融合CNN与Transformer的高分辨率遥感影像建筑物双流提取模型

刘宇鑫[1,2]，孟瑜[1]，邓毓骋[1]，陈静波[1]，刘帝佑[1]

1. 中国科学院空天信息创新研究院 国家遥感应用工程技术研发中心, 北京 100094;
2. 中国科学院大学 电子电气与通信工程学院, 北京 100049

摘 要： 卷积神经网络（Convolutional Neural Network，CNN）和Transformer已被广泛应用于高分辨率遥感影像的建筑物提取任务。然而，CNN在建模长距离空间依赖时仍存在挑战，导致提取的建筑物存在内部空洞问题；而Transformer在捕捉空间局部细节特征上存在不足，容易导致建筑物边缘模糊及小型建筑物的漏检。为解决上述问题，本文提出了一种新型的双流网络模型用于高分辨率遥感影像的建筑物提取，名为ILGS-Net（Network for the Integration of Local and Global Features Stream）。该模型将CNN与Transformer相结合，采用多层级的局部-全局特征融合模块，实现了对建筑物的局部细节特征与全局上下文特征的高效融合。同时，在目标函数中引入边缘损失函数约束模型训练，提高了建筑物边界的定位精度。在三个高分辨率建筑物数据集上的实验结果显示，所提出方法的交并比均高于本文所对比的最佳方法，平均提高了1%。

关键词： 建筑物提取，深度学习，双流网络，边缘损失，局部和全局特征融合

中图分类号： TP701

引用格式： 刘宇鑫,孟瑜,邓毓骋,陈静波,刘帝佑.XXXX.融合CNN与Transformer的高分辨率遥感影像建筑物双流提取模型.遥感学报,XX(XX):1-12

LIU Yuxin, MENG Yu, DENG Yupeng, CHEN Jingbo, LIU Diyou. XXXX. Integration of CNN and Transformer for High-Resolution Remote Sensing Image Building Extraction: A Dual-Stream Network. National Remote Sensing Bulletin, DOI:10.11834/jrs.20243307]

1 引 言

遥感影像建筑物提取在城市规划、土地管理和灾害监测等领域具有重要的应用价值（Zhu等，2017；Cooner等，2016）。近年来，随着高分辨率地面观测技术的不断发展，遥感影像的空间分辨率得到显著提升，为建筑物提取任务提供了更为详细和丰富的空间结构和细节信息，然而也带来了新的挑战。由于建筑物内部结构和材质的多样性，高分辨率遥感影像中建筑物存在明显的类内差异，容易发生识别错误。另一方面，高分辨率遥感影像上存在大量干扰信息，如阴影和植被遮挡等，给建筑物的完整提取和边界定位造成很大困难。因此，建筑物提取算法需要兼顾建筑物内部的空间细节特征和整体的结构布局特征，以减少因上述因素导致的建筑物提取不完整或不准确的问题。

深度学习模型，尤其是卷积神经网络（Convolutional Neural Network，CNN）模型在建筑物提取任务中表现出优异性能。Long等（2015）首先提出全卷积神经网络模型（Full Convolutional Network，FCN），该模型使用卷积层取代CNN中的全连接层，从而具有了像素级别的端到端识别能力。Ronneberger等（2015）提出UNet网络模型，该模型利用下采样编码层提取深层语义特征，并通过上采样解码层将图像恢复到原始大小，同时通过跨层连接有效融合了低层和高层特征。Chen等（2018）提出使用空洞空间卷积池化金字塔模块（Atrous Spatial Pyramid Pooling）提取和融合特征的多尺度上下文信息。林娜等（2022）改进特

收稿日期： 2023-XX-XX；**预印本：** XXXX-XX-XX
基金项目： 国家重点研发计划课题(2021YFB3900503)
第一作者简介： 刘宇鑫,研究方向为遥感图像智能解译。E-mail：liuyuxin213@mails.ucas.ac.cn
通信作者简介： 孟瑜,研究方向为遥感时间序列分析、遥感数据知识工程。E-mail：mengyu@aircas.ac.cn

© 《遥感学报》

征金字塔网络（Feature Pyramid Networks，FPN）（Lin等，2017）和优化非极大值抑制算法（Non-Maximum Suppression）以增强Mask-RCNN（Mask Region-based Convolutional Neural Network）在建筑物提取方面的能力。李星华等（2022）提出一种包含多路径卷积融合模块和大感受野特征感知模块的多层次特征融合网络，以多个维度提取建筑物特征和解决卷积感受野大小限制问题。许正森等（2022）将注意力机制引入胶囊网络，使模型更加关注显著性强、信息量大的特征通道和空间位置。吕少云等（2023）使用ResNet、ASPP和UNet++相结合的方式，提出Res_ASPP_UNet++模型。尽管CNN能够高效提取局部特征，但由于卷积核的感受野尺寸限制，CNN难以捕捉图像的全局特征（Strudel等，2021），导致提取的建筑物存在内部空洞和遗漏等问题。

Transformer（Vaswani等，2017）是基于自注意力机制（self-attention）构建的神经网络模型，其特点在于能够捕捉全局上下文特征。Dosovitskiy等（2020）提出vision Transformer（ViT）。该模型将图像块映射为独立的序列，用于图像分类。为了适应语义分割任务，Strudel等（2021）在ViT的基础上引入语义分割解码器Mask Transformer，提出了Segmenter模型。为了提取多尺度的全局特征信息，Liu等（2021）通过引入移动窗口并限制特征的全局关注范围，设计了Swin Transformer模型。鉴于Swin Transformer在各种计算机视觉任务中的优越表现，Yuan等（2021）和Chen等（2021）将该模型作为编码模块引入到建筑物提取任务中，有效解决了建筑物内部空洞问题。Wang等（2022）改进Swin Transformer的内部结构，并提出BuildFormer模块，以降低模型在处理高分辨率遥感影像时所需的参数数量和计算开销。与CNN不同，Transformer模型以图像块为单元进行特征学习，虽然能通过自注意力机制捕捉长距离目标之间的关联性，但由于缺乏局部感受野的约束，图块内部的细节信息无法得到很好地保留，导致提取的建筑物存在边缘模糊的问题（Xiao等，2022）。

针对以上问题，本文提出一种基于CNN与Transformer的建筑物提取模型（ILGS-Net，Network for the Integration of Local and Global Features Stream），能兼二者之长，实现卷积的局部细节特征与自注意力的全局上下文特征的高效融合。首先，采用EfficientNet-b3模型（Tan和Le，2019）和BuildFormer模型构成双流骨干网络，分别用于提取图像的局部特征和全局特征。其次，设计了一种多层级的局部-全局特征融合模块（Local and Global-Feature Fusing，LGFF），以融合局部特征和全局特征。再次，设计了一个上下文聚合解码器（Context Aggregation Decoder，CAD），来聚合和解码融合后的特征，以获得最终的提取结果。为了使模型更多地关注建筑物的边缘特征，在目标函数中引入了边缘损失，以提取更为清晰的建筑物边界。最终实现对复杂场景下遥感影像建筑物目标的高精度提取。

2 研究方法及原理

局部-全局特征双流提取模型（ILGS-Net）由局部特征流（Local Feature Stream，LFS）、全局特征流（Global Feature Stream，GFS）、局部-全局特征融合模块（LGFF）和上下文聚合解码器（CAD）组成，模型结构如图1所示。

