FLGF-UNet：融合局部—全局特征的光学遥感图像遥感建筑物提取网络

李国燕，刘涛，王丽，刘毅

天津城建大学 计算机与信息工程学院，天津 300384

摘 要：遥感图像的语义分割在城市变化检测、环境保护、地质灾害识别等领域具有重要作用。针对当前遥感建筑物提取中存在的漏检、误检、因树木遮挡或类似物体干扰导致提取不完整等问题，本文基于UNet网络提出一种改进的建筑物提取网络—融合局部—全局特征网络FLGF-UNet（Fusion of Local Global Features Network）。FLGF-UNet的并行特征融合方式确保每个阶段的特征都包含细粒度的局部信息和全局依赖，使得网络在每一阶段的特征表示中同时具备局部和全局信息，有效克服Transformer在局部信息交换上的不足，同时在全局信息建模方面优于传统CNN。此外，为弥补编码器和解码器之间的语义鸿沟，编解码器之间加入交互融合IF（Interactive Fusion）模块，增强空间细节、全局上下文和语义特征的融合效果。为验证FLGF-UNet的优越性和通用性，在WHU、Massachusetts数据集和中国典型城市建筑物实例数据集上，将所提网络与U2Net、Swin Transformer、MA-Net、HD-Net和RS-Mamba等网络进行对比。结果表明，FLGF-UNet在性能上优于其他SOTA网络，具有较高的实际应用价值。

关键词：遥感图像，建筑物提取，融合局部—全局特征网络，特征融合，交互融合模块

中图分类号：P2

引用格式：李国燕，刘涛，王丽，刘毅.2026.FLGF-UNet：融合局部—全局特征的光学遥感图像遥感建筑物提取网络.遥感学报，30(3)：696-709

Li G Y，Liu T，Wang L and Liu Y. 2026. FLGF UNet：Remote sensing building extraction network for optical remote sensing images that integrates local-global features. National Remote Sensing Bulletin，30(3)：696-709

[DOI：10.11834/jrs.20264516]

1 引 言

建筑物提取的主要目的是从高分辨率遥感图像中区分出建筑物的形状。过去几十年来，该领域取得显著进展，广泛应用于城市规划、城市动态监测和灾害监测等领域。早期研究主要依赖手工标记的建筑特征（如形状、边缘和阴影）进行二分类，这些特征一般被认为是低级特征，且高度依赖先验知识。为从大量数据中自适应地学习高层特征信息（Wang等，2022），近年来深度学习技术被引入到建筑物提取中，取得良好的效果，已经成为研究热点。

在深度学习方法中，卷积神经网络CNN（Convolutional Neural Network）作为一种强大的特征学习网络，被广泛应用于各种计算机视觉任务。Long等（2015）基于CNN提出全卷积网络FCN（Fully Convolutional Network），其优点是能够灵活接受任意大小的输入图像，并有效生成相应大小的输出。例如，Ronneberger等（2015）提出U-Net网络，利用解码器学习编码器阶段图像特征的空间相关性，从而显著提高分割性能。此外，Qin等（2020）设计具有两级嵌套U型结构的U2Net，其中的残差U块（RSU）能够提取不同尺度的感受野，从而捕获更多上下文信息。为解决特征利用率不足和多尺度信息融合不充分的问题，Li等（2022）提出多注意力网络（MANet），通过多个有效的注意力模块提取上下文依赖关系。Li等（2024b）进一步提出高分辨率去耦网络HD-Net，利用其多尺度信息并行交互，显著缓解整体与边界特征之间的耦合问题。尽管CNN网络在建筑物

收稿日期：2024-11-12；**预印本：**2025-08-01

基金项目：国家自然科学基金(编号：52178295)

第一作者简介：李国燕，研究方向为机器视觉、下一代网络技术。E-mail：ligy@tcu.edu.cn

通信作者简介：王丽，研究方向为智能传感。E-mail：wanglichengjian@tcu.edu.cn

提取上取得良好效果，但由于解码器设计的限制，CNN无法充分利用全局信息（Song等，2023）。此外，由于图像中存在复杂的背景和大量噪声，利用全局上下文信息和细致的空间特征，可以有效地进行遥感图像的语义分割（Ding等，2021）。然而，传统CNN网络在遥感图像语义分割方面仍存在不足，亟需进一步改进。

为解决传统CNN网络的这一局限性，采用注意力机制和多尺度特征融合策略，如金字塔池化模块（Zhao等，2017）、多级特征融合策略（Wang等，2018）、通道注意力和位置注意力模块（Fu等，2019）、交叉注意力模块（Chen等，2021）。然而，此类方法捕获的全局信息并不是直接从全局建模中编码捕获的（He等，2022），而是由现有CNN网络捕获的局部特征组成的；因此，全局信息可能尚未被捕获（Mou等，2020）。Transformer在自然语言处理（NLP）领域中捕获全局信息关系的能力强，为语义分割提供一个可行的方案。例如，Carion等（2020）设计双分支Transformer结构，以建模目标和全局图像背景之间的关系为主。此外，Liu等（2021）构建Swin Transformer，在图像分类和密集预测任务中展示巨大的潜力。值得注意的是，尽管Transformer可以有效地捕获全局特征的长距离依赖性关系，但局部特征信息经常被忽视。为有效地利用局部特征和全局特征对遥感图像进行分割，设计一个Transformer和CNN相结合的网络至关重要，进而既可以有效捕获全局信息，也可以捕获精细的空间细节特征（Azad等，2022）。Li等（2024a）提出一种轻量级局部全局特征提取网络LLG-Net（Lightweight Local Global Feature Extraction Network），其中自关注块和卷积块的有效结合保证全局和局部特征跨多个层次的逐步融合。Zhang等（2024）提出一种光谱空间净化网络来提高分类精度，利用全局—局部互导模块实现图像—像素级特征交互，增强提取特征的空间区分度，降低空间异质性。上述方法大多使用CNN分支提取局部特征，Transformer分支提取全局特征，然后在解码器端进行融合，这往往导致多尺度特征融合中的语义冗余，且融合不够充分。最新的研究中，Zhao等（2024）提出一种基于遥感Mamba（RSM）的大规模高分辨率遥感图像密集预测方法专门设计用于捕获具有线性复杂度的遥感影像的全局上下文，从而促进大型遥感影像的有效处理。

鉴于全局和局部信息的充分融合方面存在局限性，并且为解决遥感建筑物图像漏检、误检、被树木遮挡或被类似物体干扰造成提取不完整问题，提出一种融合局部—全局特征网络FLGF-UNet（Fusion of Local Global Features Network），直接整合局部和全局信息，通过多尺度和全局建模并行地获取特征，使得网络每一阶段的特征表示都具备局部和全局信息，能够在特征提取过程中更加自然和充分地融合不同尺度和层次的信息。

