianlin^{2,5*}, LI Jianhua⁶

1. College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China

2. Beijing Geo-Vision Information Technology Co., Ltd., Beijing 100070, China

3. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

4. Beijing Polytechnic College, Beijing 100144, China

5. North China University of Water Resources and Electric Power, Zhengzhou 450045, China

6. Zhongguancun Smart City Co., Ltd., Beijing 100081, China

Abstract The accurate and efficient extraction of building from remote sensing images is fundamental for applications such as fine urban management, high-precision mapping, and land resource investigation. It is essential to investigate how to leverage image features for intelligent interpretation. This study introduces a global self-attention network with edge-enhancement (E-GSNet) for remote sensing building extraction. The network integrates the edge enhancement module into the encoder backbone, providing the network with a priori knowledge about boundaries, and then establishes long-distance dependency relationships between features using the global self-attention feature expression module, enabling the fusion of salient features with edge-enhanced features. A stepwise up-sampling decoding module is designed to fuse the shallow features with rich spatial detail information and the deep features with high-order semantic information to obtain accurate extraction results of buildings. The comparison experiments between E-GSNet and the current mainstream methods are conducted based on two open-source remote sensing building datasets. The quantitative analysis and qualitative demonstrations prove that E-GSNet achieves optimal results across all evaluation metrics, yielding more complete building extractions, precise edges, and higher accuracy. Additionally, network structure ablation experiments and analysis demonstrate the effectiveness of each module.

Keywords remote sensing image; deep learning; edge enhancement; global self-attention; building extraction ●



(责任编辑 王微)