本模型的不同结构具有不同功能。LFS的主要功能是捕获多层次的空间细节特征，这种特征蕴含着局部范围内的纹理、边缘等微小信息。GFS的主要功能是提取影像中的长距离依赖特征，以更好地理解建筑物之间的整体结构及布局关系。LGFF模块旨在高效地融合局部特征和全局特征，以集成二者的优势。CAD旨在根据融合特征理解建筑物的空间语义信息，并输出识别结果。

2.1 局部特征流结构

卷积神经网络通过构建多层卷积层、池化层以及一些非线性激活函数，能够提取图像中的局部空间、纹理等细节特征。而在高分辨率遥感影像上提取的信息越丰富，相应的卷积神经网络模型的参数和计算开销则越大。为了在计算效率和特征提取能

力之间做到权衡，本文选择了一种轻量级CNN模型——EfficientNet-b3作为LFS。

EfficientNet网络由多个堆叠的倒残差模块（MBConv）组成。利用MBConv模块中的通道注意力机制和一系列的深度可分离卷积，该模型能够在较小的计算开销下充分提取局部特征。EfficientNet-b3是EfficientNet的一个变体模型，相

© 《遥感学报》

较于EfficientNet-b0，它具有更大的网络宽度、深度和输入图片分辨率，旨在提高模型的准确性和表达能力。EfficientNet-b3总共包含九个阶段，除了第一个和最后一个阶段外，第二至八阶段由连续的MBConv模块堆叠而成。每个MBConv模块中深度可分离卷积的卷积核尺寸为3或5，卷积扩张因子为1或6。

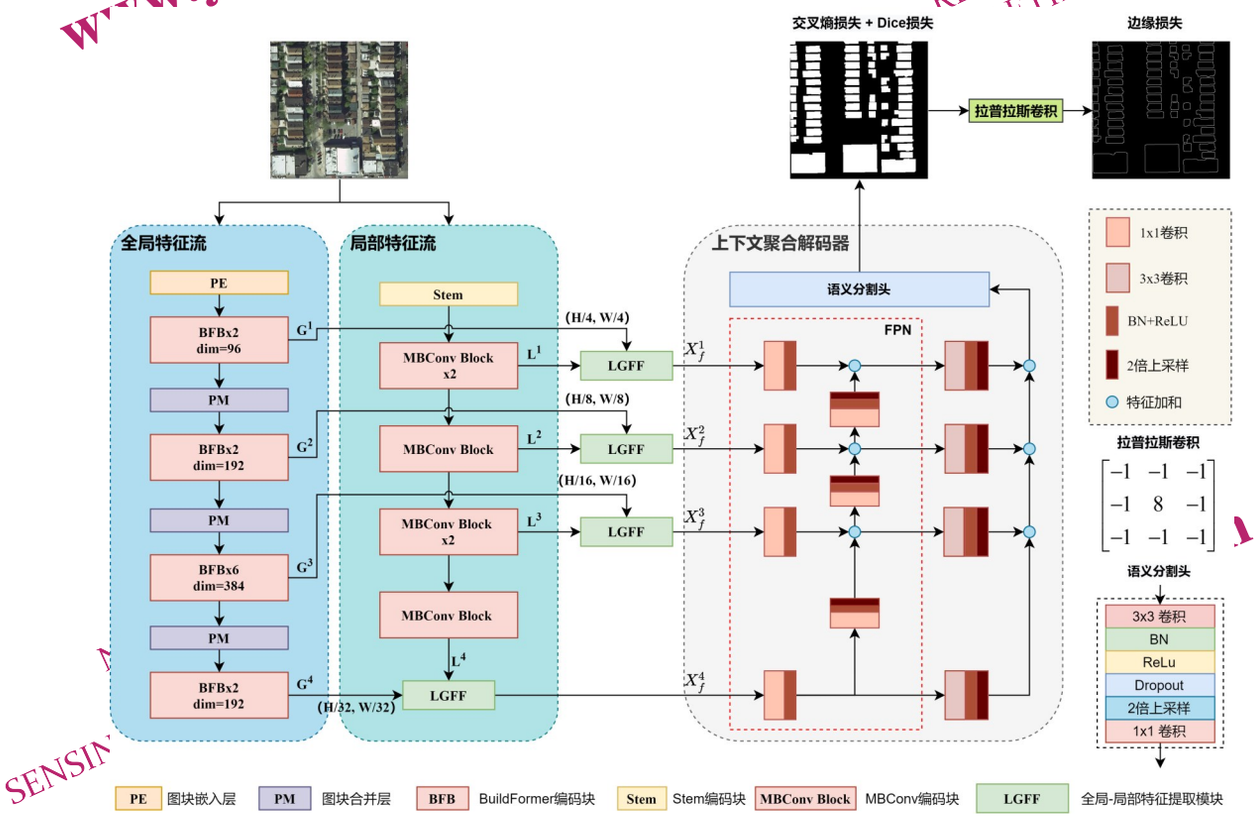


图1 局部-全局特征双流融合模型总体结构

Fig.1 The overall of ILGS-Net

2.2 全局特征流结构

Transformer模型通过其内部的多头自注意力机制，在处理输入图片数据时动态地为每个图片块分

配不同的注意力权重。这使得模型能够捕获图像中不同位置的关系，特别适用于长距离依赖关系。当处理高分辨率遥感影像时，为了更好提取影像中的全局依赖特征，需对影像分块得更多，带来更高的模型计算开销。为了在同等提取能力下尽可能减少模型的计算开销，本文使用Transformer系列模型BuildFormer作为GFS。

BuildFormer是一个改进版的Swin-Transformer。它主要改进了自注意力机制和跨窗口交互模块的计算方式，旨在不影响模型精度的情况下进一步减少Swin Transformer的参数。如图1所示，BuildFormer由基本的编码块（BuidFormer Block，BFB）、图块嵌入层（Patch Embedding，PE）和图块合并层（Patch Merging，PM）组成。PE的作用是将原始图像分成多个相互重叠的图片块。PM的目的是获得层级式的全局注意力特征。如图2所示，BFB主要包含一个基于窗口的线性多头自注意力机制（W-LMHSA）和2个卷积多层感知机层（C-MLP）。在BuildFormer中，W-LMHSA将经过PE层的特征划分为互不重叠的窗口，并对每个窗口内的特征进行线性多头注意力操作。对于每个局部窗口内的特征，线性多头注意力机制可以定义为：

$$LMHSA(\boldsymbol{X}) = Concat\left(head_1,\cdots,head_h\right)\boldsymbol{W}_o \quad (1)$$

其中$\boldsymbol{X}$是输入特征向量，h代表自注意力头的数量。$\boldsymbol{W}_o \in \mathbb{R}^{N\times D}$是一个投影映射矩阵，$D$是输入特征向量的维度，$N$是窗口的尺寸。每个$head$都代表一个自注意力操作，可以被定义为：