FLGF-UNet网络的编码器端采用残差注意力RA（Residual Attention）模块作为初始层，以有效提取浅层特征，并提出线性融合LF（Linear Fusion）模块作为主干，用于捕获空间细节和全局上下文信息。LF模块包含多尺度融合MSF（Multi-scale fusion）和全局注意力建模GM（Global Attention Modeling）模块，分别用于提取不同尺度的局部信息和全局上下文信息，确保特征依赖关系的完整性，从而弥补单一模块在特征提取中的不足。解码器端同样采用残差注意力模块来融合多层特征信息，进一步提高建筑物提取的精确度。此外，编解码器之间引入交互融合IF（Interactive Fusion）模块，以弥补语义鸿沟，增强空间细节、全局上下文和语义特征的融合效果。实验在WHU、Massachusetts数据集和中国典型城市建筑物实例数据集上对FLGF-UNet的有效性和普适性进行验证。

2 融合局部—全局特征网络

建筑物屋顶材质多样，且与道路、广场、裸土等地物的材质类似。因此在高分辨率遥感影像上，建筑物与道路、广场、裸土等地物在光谱上容易混淆。但植被、阴影、水体与建筑物的光谱特征差别较大。建筑物与道路、广场、裸土的均值较为接近。建筑物的方差与道路较为接近。建筑物与道路对比度较为接近。总之，建筑物与道路、广场、裸土在部分纹理特征上出现了相似性，与阴影、植被、水体的纹理差别较大。

由于建筑物在几何形态上具有显著的多尺度差异性（如以及尺寸分布的跨度性），且其清晰的建筑物边界易受复杂环境干扰（如植被遮挡或邻域建筑粘连），这对分割模型提出双重要求：一方面需通过精细的边缘感知机制捕捉像素级轮廓特征，另一方面需建立多层次特征交互机制来处理

目标尺度的剧烈变化。FLGF-UNet针对此类特性创新性地设计线性—交互双模态融合架构：线性融合层通过可学习的参数化通道加权策略实现跨尺度特征的渐进式对齐，有效缓解建筑目标尺度跳变带来的特征失配问题；交互融合模块则通过建立空间—通道双维度的注意力交互机制，在增强局部细节敏感度的同时保持全局结构一致性，弥补编码解码过程的语义鸿沟。

FLGF-UNet总体架构见图1所示。编码器由1个RA模块和4个LF模块组成，解码器由4个RA模块组成。LF模块由3个3 × 3*Conv*卷积层、MSF模块和GA模块组成。FLGF-UNet采用IF模块代替编码器和解码器之间的跳跃连接，从而实现跨层的多尺度融合。当特征图进入IF模块之前，使用*Up-Conv*先对影像进行双线性上采样，再将低分辨率的特征图的通道压缩，去除冗余信息。随后将其馈送到IF模块，以进一步解码上下文信息，IF模块实现编解码器之间的无缝连接。网络的预测输出和标签（GT）之间的差异程度决定损失值。本文采用Sigmoid IoU loss损失函数进行计算。FLGF-UNet将损失函数添加到其5个分支可以监督网络训练，进一步从聚合特征图特征表示。当网络的预测输出与GT的形状尺寸不一致时，考虑到对网络输出进行下采样后会出现语义丢失的情况，对预测输出进行上采样，使其形状尺寸与GT相同，然后使用损失函数计算损失。上采样网络分支可以平滑其噪声，网络输出与GT相似，特征图的分支就可以更好地监督学习。

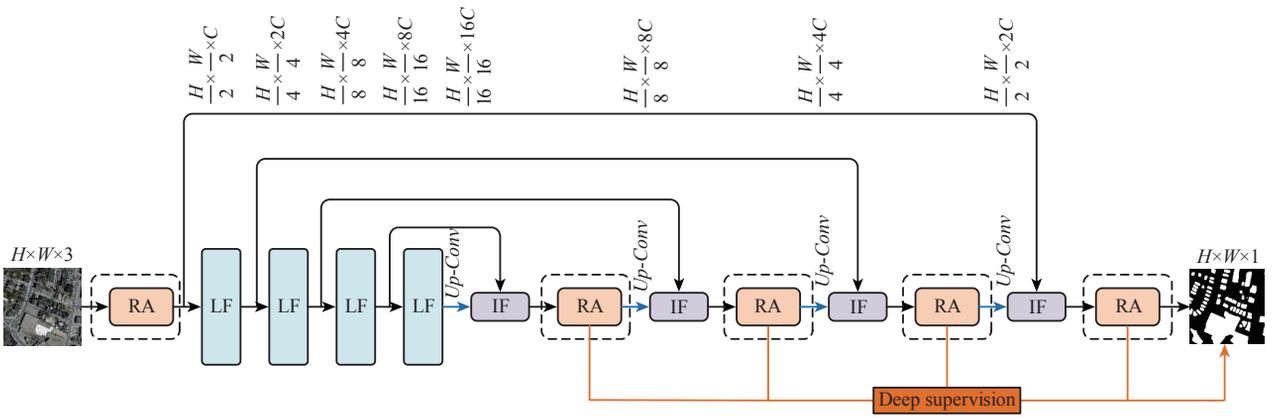

图1 FLGF-UNet网络模型

Fig. 1 FLGF-UNet network model

2.1 残差注意力模块

为应对网络层数增加带来的梯度消失、特征提取不充分以及参数分配不足等问题，本文提出残差注意力模块RA（Residual Attention Module）。该模块的主要组成部分包括残差卷积结构和注意力模块，如图2所示。注意力模块处理按如下顺序执行。特征映射$\boldsymbol{X} \in \boldsymbol{R}^{H \times W \times C}$作为通道注意力的输入，进而生成1-D注意力图$\boldsymbol{X}_1$，并与$\boldsymbol{X}$逐元素相乘以创建通道注意力特征$\boldsymbol{X}'$。通道注意力的表达式如下：

$$\begin{aligned} \boldsymbol{X}_1 &= \sigma\left(\boldsymbol{MLP}(\boldsymbol{X})\right) \\ \boldsymbol{X}' &= \boldsymbol{X}_1 \otimes \boldsymbol{X} \end{aligned} \tag{1}$$

上一步的输出$\boldsymbol{X}'$用作空间注意力的输入，随后生成一个二维空间注意力图$\boldsymbol{X}_2$，与$\boldsymbol{X}'$逐元素相乘，得到空间注意力特征$\boldsymbol{X}''$。其中，σ表示*sigmoid*函数，表示逐元素乘法。空间注意力的表达式如下：

$$\begin{aligned} \boldsymbol{X}_2 &= \sigma\left(\boldsymbol{MLP}\left(\boldsymbol{Conv}\left(\boldsymbol{MLP}\left(\boldsymbol{Conv}(\boldsymbol{X}')\right)\right)\right)\right) \\ \boldsymbol{X}'' &= \boldsymbol{X}_2 \otimes \boldsymbol{X}' \end{aligned} \tag{2}$$

2.2 线性融合模块

传统的卷积神经网络CNN（Convolutional Neural Network）在遥感图像处理中的局限性在于其缺乏全局建模能力，仅能对图像的局部区域进行特征提取，导致在复杂背景下容易受到噪声干扰。此外，由于多次下采样操作，网络深层次中的遥感建筑物特征可能逐渐丢失，从而影响建筑物提取的准确性和完整性。Transformer的引入为遥感图像中的建筑物提取带来新的可能性，其自注意力机制具有强大的全局建模能力，可以有效捕获远距离特征间的依赖关系。然而，遥感建筑物具有边界复杂、形状多样且受背景干扰明显的特点，

单纯依赖全局特征往往不足以准确提取建筑物。因此，局部特征的提取仍然是遥感建筑物提取过程中不可或缺的部分。为克服上述局限，可以设计出线性融合模块见图3所示，既具备全局特征建模能力，又保留局部特征的细节表示，以提高遥感图像中建筑物的提取精度和鲁棒性。LF模块是由3个3×3*Conv*卷积层（捕获大邻域内的上下文语义）、多尺度融合MSF（Multi-Scale Fusion）模块（可以抑制来自无关区域的信息干扰，有选择地强调多尺度特征）和全局注意力GA（Global Attention Modeling）模块（增强所获得的特征图中的全局信息）组成。全局建模模块负责计算注意力矩阵，结合特征重整形和逐点卷积操作，更高效地实现全局与局部特征的融合，保留空间细节。给定一个输入特征 $\boldsymbol{F}\in R^{H\times W\times C}$，分别经过1 × 1卷积后得到 $\boldsymbol{F}_1\in R^{H\times W\times C}$ 和 $\boldsymbol{F}_2\in R^{H\times W\times C}$。然后对 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$ 进行整形，得到 $\boldsymbol{q}\in R^{(H\times W)\times 1}$ 和 $\boldsymbol{k}\in R^{(H\times W)\times 1}$，再通过全连接层对 $\boldsymbol{q}$ 和 $\boldsymbol{k}$ 进行计算和整形，得到 $\boldsymbol{q}\in R^{H\times 1}$ 和 $\boldsymbol{k}\in R^{1\times H}$。在 $\boldsymbol{q}$ 和 $\boldsymbol{k}$ 上进行矩阵乘法，获得注意力 $\boldsymbol{F}_3\in R^{H\times H}$。最后，对注意力进行1 × 1卷积和*softmax*运算，得到输出注意力 $\boldsymbol{F}_4\in R^{H\times H\times C}$。其中1 × 1*Conv*表示逐点卷积，$\boldsymbol{FC}(\cdot)$表示全连接层。该模块可以更直接地学习和优化特征图中目标位置的注意力权重，适应遥感图像中物体复杂的空间分布。相比之下，Transformer在对图像块进行信息交换时，局部空间信息可能会丢失，尤其在单个块内缺乏细节感知。注意力矩阵经过不断的优化学习，能够有效感知目标在特征图中的位置。