© 《遥感学报》

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\frac{\sum_j \boldsymbol{V}_{ij}+\left(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right)\left(\left(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2}\right)^T\boldsymbol{V}\right)}{N+\left(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right)\sum_j\left(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2}\right)^T_{ij}} \tag{2}$$

$$\boldsymbol{Q}=\boldsymbol{X}_m\boldsymbol{W}_q\in\mathbb{R}^{N\times d} \tag{3}$$

$$\boldsymbol{K}=\boldsymbol{X}_m\boldsymbol{W}_k\in\mathbb{R}^{N\times d} \tag{4}$$

$$\boldsymbol{V}=\boldsymbol{X}_m\boldsymbol{W}_v\in\mathbb{R}^{N\times d} \tag{5}$$

$\boldsymbol{X}_m$是第m个头的输入特征向量。$\boldsymbol{Q}$，$\boldsymbol{K}$和$\boldsymbol{V}$分别是query特征，key特征以及value特征，它们分别是由三个投影矩阵$\boldsymbol{W}_q$，$\boldsymbol{W}_k$和$\boldsymbol{W}_v$与输入特征向量相乘所得。d是第m个头的维度，$d=D/h$。

卷积多层感知机（C-MLP）的目的是加强窗口之间特征的交互，其主要由两个常规1×1卷积、一个Depth-wise卷积构成。

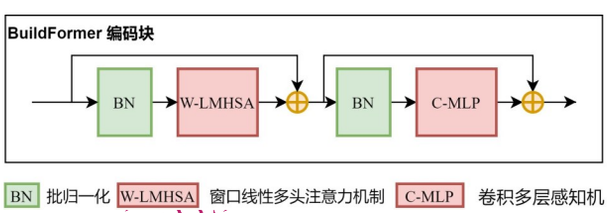


图2　BuildFormer编码块的内部结构

Fig.2　The structure of BuildFormer Block

2.3　局部-全局特征融合模块

LFS和GFS分别输出图像的局部和全局特征，此小节将重点论述这两种特征具体的融合方法。由于卷积操作的局限性，基于CNN的编码器难以对全局上下文信息进行建模。相比之下，基于Transformer的编码器在提取远距离上下文信息方面具有鲁棒性，但在捕获细粒度空间特征方面存在不足。因此，本文提出LGFF来融合这些不同层级的语义特征，以提升模型对目标对象的提取能力。从图1中可以看到，LGFF模块将多个层级上的GFS提取的全局特征和LFS提取的局部特征进行融合。从图3可以

看出，为了与GFS中四个全局特征的通道进行匹配，首先对从LFS中得到的四个局部特征图进行核尺寸为1的卷积操作。之后，为了增强空间局部细节并抑制不相关区域，对局部特征使用空间注意力机制。具体操作包括对特征在通道维度上进行最大池化和平均池化，然后将这两个池化特征进行拼接，经过一个核为7的卷积和Sigmoid激活函数，最后与最初的局部特征进行哈达玛乘积（Hadamard Product）。为了进一步利用来自全局特征流的全局信息，模型中加入了通道注意力机制。具体操作如下：首先，对全局特征在空间维度上进行最大池化（Max Pooling）和平均池化（Average Pooling）操作。接着，通过两个卷积操作对经过池化后的特征进行特征加和。然后，经过Sigmoid激活函数后与最初的全局特征进行哈达玛乘积。最后，为了避免引入过多的参数和增加模型复杂度，本文选择将空间注意力局部特征和通道注意力全局特征相加，并将其输入到多尺度上下文聚合解码器中，从而得到最终的分割结果。具体的计算过程如下公式所示：

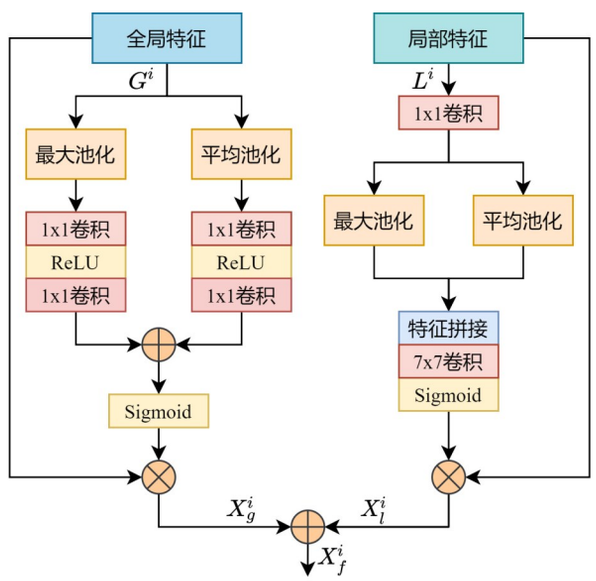


图3　局部-全局特征融合模块具体细节

Fig.3　The detail of LGFF

$$\boldsymbol{X}_l^i=\boldsymbol{L}^i\odot\mathrm{SA}(\boldsymbol{L}^i) \tag{6}$$

$$\boldsymbol{X}_g^i=\boldsymbol{G}^i\odot\mathrm{CA}(\boldsymbol{G}^i) \tag{7}$$

$$\boldsymbol{X}_f^i=\boldsymbol{X}_l^i\oplus\boldsymbol{X}_g^i \tag{8}$$

其中$\boldsymbol{L}^i$代表局部特征流第i层的特征，$\boldsymbol{G}^i$代表全局特征流第i层的特征，$\boldsymbol{X}_l^i$和$\boldsymbol{X}_g^i$分别代表经过注意力机制后第i层输出的局部特征和全局特征。SA表示空间注意力机制，CA表示通道注意力机制。$\odot$指哈达玛乘积算子，$\oplus$指矩阵加法运算。

2.4　上下文聚合解码器

为了得到最后的分割结果，需要对经过LGFF模块输出的融合特征进行解码。本文使用类似FPN的特征融合策略。考虑到FPN解码结构仅仅

www.ygxb.ac.cn
SENSING BULLETIN
NATIONAL REMOTE SENSING BULLETIN 遥感学报

使用最后一层的输出特征作为解码输出特征，可能会丢失其他层级的上下文语义特征，本文额外对每层金字塔解码特征进行加和操作。图1的最右侧是上下文聚合解码器部分，其中左半部分是常规的FPN结构，右半部分是改进之处。具体来说，首先，对经过LGFF模块融合后的四个特征进行带有批归一化（Batch Normalization，BN）和ReLU激活函数的卷积层处理，以进一步挖掘这些融合特征的表征能力。其次，为了使不同层级之间的特征在通道上达成一致以进行特征加和，使用相同的卷积操作进行特征维度变换。然后，使用上采样操作使不同层级之间的特征尺度一致，并对这些尺度相同的特征进行加和。接下来，对各层不同尺度的上下文特征进行带有批归一化和ReLU激活函数的卷积操作，以统一各层级的特征维度。最后，通过加和操作将这些维度相同的各层级上下文特征进行融合，并将其送入语义分割头中生成最终的结果。