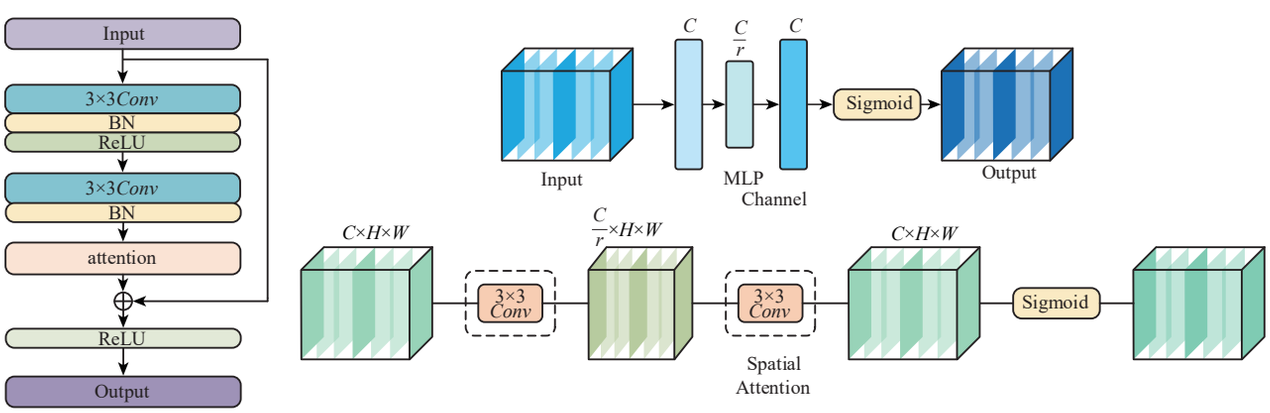

图2　残差注意力模块

Fig. 2　Residual attention module

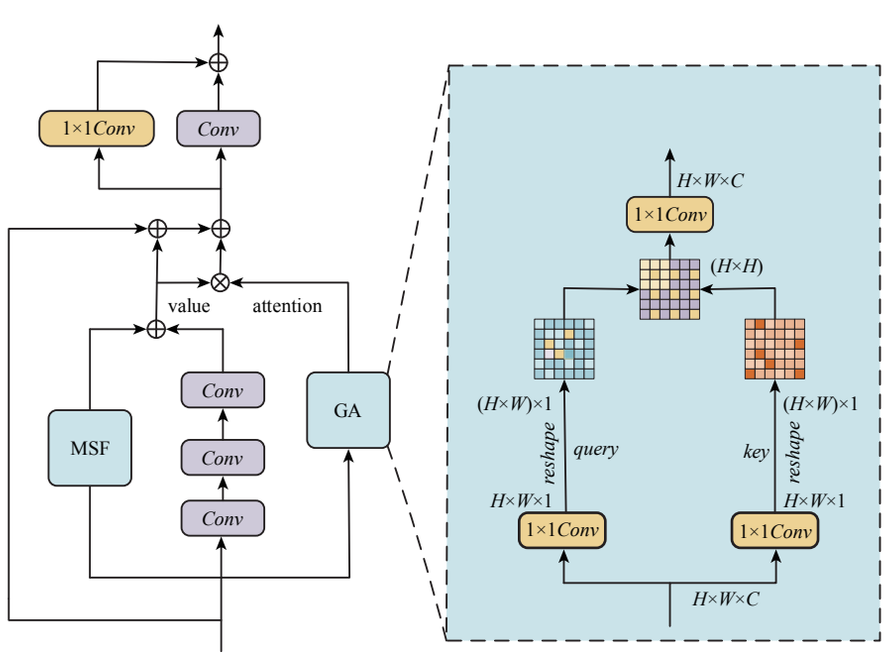

图3　线性融合模块

Fig. 3　Linear fusion module

$$\begin{aligned} \boldsymbol{q} &= \boldsymbol{FC}\left(\boldsymbol{reshape}\left(1\times 1\boldsymbol{Conv}(\boldsymbol{F}_1)\right)\right)\\ \boldsymbol{k} &= \boldsymbol{FC}\left(\boldsymbol{reshape}\left(1\times 1\boldsymbol{Conv}(\boldsymbol{F}_2)\right)\right)\\ \boldsymbol{F}_4 &= \boldsymbol{softmax}\left(1\times 1\boldsymbol{Conv}(\boldsymbol{q}\times\boldsymbol{k})\right)\end{aligned} \tag{3}$$

LF模块中的value(v)通过3个3×3*Conv*卷积层和多尺度融合模块进行计算。堆叠3个3×3*Conv*相当于一个7×7*Conv*的感受野，可以有效地提取更丰富的局部特征。如图4所示，MSF模块能够获取不同尺度的特征信息，使得特征提取更加细致精确。卷积层与MSF模块相辅相成，互相弥补不足，从而提取出更为精细的特征。这种结构设计进一步提升特征的表达能力，有助于更好地应对复杂的建筑物形状和背景干扰。

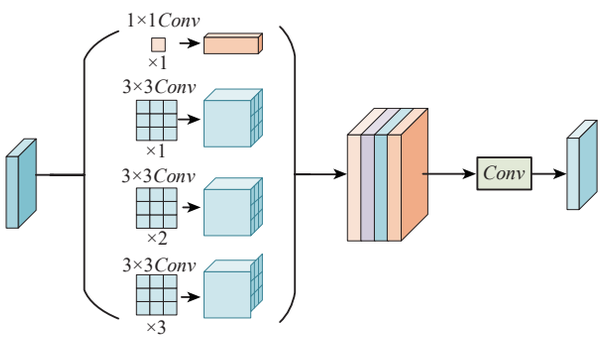

图4 多尺度融合模块

Fig. 4 Multi-scale fusion module

给定输入特征$\boldsymbol{F}\in\boldsymbol{R}^{H\times W\times C}$，通过卷积层和MSF模块分别计算得到$\boldsymbol{F}_{Conv}\in\boldsymbol{R}^{H\times W\times C}$和$\boldsymbol{F}_{LConv}\in\boldsymbol{R}^{H\times W\times C}$，然后将$\boldsymbol{F}_{Conv}$和$\boldsymbol{F}_{LConv}$相加得到$\boldsymbol{v}\in\boldsymbol{R}^{H\times W\times C}$。最后，执行attention和$v$的矩阵乘法，得到输出$\boldsymbol{W}_{\text{attention}}\in\boldsymbol{R}^{H\times W\times C}$。

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{Conv}(\boldsymbol{F})+\boldsymbol{LConv}(\boldsymbol{F})\\ \boldsymbol{W}_{\text{attention}} &= \boldsymbol{F}_4\times\boldsymbol{v}\end{aligned} \tag{4}$$

式中，$\boldsymbol{Conv}(\cdot)$表示3×3卷积层组，$\boldsymbol{LConv}(\cdot)$表示MSF模块。最后，将提取的局部特征和全局特征进行融合，该模块通过卷积层和1×1卷积层组成的前馈层来计算最终输出$\boldsymbol{W}_{\text{attention}}\in\boldsymbol{R}^{H\times W\times 2C}$。

$$\begin{aligned} \hat{\boldsymbol{W}} &= \boldsymbol{F}+\boldsymbol{v}+\partial\boldsymbol{W}_{\text{attention}}\\ \boldsymbol{W} &= \boldsymbol{Conv}(\hat{\boldsymbol{W}})+1\times 1\boldsymbol{Conv}(\hat{\boldsymbol{W}})\end{aligned} \tag{5}$$

式中，∂表示可学习的权重，$\boldsymbol{Conv}(\cdot)$表示3×3卷积操作，$1\times1\boldsymbol{Conv}(\cdot)$表示1×1卷积操作，$\boldsymbol{W}\in\boldsymbol{R}^{H\times W\times C}$表示特征融合后的输出。$v$是通过卷积和MSF模块提取的特征图，由于缺乏全局信息，无法完全抑制背景噪声。然而，在与全局注意力建模相结合后，通过矩阵相乘得到的特征矩阵包含全局信息，能够有效指导目标特征的提取。

LF模块可以减少对背景信息的关注，增强对遥感建筑物的专注，从而提高建筑物提取的准确性。这种结合既保留局部细节，又充分利用全局上下文信息，使得LF模块在复杂的遥感场景中表现出色。

2.3 交互融合模块

为进一步突出全局上下文特征信息，提出一个交互融合IF（Interactive Fusion）模块，替代传统编解码器结构中的跳跃连接层，以保留更多来自编码器和解码器层的上下文信息。如图5所示，IF模块通过学习低级细节特征与高级语义特征，捕获像素间的长程依赖关系。具体而言，IF模块在特征融合过程中，通过输出的每个通道隐含的语义特征来建模像素与对象之间的长距离依赖，这些通道特征可以用作特定对象的映射，并且不同的语义特征之间存在相互关联。通过这种方式，IF模块能够对全局上下文信息进行编码，以合理利用特征图之间的相互依存关系，增强特征表示能力，从而有效捕获遥感图像中建筑物与其周围背景之间的关系。交互融合模块不仅强化局部特征的细节保留，还通过全局建模实现遥感建筑物背景和建筑物对象之间的协同建模，显著提高建筑物提取的精度。

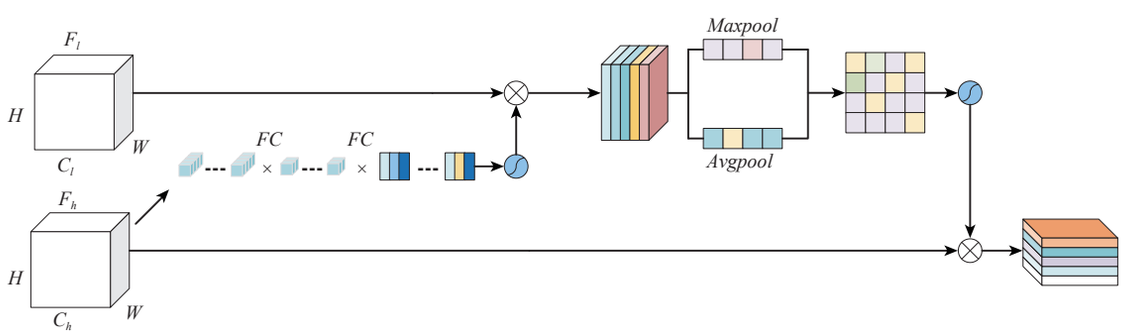

图5 交互融合模块

Fig. 5 Interactive fusion module

首先采用AdaptiveAvgPool2d对特征图$\boldsymbol{u}\in\boldsymbol{R}^{W\times H\times C}$进行操作，其中$C$为通道总数，$H$、$W$分别是特征的高和宽。

$$\boldsymbol{u}=\frac{1}{\boldsymbol{HW}}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{u}(i,j) \tag{6}$$

之后，使用激发操作（$C\rightarrow C/r$，$C/r\rightarrow C$，r=4）来重塑高级特征（$\boldsymbol{F}_H$），同时减少网络参数量。其中$\boldsymbol{\delta}(\cdot)$和$\boldsymbol{BN}(\cdot)$分别是（ReLU）和批量归一化（BN）。$\boldsymbol{T}_1\in\boldsymbol{R}^{C\times\frac{C}{r}}$和$\boldsymbol{T}_2\in\boldsymbol{R}^{\frac{C}{r}\times C}$分别代表$C\rightarrow C/r$和$C/r\rightarrow C$激励操作。

$$\boldsymbol{U}_{IF}=\boldsymbol{\sigma}\left(\boldsymbol{B}\left(\boldsymbol{T}_2\boldsymbol{\delta}\left(\boldsymbol{B}\left(\boldsymbol{T}_1(\boldsymbol{u})\right)\right)\right)\right) \tag{7}$$

最后，跨通道特征通过具有低级特征权重的加权和来实现，其中⊗表示逐元素乘法，并且$\boldsymbol{F}_L$表示解码器中的低级别特征。

$$\boldsymbol{F}_{IF}=\boldsymbol{U}_{IF}\otimes\boldsymbol{F}_L \tag{8}$$

为增强目标与背景的可分辨性性，进一步注重空间信息。首先对F_{IC}执行激励操作（（$C\rightarrow C/r$，r=4），然后使用平均池化（Pavg）和最大池化（Pmax）操作对其进行聚合。其中$C_{3\times3}$是卷积运算。$\left(F_{IF}^{1\times1}\right)$代表1 × 1卷积后的$F_{IF}$。

$$U'_{IF}=\sigma\left(C_{3\times3}\left(P_a\left(F_{IF}^{1\times1}\right);P_{IF}^{1\times1}\left(\left(F_{IF}^{1\times1}\right)\right)\right)\right) \tag{9}$$

空间特征通过具有高级特征权重的加权和来实现，其中⊗表示逐元素乘法，并且$\boldsymbol{F}_H$表示编码器中的高级特征。

$$\boldsymbol{F}'_{IF}=\boldsymbol{U}'_{IF}\otimes\boldsymbol{F}_H \tag{10}$$

3 实验与分析

3.1 数据集

为验证所提出网络的优越性，选择3个公开的遥感建筑物数据集进行广泛实验，包括WHU建筑物数据集（Ji等，2019）、Massachusetts建筑物数据集（Mnih，2013）和中国典型城市建筑物实例数据集（吴开顺 等，2021）。在训练之前，首先对所有输入图像进行归一化处理，随后依次进行随机翻转、高斯模糊和数据增强。3个数据集的详细信息如下：

（1）Massachusetts建筑物数据集包含151张空间分辨率为1 m的波士顿地区航空图像，覆盖约340 km²的城市和郊区场景。这些图像均为1500×1500像素大小。官方数据集划分为训练集（137张图像）、验证集（4张图像）和测试集（10张图像）。为便于模型训练，将这些图像裁剪为适合网络输入的256×256像素小块，生成4932张训练图像、144张验证图像和360张测试图像。

（2）WHU建筑物数据集包含8189张图像（其中4736张用于训练，1036张用于验证，2416张用于测试），空间分辨率为0.3 m，覆盖面积超过450 km²，包含约22000座建筑物。为便于训练，将图像裁剪为适合网络输入的256×256像素小块，得到18944张训练影像、4144张验证影像和9664张测试影像。

（3）中国典型城市建筑物实例数据集覆盖中国多个典型城市（含北京、上海、深圳、武汉），空间分辨率达0.29 m。数据源采自公开卫星影像，共包含63886个精细化标注的建筑实例，地理覆盖范围约120 km²（其中5985张用于训练，1275张用于测试）。每张图片的分辨率为500×500像素。为便于训练，将图像裁剪为适合网络输入的256×256像素小块，得到23940张训练影像和5100张测试影像。

3.2 评价指标

为对所提出的网络进行广泛而全面的评估，选用4个指标：交并比（IoU）、F1、精确度（Precision）和召回率（Recall）。其中，公式中的TP、FP和FN分别表示真阳性、假阳性和假阴性。

$$\mathrm{IoU}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{11}$$

$$\mathrm{F1}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{12}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{13}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{14}$$

所有相关实验均在PyTorch 2.01（CUDA 11.7）环境下，在配备12 GB内存的NVIDIA GeForce RTX 4070 GPU上进行。在训练阶段，选择SGD优化器，并采用余弦策略来调整学习率。为确保实验的公平性和一致性，在WHU数据集和Massachusetts数据集和中国典型城市建筑物实例数据集上，所提出的网络与其他网络（包括U2Net（Qin等，2020）、Swin-Transformer（Liu等，2021）、MANet（Li等，2022）、HD-Net（Li等，2024b）、RS-Mamba（Zhao等，2024））的参数设置均为：初始学习率为0.05，批量大小为6，训练轮数（epoch）为500。

3.3 结果分析

3.3.1 消融实验

为探索提出的模块和组件在FLGF-UNet中的有效性，本文在Massachusetts数据集上进行广泛的实验。前4组FLGF-UNet的消融实验中，通道数选择以C=16为基准，第5组消融实验则用于验证通道选择的影响。在接下来的部分中，将提供详细的分析。

（1）模块有效性。为探索所提出模块在FLGF-UNet中的有效性，本文在Massachusetts建筑数据集上进行广泛实验，依次验证残差注意力（RA）模块、线性融合（LF）模块和交互融合（IF）模块的效果，如表1所示。实验表明，当在网络中加入LF模块时，网络性能有所提升，表明LF模块可以有效处理网络的深层特征。在交互融合模块的作用下，低级特征能够逐通道引导高级特征学习空间和语义关系，从而显著提高提取精度。单独加入LF或IF模块都能增强FLGF-UNet的性能，而将LF和IF模块结合使用时，分割效果进一步提升，验证LF和IF模块在结构上的互补性。