2.5 损失函数

为了综合各个像素分类准确度、像素间类平衡以及边缘像素提取等多个方面进行优化，本文使用联合损失函数进行模型训练。该联合损失函数由交叉熵损失、Dice损失和边缘损失组成，具体定义如下：

$$\mathcal{L} = \alpha \mathrm{CE}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) + \beta \mathrm{DICE}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) + \gamma \mathrm{BCE}\left(\boldsymbol{Y}_b, \hat{\boldsymbol{Y}}_b\right) \tag{9}$$

$$\boldsymbol{Y}_b = \mathrm{Lap}(\boldsymbol{Y}) \tag{10}$$

$$\mathrm{CE}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} \hat{\boldsymbol{Y}}_{ic} \log\left(\boldsymbol{P}_{ic}\right) \tag{11}$$

$$\mathrm{DICE}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{12}$$

$$\mathrm{BCE}\left(\boldsymbol{Y}_b, \hat{\boldsymbol{Y}}_b\right) = \frac{1}{N}\sum_{i} -\left[\hat{\boldsymbol{Y}}_i \log(\boldsymbol{P}) + \left(1 - \hat{\boldsymbol{Y}}_i\right)\log(1 - \boldsymbol{P})\right] \tag{13}$$

上式中，α、β和γ分别为三个损失函数的平衡因子。在本文中，它们都被设置为1。$\boldsymbol{Y}$代表预测标签，$\hat{\boldsymbol{Y}}$代表真实标签，$\boldsymbol{Y}_b$和$\hat{\boldsymbol{Y}}_b$指经过核为$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$的拉普拉斯卷积（Lap）（Fan等，2021）后所得到的关于建筑物的边缘和真实标签。CE代表交叉熵损失函数，目的是使模型能够更好地区分建筑物像素和背景像素，其中N指样本总数，M为类别数量，在本文中M为2，$\hat{\boldsymbol{Y}}_{ic}$指第$i$个样本的独热编码标签，值为1或0，$\boldsymbol{P}_{ic}$指第$i$个样本类别数为$c$的概率。DICE指Dice损失函数，目的是应对数据集中背景像素和建筑物像素之间可能存在的类别不平衡问题。为解决提取结果中模糊的建筑物边缘问题，本文用BCE损失函数对提取出的边缘进行约束，使模型在训练时关注建筑物边缘的像素。

3 实验与分析

3.1 实验数据集

为验证所提出方法的有效性，将在三组公开数据集进行测试。

（1）武汉大学建筑物数据集（WHU Building Dataset）（Ji等，2018）。该套数据包括从空间分辨率为0.075m、面积为450平方公里的航空影像中提取出的22万多座独立建筑，主要覆盖区域为新西兰Christchurch。该数据集原始影像空间分辨率被下采样到0.3m，并且被切分为8189张尺寸大小为512×512像素的影像图片，其中训练集包含4736张（130500座建筑），验证集包含1036张（14500座建筑）以及测试集包含2416张（42000座建筑）。

（2）Inria航空影像标注数据集（Maggiori等，2017）。该套数据包含360张精细的航空遥感影像，采集区域来自5个城市（Austin，Chicago，Kitsap，Tyrol，and Vienna）。由于测试数据集的标签不是公开可用的，本文只使用原始数据集。按照官方的划分，每个城市的1–5副图片被挑选出作为验证集而剩下的作为训练集。本次实验中，首先填充原来尺寸为5000×5000像素的图像到5120×5120像素然后再裁剪到512×512像素。为了高效训练，裁剪后的图片中不包含建筑物目标的部分被人为移除。最后，分别生成了9737张训练图像和1942张验证图像。

（3）Massachusetts建筑物数据集。该套数据包含151张美国波士顿市的航空影像，每张影像的尺寸为1500×1500像素，空间分辨率为1m。数据集被划分为三部分，其中训练集包含137张影像，验证集包含4张影像，测试集包含10张影像。本文遵循Wang等（2022）对此数据集的处理策略，使

© 《遥感学报》

用一些数据增强策略例如垂直和水平方向翻转进一步扩充了训练影像，最后得到了411张训练影像，4张验证影像以及10张测试影像。

3.2 实验评价指标

本文使用交并比（IoU）、精确率（Precision）、召回率（Recall）和F1分数（F1）来验证所提出模型的性能。IoU表示预测标签中建筑物像素与真实标签建筑物像素的重合程度。Precision表示被正确预测的建筑物像素与所有预测为建筑物像素的比值。Recall表示正确预测的建筑物像素与真实标签中所有建筑物像素的比值。F1是精确率和召回率的调和平均。具体计算公式如式（14）—（17）所示。

$$IoU = \frac{TP}{TP + FP + FN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

TP表示预测的建筑物被正确识别为建筑物的像素数量；FP表示预测的建筑物被错误地识别为建筑物的像素数量；TN表示预测的非建筑物被正确识别为非建筑物的像素数量；FN表示真实的建筑物被错误地识别为非建筑物的像素数量。

3.3 实验设置

实验硬件环境使用配有两张NVIDIA RTX3090 GPU的服务器。使用Pytorch深度学习框架，并采用端到端监督学习的方式来训练模型。为了训练模型，使用AdamW优化器和余弦学习率调整策略。在进行模型训练之前，对数据使用随机水平、垂直翻转以及归一化等数据增强策略。对于武汉大学建筑物数据集，最大训练轮数设置为255，学习率设置为0.001，批次大小设置为16。对于Inria和Massachusetts数据集，使用在武汉大学建筑物数据集上的预训练权重精调模型，设置学习率为0.0005，最大训练轮数为105。

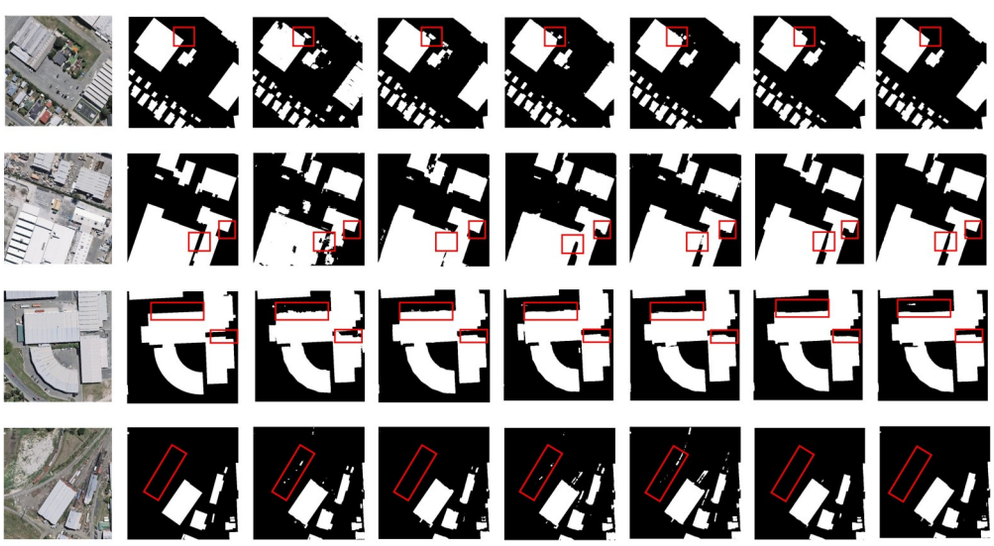


图4 UNet，SegNet，DeeplabV3+，SwinUperNet，STT和ILGS-Net在武汉大学建筑物数据集上的可视化结果

Fig4 Visualized results of the UNet，SegNet，DeeplabV3+，SwinUperNet，STT and ILGS-Net on the WHU Building dataset