表1 模块有效性

Table 1 Module validity

/%

RA	LF	IF	IoU	F1	P	R
✓			71.12	83.12	84.02	82.24
✓	✓		71.59	83.45	84.59	82.35
✓		✓	71.75	83.55	82.96	84.16
✓	✓	✓	**73.27**	**84.57**	**84.98**	**84.17**

注：黑体数值表示该指标的最好结果。

图6展示Massachusetts数据集中不同网络的遥感建筑物的可视化结果。RA模块构成FLGF-UNet的基线网络。当背景中存在较多噪声干扰时，网络借助LF模块的出色噪声抑制能力，有效避免误检情况。FLGF-UNet首先通过LF模块对目标进行特征增强，防止目标丢失；接着通过IF模块进行细化，有效检测出非常微小的遥感建筑物，使分割结果更加精确。当仅使用LF或IF模块时，由于特征提取和融合不够充分，建筑物提取结果中出现空洞以及误检和漏检现象。而当同时加入LF和IF模块后，分割结果显示建筑物的提取图像更加贴合标签，大大减少上述问题。

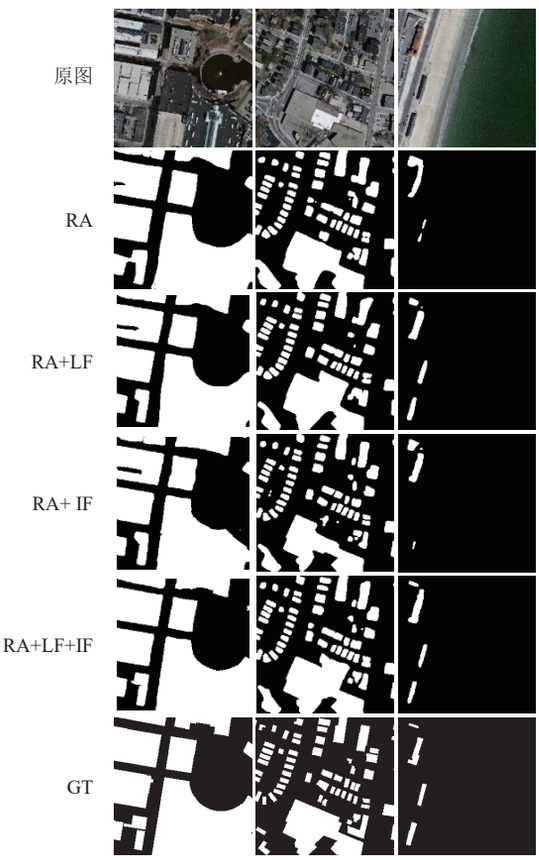

图6 在Massachusetts数据集上不同网络的可视化

Fig. 6 Visualization of different networks on the Massachusetts dataset

（2）LF模块有效性。对Massachusetts数据集进行的LF模块消融研究结果如表2所示。第2行表示在LF模块中去除GA后的效果，此时LF模块仅保留局部特征提取能力；第3行则是仅保留全局特征提取能力后的结果。与基线相比，这两种变体的性能均有所提升，但未达到完整LF模块的性能。第4行显示使用完整LF模块后网络性能的显著改善，表明这两种子模块在LF模块中各自发挥独特作用，共同提升网络的整体性能。

表2 LF模块有效性

Table 2 LF module effectiveness

/%

方法	IoU	F1	Precision	Recall
baseline	71.12	83.12	84.02	82.24
MSF	71.39	83.30	84.36	82.27
GA	71.22	83.19	84.22	81.18
MSF+GA	**71.59**	**83.45**	**84.59**	**82.35**

注：黑体数值表示该指标的最好结果。

（3）编码器有效性。如表3所示，将编码器第一层中的RA模块替换为LF模块后，网络性能显

著下降。由于遥感图像缺乏明确的语义信息，且信噪比较低，LF模块在计算注意力矩阵时会被维度压缩。当直接使用复杂的原始遥感图像作为输入时，LF模块在计算注意力矩阵时容易受到干扰，从而传递出错误的全局信息。因此，首先使用残差注意力模块对遥感图像进行初步特征提取，去除部分背景杂波，使输入到LF模块的特征图干扰更少。实验证明，LF模块在处理特征图方面比直接处理原始建筑图像效果更好。

表3　编码器有效性

Table 3　Encoder validity

/%

RA	IoU	F1	P	R
N	71.29	83.24	84.04	82.45
Y	**71.58**	**83.44**	**84.37**	**82.52**

注：黑体数值表示该指标的最好结果。

（4）监督学习的选取。由于单分支监督学习（仅对网络末端输出计算损失）效果显著受限，其根本原因在于深层特征传递过程中的信息衰减与梯度优化不足。在遥感建筑物分割任务中选取上采样监督学习策略，主要源于下采样过程中的语义信息丢失问题：传统下采样操作会因分辨率的降低过滤高频细节，导致损失计算时特征图与原始标签无法对齐。为解决这一问题，本文采用上采样GT方法，通过双线性插值将标签与中间层特征图进行像素级对齐监督，使平均损失值降低0.0147，并构建多尺度深度监督机制在解码器的不同上采样阶段注入监督信号，浅层约束纹理细节），深层保障全局语义，实验数据表明，该策略有效协调了多尺度语义表达，缓解了遥感影像中普遍存在的类间相似性与尺度敏感性问题。

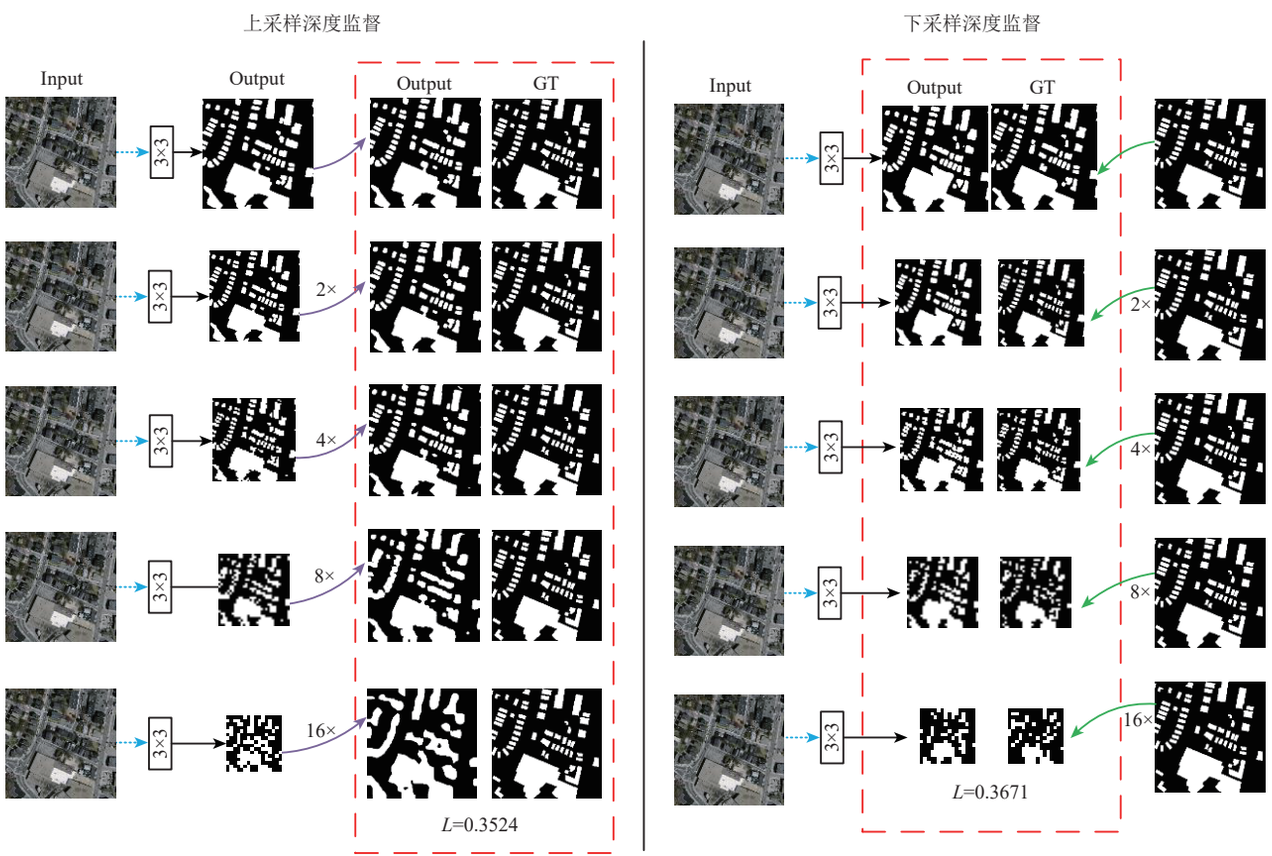

图7　监督学习方式

Fig. 7　Supervised learning method

（5）网络通道选择。如表5所示，根据不同的输入维度C将网络划分为T（C=16）、S（C=32）、B（C=64）和L（C=128）。并在Massachusetts数据集上进行通道数的消融实验。研究不同输入维度C对网络性能的影响，最终选取通道数C=64作为最佳配置。

当C分别设置为16、32和64时，网络性能逐步提升。但本文发现，当C设置为128时，性能略

有下降。

由于遥感图像没有明确的语义信息，当输入维数过大时，会影响网络的归纳偏差，从而导致网络不能有效地提取目标特征。性能略有下降。

表4 监督学习的选取

Table 4 Selection of supervised learning

/%

Deep supervision	IoU	F1	P	R
单分支监督	72.92	84.34	84.66	84.02
上采样监督	**73.27**	**84.57**	**84.98**	**84.17**
下采样监督	73.12	84.47	84.82	84.13

注：黑体数值表示该指标的最好结果。

表5 网络通道选择

Table 5 Network channel selection

/%

Channel	IoU	F1	Precision	Recall
16	73.27	84.57	84.98	84.17
32	74.74	85.54	85.88	85.19
64	**75.47**	**86.02**	**86.47**	**85.58**
128	73.81	84.93	85.87	84.02

注：黑体数值表示该指标的最好结果。

3.3.2 对比实验

为验证本文网络的有效性，本文在建筑物数据集上将其与典型网络和先进网络进行对比。具体比较结果如下：U2Net（Qin等，2020）、Swin-Transformer（Liu等，2021）、MANet（Li等，2022）、HD-Net（Li等，2024b）和RS-Mamba（Zhao等，2024），FLGF-UNet在3个数据集中取得更高精度。U2Net：通过双级嵌套U型结构和RSU模块提取多尺度特征，表现出色，适合建筑物边缘和细节提取。用于验证FLGF-UNet在多尺度特征提取和细节保持方面的改进。Swin-Transformer：利用自注意力机制捕获全局特征，适合复杂场景和大尺度物体提取。用于验证FLGF-UNet在融合全局信息和局部细节方面的能力。MANet：采用多注意力机制增强上下文依赖，适合多尺度特征融合和建筑物精度提取。用于展示FLGF-UNet在噪声抑制和细节增强方面的改进。HD-Net：通过高分辨率特征的多尺度交互，解决整体与边界特征耦合问题，提升建筑物提取性能。用于验证FLGF-UNet在多尺度特征交互与全局建模上的优势。RS-Mamba：专为大规模遥感图像密集预测设计，处理复杂场景和高分辨率影像表现优异。用于评估FLGF-UNet在大规模遥感图像处理和细节丰富场景中的提取能力。

（1）对Massachusetts建筑物数据集的分析。定量分析：表6列出在Massachusetts数据集上的不同网络的总体定量评价结果。与其他SOTA网络相比，FLGF-UNet实现最佳性能。建筑物和非建筑物的表征特征之间的差异更大，导致它们之间的区分度更高。具体来说，与SOTA网络RS-Mamba相比，在IoU、F1、Precision和Recall指标上分别提升0.49%、0.32%、0.13%、0.48%，与其他网络相比，FLGF-UNet取得了明显优势，说明所提方法具有较好的有效性。相比之下，所提出的FLGF-UNet，正确提取更多的建筑物，同时获得更少的错误预测。这主要是由于以下两个方面。一方面，引入一种有效的融合模块，利用FLGF-UNet模块的优势，这可以推动网络学习更多的特征。另一方面，跨层的融合策略将有助于更好地表征不同尺度的建筑物。

表6 在Massachusetts数据集不同网络定量分析

Table 6 Quantitative analysis of different networks on the Massachusetts dataset

/%

Network	IoU	F1	Precision	Recall
U2Net	74.78	85.57	86.21	84.94
Swin-T	74.65	85.48	85.97	85.01
MANet	73.66	84.83	84.99	84.68
HD-Net	74.81	85.59	86.11	85.07
RS-Mamba	74.98	85.70	86.34	85.10
FLGF-UNet	**75.47**	**86.02**	**86.47**	**85.58**