3.4 建筑物提取

为了验证所提出模型的有效性，本文在上述介绍的3个公开数据集上，分别将所提出方法的结果与目前最先进的一些建筑物提取方法进行对比。这些方法包括UNet（Ronneberger等，2015）、SegNet（Badrinarayanan等，2017）、DeepLabV3+（Chen等，2018）、MAFCN（Wei等，2019），SwinUperNet（Liu等，2021）、MANet（Li等，2021）、MSST-Net（Yuan等，2021）、STT（Chen等，2021）、BOMSC-Net（Zhou等，2022）、B-

© 《遥感学报》

FGC-Net（Wang 等，2022）、CBRNet（Guo 等，2022）、DCSwin（Wang 等，2022），其中 UNet、SegNet、DeepLabV3+、MSST-Net、B-FGC-Net 和 BOMSC-Net 是 CNN 模型，其余的属于 Transformer 模型。

表1 武汉大学建筑物数据集上与最先进方法的对比

Table 1 Comparison with SOTA Methods on the WHU Building Dataset

方法	IoU	Precision	Recall	F1
UNet	88.48	93.27	94.51	93.89
SegNet	89.28	95.14	93.55	94.34
DeepLabV3+	87.67	92.94	93.92	93.43
SwinUperNet	90.34	93.61	**96.28**	94.93
MSST-Net	88.00	–	–	88.20
STT	90.48	–	–	94.97
B-FGC-Net	90.04	95.03	94.49	94.76
BOMSC-Net	90.15	95.14	94.50	94.80
ILGS-Net	**91.55**	**95.77**	95.41	**95.59**

表2 Inria航空影像标注数据集上与最先进方法的对比

Table 2 Comparison with SOTA Methods on the Inria Aerial Image Labeling Dataset

方法	IoU	Precision	Recall	F1
UNet	70.78	85.18	80.72	82.89
SegNet	76.32	87.64	84.09	85.83
DeepLabV3+	76.80	87.35	86.40	86.88
SwinUperNet	79.53	87.55	**89.67**	88.60
STT	79.42	–	–	87.99
CBRNet	81.10	89.93	89.20	89.56
B-FGC-Net	78.18	87.82	89.12	88.46
BOMSC-Net	78.18	87.93	87.58	87.75
ILGS-Net	**81.66**	**90.52**	89.31	**89.91**

（1）武汉大学建筑物数据集：对于以上方法，本文首先在武汉大学建筑物数据集上进行提取，结果见图4所示，图中列依次为原图、原图标签、UNet 模型提取结果、SegNet 模型提取结果、DeepLabV3+模型提取结果、SwinUperNet 提取结果、STT 模型提取结果以及本文 ILGS-Net 模型提取结果。红框标记了各模型提取结果中明显的特征差异。分析标记可知，第一行显示本文模型具有最清晰的建筑物角点，第二行显示本文模型更好地提取了建筑物的整体特征，第三行显示本文模型有更清晰的建筑物边缘，最后一行显示本文模型能够更好地抵抗其他非建筑物的干扰，这充分证明本文所提模型的强大优势。此外，本文还对武汉大学建筑物数据集进行了精度指标的定量分析。从表1可以看出，本文提出的模型在IoU达到91.55%，Precision达到95.77%，Recall达到95.41%，F1达到95.59%，各项指标均超过了除SwinUperNet以外的其他方法。虽然SwinUperNet在Recall指标上表现最好，但其在其他指标方面比本文提出的模型要低。本文模型在IoU、Precision、Recall以及F1上分别比以上最好的CNN模型BOMSC-Net要高1.4%、0.63%、0.91%以及0.79%，比以上最好的Transformer模型SwinUperNet要高1.21%、2.16%、−0.87%以及0.63%。上述对比结果显示，本文提出的ILGS-Net模型在整体上优于仅使用CNN模型或Transformer模型的方法。

表3 Massachusetts数据集上与最先进方法的对比

Table 3 Comparison with SOTA Methods on the Massachusetts Dataset

方法	IoU	Precision	Recall	F1
UNet	67.61	79.13	82.29	80.68
SegNet	66.57	82.40	77.60	79.93
DeepLabV3+	69.23	84.73	79.10	84.93
MAFCN	73.80	87.07	82.89	84.93
CBRNet	74.55	86.50	84.36	85.42
MANet	70.76	82.00	84.66	83.86
DC-Swin	72.59	83.07	85.19	84.12
BOMSC-Net	74.71	86.64	83.68	85.13
ILGS-Net	**75.75**	**87.23**	**85.20**	**86.20**

（2）Inria建筑物数据集：为了验证ILGS-Net的泛化性和稳定性，本文使用Inria建筑物数据集进行实验。实验结果如图5所示，对于所提取结果，

在IoU、Precision、Recall和F1等指标上进行了定量分析。从表2可以看出，大多数所对比方法都表现良好，交并比均高于75%。而本文方法的交并比、精确率、召回率和F1分数分别为81.66%、90.52%、89.31%和89.91%。除召回率外，这些指标均优于其他模型。与最近提出的CNN模型CBRNet相比，本文的交并比高出0.56%，精确率高出0.59%，召回率高出0.11%，F1分数高出0.35%。相较于对比模型中表现最好的

© 《遥感学报》

Transformer模型SwinUperNet，本文方法的交并比提高了2.13%，精确率提高了2.97%，

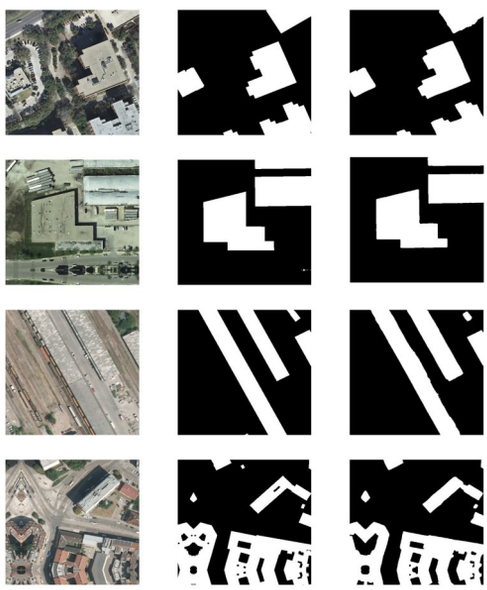


图5　ILGS-Net模型在Inria航空数据集上的结果

Fig.5　Predicted results of ILGS-Net on Inria dataset

原图 标签 ILGS-Net

F1分数提高了1.31%。

（3）Massachusetts建筑物数据集：Massachusetts数据集中建筑物有着更为复杂的形状、颜色和纹理，并且可用于训练的样本较少。提取结果如图6所示。对于所提取出的结果，对各个模型在不同指标上进行定量分析。从表3中可以看出，本文方法在此数据集上的提取结果在交并比、精确率、召回率和F1分数方面分别为75.75%，87.23%，85.20%和86.20%，优于所对比的CNN和Transformer模型。

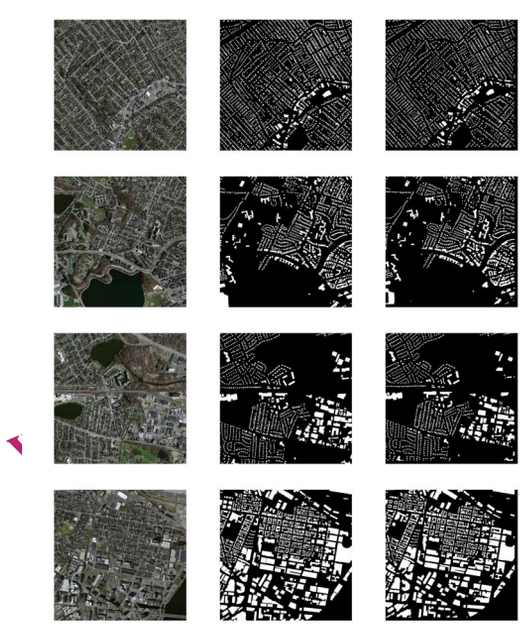


图6　ILGS-Net模型在Massachusetts数据集上的结果

Fig.6　Predicted results of ILGS-Net on Massachusetts dataset

原图 标签 ILGS-Net

对以上结果进行分析，本文模型性能优势主要表现在以下几个方面：首先，与仅提取局部空间特征的CNN模型或仅建模特征的长距离依赖关系的Transformer模型相比，本文的模型通过双流网络结构同时提取局部空间特征和长距离特征。此外，还设计了LGFF模块，通过空间注意力机制和通道注意力机制多层次地融合提取出的局部特征和全局特征。融合后的特征不仅能够使模型更准确地定位高分辨率遥感影像中的建筑物对象，而且使模型能够更好地抵抗非建筑物的干扰，集中关注建筑物对象的整体结构。最后，在训练过程中，模型使用边缘损失进行约束，可更好提取建筑物边缘特征。

表4　模型针对局部特征流和全局特征流的消融实验

Table 4　The ablation study of LFS and GFS

方法	交并比	F1
BuildFormer	90.31	94.91
EfficientNet-b3	89.97	93.64
ResNet+BuildFormer	91.03	95.32
EfficientNet-b3+Swin-T	91.36	95.43
ILGS-Net	**91.55**	**95.59**

表5　模型针对融合模块的消融实验

Table 5　The ablation study of LGFF

方法	交并比	F1
直接特征加和	90.75	95.15
使用LGFF	**91.55**	**95.59**

3.5　消融实验

为了验证所提出模型中各模块的有效性，在武汉大学建筑物数据集上进行消融实验。

（1）局部特征流和全局特征流的有效性：在提出的ILGS-Net模型中，局部特征流用于提取多层次

的局部空间细节特征，全局特征流则用于提取遥感影像中建筑物的长距离依赖特征，以实现

© 《遥感学报》

更好的建

表6 针对联合损失函数的平衡因子进行消融实验

Table 6 The ablation study for balance fators of the join loss

α	β	γ	交并比	F1
1	1	0	91.35	95.50
1	1	0.5	91.46	95.49
1	**1**	**1**	**91.55**	**95.59**
1	1	5	91.02	95.30
1	0.5	1	90.94	95.26
0.5	1	1	90.49	95.01

表7 各模型参数和推理速度对比

Table 7 Comparison of model parameters and inference speed

模型	参数	推理速度(FPS)
UNet	**28.99M**	**43.96**
DeepLabV3+	41.22M	35.58
SwinUperNet	59.83M	29.77
BuildFormer	37.90M	23.17
ILGS-Net	47.50M	18.56

筑物分割效果。为了测试局部特征流的有效性，本实验移除此模块。从表4可以看出，去除局部特征流模块后，IoU下降了1.24%，F1下降了0.68%，这证明了此模块的有效性。接着，为了测试全局特征流模块的有效性，本实验移除BuildFormer全局特征流。去除全局特征流后，IoU下降了1.58%，F1下降了1.95%，这证明了此模块的有效性。此外，为了验证EfficientNet-b3局部特征流的优势，本文选择ResNet作为局部特征提取流。从表4的第三行和最后一行可以看出，选择ResNet作为局部特征流时，交并比和F1分数较EfficientNet-b3更低。同时，为了验证BuildFormer全局特征流的优势，本文选择Swin-T模型（Liu等，2021）作为全局特征流进行对比。对比表4第四行和最后一行，本文模型具有更高的交并比和F1分数。上述结果还表明，双流结构相比单流结构具有更优秀的性能。

（2）局部-全局特征注意力融合模块的优势：在所提出的模型中，LGFF模块能够多层次地融合局部特征和全局特征，使得提取出的建筑物具有更少的内部孔洞以及更为完整的轮廓。为了验证此模块中多层次融合的有效性，本文改变此模块的结构，不对全局特征进行通道注意力机制操作，也不对局部特征进行空间注意力操作，而是选择直接对全局特征和局部特征进行加和。表5显示，使用直接加和操作进行特征融合相比使用LGFF模块进行特征融合，交并比下降了0.8%，F1分数下降了0.44%，这证明了所提出的融合模块的有效性。

（3）联合损失函数平衡因子的影响：为选择适合ILGS-Net模型的联合损失函数，本文对平衡因子α、β和γ进行了不同值的设置。首先，为了验证提出的边缘损失的重要性，固定α和β的值为1，并为γ设置多组不同值。从表6可以看出，当γ为1时，模型表现出最佳性能。当γ过大或过小时，模型性能都下降，这是由于过于关注建筑边缘会降低模型对建筑物像素的分类准确度，而对建筑边缘关注不足或忽视将导致模型提取出模糊的建筑边缘。其次，为了验证Dice损失函数的影响，本文将β设为0.5。从表6可以看出，模型性能有所下降，这突显了Dice损失的重要性。最后，为了验证交叉熵损失函数的影响，本文将α设为0.5。从表8可以看出，模型性能下降明显，这表明交叉熵损失在联合损失中起主导作用。

（4）模型复杂度和时效性分析：为了验证所提出模型的高效性，本文将其与一些先进模型基于模型参数量和推理速度进行对比。从表7可以看出，相比于只有单流结构的CNN模型UNet和DeepLabV3+，本文模型具有更多的参数和更慢的推理速度。与其他Transformer模型相比，虽然参数量少于SwinUperNet，但推理速度也相对较慢。通过分析，本文发现ILGS-Net模型的低帧率主要是由于双流结构的性质所致，模型的总推理时间是在两个流上推理时间之和。虽然双流结构使模型的推理速度较低，但本文的目标是探索CNN和Transformer模型相结合以提高建筑物提取性能的可能性，表1和表4的实验结果有力地证明了这种结合方式的有效性。上述实验结果和对比分析为本文未来的研究方向提供了思路，即在保证提取精度的同时，探索采用一些轻量级网络实现双流结构模型的可能性。

4 结论

为了提高深度学习模型对建筑物地物的提取

© 《遥感学报》

精度，本文提出一个融合局部特征和全局特征的双流建筑物提取模型ILGS-Net。模型的编码部分利用全局特征流生成具有长距离依赖关系的全局特征，并通过局部特征流提取具有定位对象精确位置能力的空间细节特征。为了更好建模全局特征和局部特征之间的语义关系，模型引入局部-全局特征融合模块。在解码部分，模型使用上下文聚合模块来多层级聚合融合特征。此外，本文还引入边缘损失函数来约束模型对建筑物边缘细节的关注。实验证明，与最先进的方法相比，ILGS-Net模型具有出色建筑物目标精细提取和边缘感知能力。为验证算法的适用能力，本文在Inria建筑物数据集和Massachusetts数据集上训练了模型。实验结果优于当前最先进的一些模型，展示了该模型对不同数据的适应性。在武汉大学建筑物数据集上的消融实验充分证明了模型中各组件设计的有效性。展望未来的研究方向，一方面，考虑将此模型应用于其他遥感任务，如多类地物提取、道路检测、变化检测等，以进一步验证其泛化能力和拓展其通用性。另一方面，考虑借助一些轻量级模型来提高模型推理速度。

参考文献（References）

Badrinarayanan V, Kendall A and Cipolla R. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12): 2481-2495. [DOI: 10.1109/TPAMI.2016.2644615]

Chen K Y, Zou Z X and Shi Z W. 2021. Building Extraction from Remote Sensing Images with Sparse Token Transformers. Remote Sensing, 13(21): 4441. [DOI: 10.3390/rs13214441]

Chen L C, Zhu Y, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation//15th European Conference on Computer Vision. Munich: Springer: 833-851. [DOI: 10.1007/978-3-030-01234-2_49]

Cooner A J, Shao Y and Campbell J B. 2016. Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. Remote Sensing, 8(10): 868. [DOI: 10.3390/rs8100868]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020). [DOI: arXiv:2010.11929]

Fan M Y, Lai S Q, Huang J S, Wei X M, Chai Z H, Luo J F and Wei X L. 2021. Rethinking BiSeNet for Real-Time Semantic Segmentation//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE: 9711-9720. [DOI: 10.1109/CVPR46437.2021.00959]

Guo H N, Du B, Zhang L P and Su X. 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 183: 240-252. [DOI: 10.1016/j.isprsjprs.2021.11.005]

Ji S, Wei S and Lu M. 2018. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. IEEE Transactions on Geoscience and Remote Sensing, 57(1): 574-586. [DOI: 10.1109/TGRS.2018.2858817]

Li R, Zheng S Y, Zhang C, Duan C X, Su J L, Wang L B and Atkinson P M. 2021. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-13. [DOI: 10.1109/TGRS.2021.3093977]

Li X H, Bai X C, Li Z J, Zuo Z Y. High-Resolution Image Building Extraction Based on Multi-level Feature Fusion Network. Geomatics and Information Science of Wuhan University. 2022, 47(8): 1236-1244.

李星华, 白学辰, 李正军, 左芝勇 . 面向高分影像建筑物提取的多层次特征融合网络 . 武汉大学学报 (信息科学版). 2022. 47(8): 1236-1244 [DOI: 10.13203/j.whugis20210506]

Lin N, Huang T, Sun P L and Wang Y Y. 2022. Building Extraction of High-resolution Remote Sensing Imagery on Optimized Mask-RCNN. Remote Sensing Information, 03:37.

林娜, 黄韬, 孙鹏林, 王玉莹 . 2022. 优化 Mask-RCNN 的高分遥感影像建筑物提取 . 遥感信息 . 003: 037 [DOI: 10.3969/j.issn.1000-3177.2022.03.001]

Lin T Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//2017 IEEE conference on computer vision and pattern recognition. Honolulu: IEEE: 936-944. [DOI: 10.1109/CVPR.2017.106]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: Hierarchical vision transformer using shifted windows//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE: 9992-10002. [DOI: 10.1109/ICCV48922.2021.00986]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//2015 IEEE conference on computer vision and pattern recognition. Boston: IEEE: 3431-3440. [DOI: 10.1109/CVPR.2015.7298965]

Lyu S Y, Li J T, A X H, Yang C, Yang R C and Shang X M. Res＿ASPP＿UNet++: Building an extraction network from remote sensing imagery combining depthwise separable convolution with atrous spatial pyramid pooling, 27(02): 502-519

吕少云, 李佳田, 阿晓荟, 杨超, 杨汝春, 尚晓梅 . 2023. Res＿ASPP＿UNet++：结合分离卷积与空洞金字塔的遥感影像建筑物提取网络 . 遥感学报, 27(02): 502-19

Maggiori E, Tarabalka Y, Charpiat G and Alliez P. 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark//2017 IEEE International Geoscience and Re-

© 《遥感学报》

mote Sensing Symposium (IGARSS). Fort Worth: IEEE: 3226-3229. [DOI: 10.1109/IGARSS.2017.8127684]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation//Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference. Munich: Springer: 234-241. [DOI: 10.1007/978-3-319-24574-4_28]

Strudel R, Garcia R, Laptev I and Schmid C. 2021. Segmenter: Transformer for semantic segmentation//2021 IEEE/CVF international conference on computer vision. Montreal: IEEE: 7242-7252. [DOI: 10.1109/ICCV48922.2021.00717]

Tan M X and Le Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks//36th International Conference on Machine Learning. PMLR: 6105-6114. [DOI: 10.48550/arXiv.1905.11946]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez N A, Kaiser Ł and Polosukhin I. 2017. Attention is all you need. arxiv. [DOI: 10.48550/arXiv.1706.03762]

Wang L B, Fang S H, Meng X L and Li R. 2022. Building Extraction With Vision Transformer. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-11. [DOI: 10.1109/tgrs.2022.3186634]

Wang L B, Li R, Duan C X, Zhang C, Meng X L and Fang S H. 2022. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 19: 1-5. [DOI: 10.1109/lgrs. 2022. 3143368]

Wang Y, Zeng X Q, Liao X H and Zhuang D F. 2022. B-FGC-Net: A Building Extraction Network from High Resolution Remote Sensing Imagery. Remote Sensing, 14(2): 269. [DOI: 10.3390/rs14020269]

Wei S Q, Ji S P and Lu M. 2019. Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization. IEEE Transactions on Geoscience and Remote Sensing, 58(3): 2178-2189. [DOI: 10.1109/tgrs.2019.2954461].

Xiao X, Guo W L, Chen R, Hui Y L, Wang J N and Zhao H Y. 2022. A Swin Transformer-Based Encoding Booster Integrated in U-Shaped Network for Building Extraction. Remote Sensing, 14 (11): 2611. [DOI: 10.3390/rs14112611]

Xu Z S, Guan H Y, Yu Y T, Lei X D and Zhao H H. 2022. A dual-attention capsule network for building extraction from high-resolution remote sensing imagery. Journal of Remote Sensing, 26(08): 1636-49

许正森, 管海燕, 彭代锋, 于永涛, 雷相达, 赵好好. 2022. 高分辨率遥感影像建筑物提取的注意力胶囊网络算法. 遥感学报, 26(08): 1636-1649 [DOI: 10.11834/jrs.20221577]

Yuan W and Xu W B. 2021. MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer. Remote Sensing, 13(23): 4743. [DOI: 10.3390/rs13234743]

Zhou Y, Chen Z L, Wang B, Li S J, Liu H, Xu D Z and Ma Chao. 2022. BOMSC-Net: Boundary Optimization and Multi-Scale Context Awareness Based Building Extraction From High-Resolution Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-17. [DOI: 10.1109/tgrs.2022.3152575]

Zhu X X, Tuia D, Mou L C, Xia G S, Zhang L P, Xu F and Fraundorfer F. 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geoscience and Remote Sensing Magazine, 5(4): 8-36. [DOI: 10.1109/mgrs.2017.2762307]

Integration of CNN and Transformer for High-Resolution Remote Sensing Image Building Extraction: A Dual-Stream Network

LIU Yuxin[1,2], MENG Yu[1], DENG Yupeng[1], CHEN Jingbo[1], LIU Diyou[1]

1.National Engineering Research Center for Geoinformatics, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;

2.2. School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Convolutional Neural Networks (CNNs) and Transformers have emerged as pivotal tools in the realm of building extraction tasks within high-resolution remote sensing images. While these techniques have seen widespread application, challenges persist for CNNs in effectively modeling long-range spatial dependencies, often leading to complications such as the emergence of internal holes in the extracted building structures. Conversely, Transformers exhibit limitations in capturing spatial local details, potentially resulting in the production of blurry building edges and the oversight of smaller structures. In response to these challenges, this paper presents an innovative dual-stream network model tailored for building extraction in high-resolution remote sensing images, denominated as ILGS-Net (Network for the Integration of Local and Global Features Stream). ILGS-Net is designed to capitalize on the strengths of both CNNs and Transformers. The model incorporates multi-level local-global feature fusion modules to seamlessly blend intricate local details and expansive global context features of buildings. In tandem, an edge loss function is integrated into the objective function, contributing to the refinement of building boundary localization precision. The proposed ILGS-Net endeavors to address the shortcomings of existing methodologies by efficiently combining the unique attributes of CNNs and Transformers. Multi-level local-global feature fusion modules

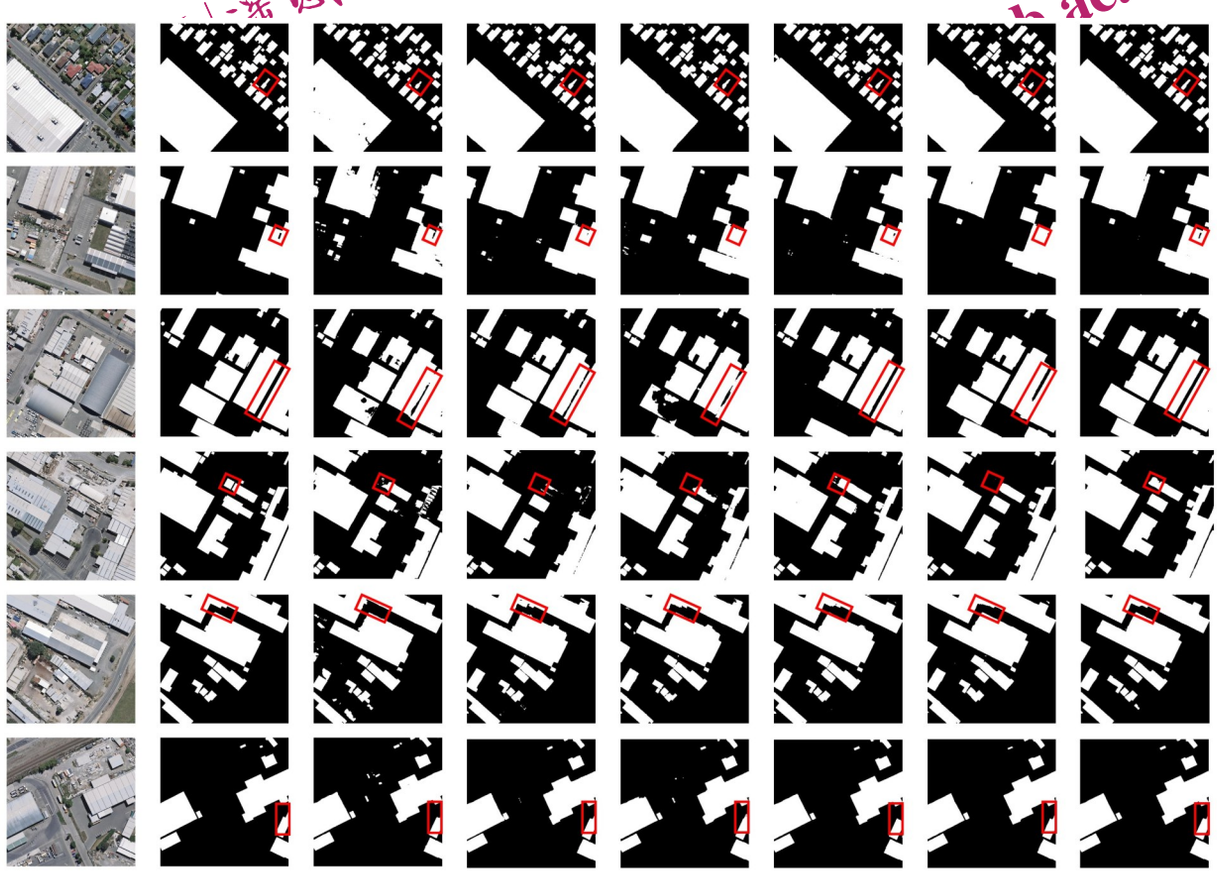
© 《遥感学报》

play a pivotal role in striking a harmonious balance between capturing fine-grained local details and incorporating broader global context features of buildings. Simultaneously, the inclusion of an edge loss function serves as a guiding mechanism in model training, augmenting the precision of building boundary localization.Extensive experiments conducted across three high-resolution building datasets consistently demonstrate the superior performance of the proposed ILGS-Net compared to benchmark methods outlined in this paper. Notably, the proposed method achieves, on average, a remarkable 1% increase in Intersection over Union (IoU) across all three datasets.In conclusion, ILGS-Net emerges as a groundbreaking dual-stream network model expressly designed for building extraction in high-resolution remote sensing images. By seamlessly integrating CNNs and Transformers, along with the implementation of multi-level local-global feature fusion and the inclusion of an edge loss function, the model adeptly addresses challenges associated with spatial dependencies and local details, resulting in a marked improvement in the accuracy of building extraction. The experimental results underscore the efficacy of the proposed method, positioning it as a promising and influential approach for achieving high-precision building extraction in high-resolution remote sensing images. The confluence of advanced methodologies and innovative techniques within ILGS-Net marks a significant stride forward in the field of remote sensing image analysis. As technology continues to evolve, ILGS-Net represents a pivotal contribution that holds promise for further advancements in building extraction accuracy, providing a solid foundation for continued research and application in the realm of high-resolution remote sensing imagery analysis.Looking ahead, the success of ILGS-Net prompts further exploration and research avenues. Investigating the potential of similar integrative approaches in other remote sensing tasks holds promise. Additionally, refining and expanding the current model architecture to accommodate varying scales and complexities of urban landscapes is a logical progression. Future work should focus on translating these advancements into tangible benefits for decision-makers and stakeholders in urban development and disaster response.

Key words: building extraction, deep learning, dual-stream network, local-global feature fusion

Supported by Supported by National Key R&D Program of China(2021YFB3900503)

附录



附图 1　UNet，SegNet，DeeplabV3+，SwinUperNet，STT 和 ILGS-Net 在武汉大学建筑物数据集上的可视化结果

Fig.1 Visualized results of the UNet，SegNet，DeeplabV3+，SwinUperNet，STT and ILGS-Net on the WHU Building datase

© 《遥感学报》