注：黑体数值表示该指标的最好结果。

可视化分析：为更直观地比较FLGF-UNet与其他SOTA网络的性能，将所有网络的提取结果进行可视化，如图8所示。图8中展示FLGF-UNet与其他网络在Massachusetts建筑物数据集上的定性结果。由于数据集的图像分辨率较低且建筑物分布密集，FLGF-UNet在提取效果上表现更为优异。相比之下，U2Net、Swin-T、HD-Net和RS-Mamba都出现明显的空洞现象，而FLGF-UNet在很大程度上避免这一问题，得益于更充分的特征融合。此外，在应对类似物体对建筑物的干扰时，U2Net、Swin-T、MANet、HD-Net和RS-Mamba的提取效果均不理想，而FLGF-UNet的表现更为出色。这

得益于其融合局部和全局信息的优势，使得FLGF-UNet能够更准确地区分建筑物与周围的干扰物体。

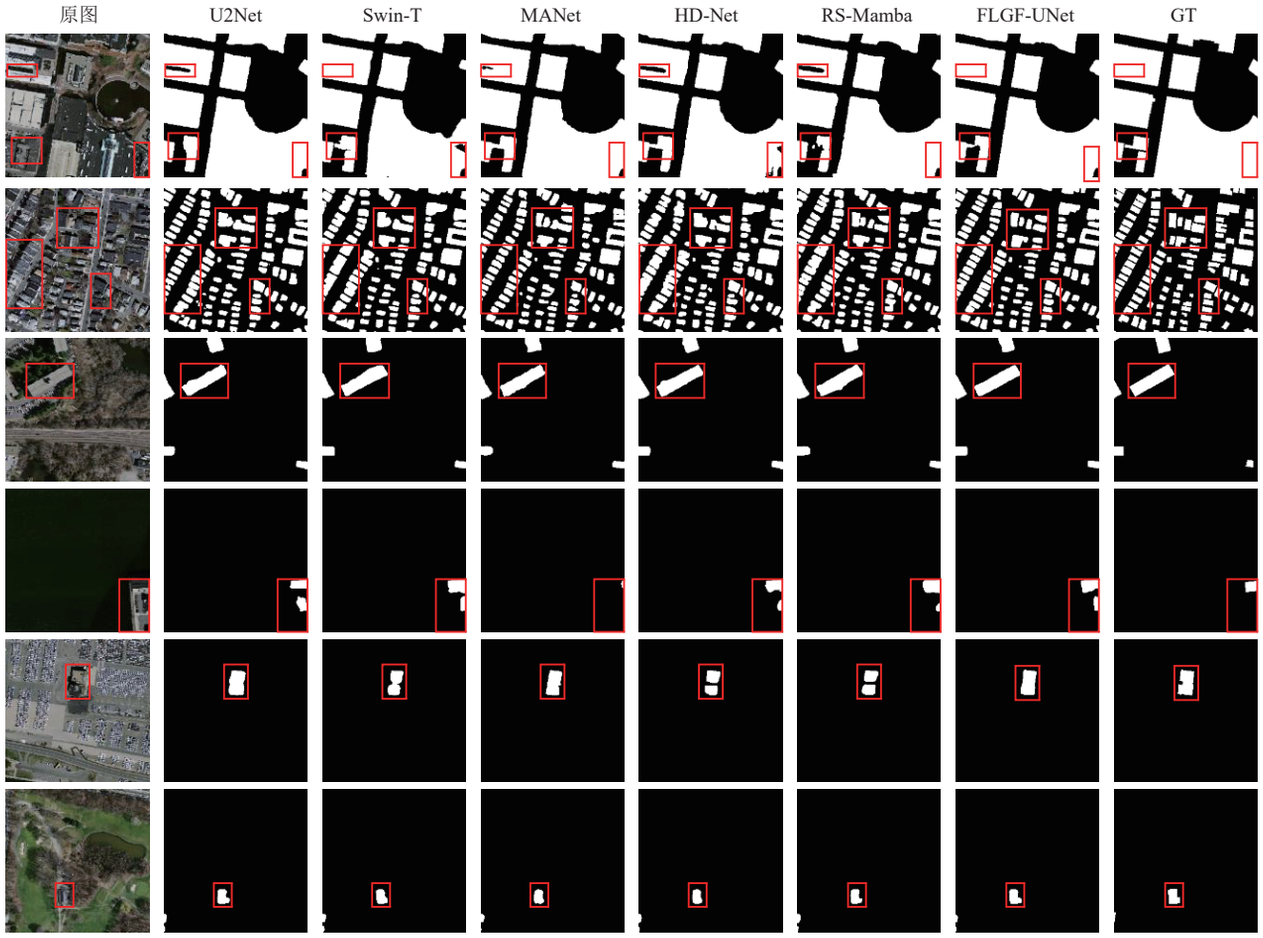

图8　在Massachusetts数据集不同网络可视化分析

Fig. 8　Visual analysis of different networks on the Massachusetts dataset

（2）对WHU建筑数据集的分析。定量分析：表7展示在WHU建筑数据集上，不同网络的总体定量评估结果。与其他SOTA网络（如RS-Mamba）相比，FLGF-UNet在各项指标上表现出最佳性能。具体而言，FLGF-UNet在IoU、F1、Precision和Recall指标上分别提升0.56%、0.31%、0.48%和0.13%。这些显著的提升表明，FLGF-UNet显著提高建筑物提取的效果，验证该网络的优越性。

可视化分析：在WHU建筑数据集上，图9中展示所有对比网络与本文网络的6组可视化示例。在建筑物被树木遮挡或受到类似物体干扰的情况下，各网络的提取结果均出现一定误差，且大部分对比网络未能完整提取出建筑物的轮廓。相比之下，FLGF-UNet的提取结果非常接近标签，证明其在复杂环境中准确提取建筑物的能力。在面对建筑物被遮挡的情况时，U2Net、Swin-T、MANet、HD-Net和RS-Mamba的提取效果欠佳，未能充分利用全局上下文信息。具体来说第六行建筑物被树木遮挡以及受类似物体干扰，相比下之下FLGF-UNet能够提取完整的建筑物。FLGF-UNet能更有效地应对此类复杂场景，展现出对全局信息的利用优势。

表7　在WHU数据集不同网络定量分析

Table 7　Quantitative analysis of different networks on the WHU dataset

/%

Network	IoU	F1	Precision	Recall
U2Net	90.46	94.99	95.13	94.85
Swin-T	90.07	94.78	95.16	94.79
MANet	89.54	94.48	93.92	95.05
HD-Net	90.43	94.97	95.16	94.79
RS-Mamba	90.77	95.16	94.78	95.55
FLGF-UNet	**91.33**	**95.47**	**95.26**	**95.68**

注：黑体数值表示该指标的最好结果。

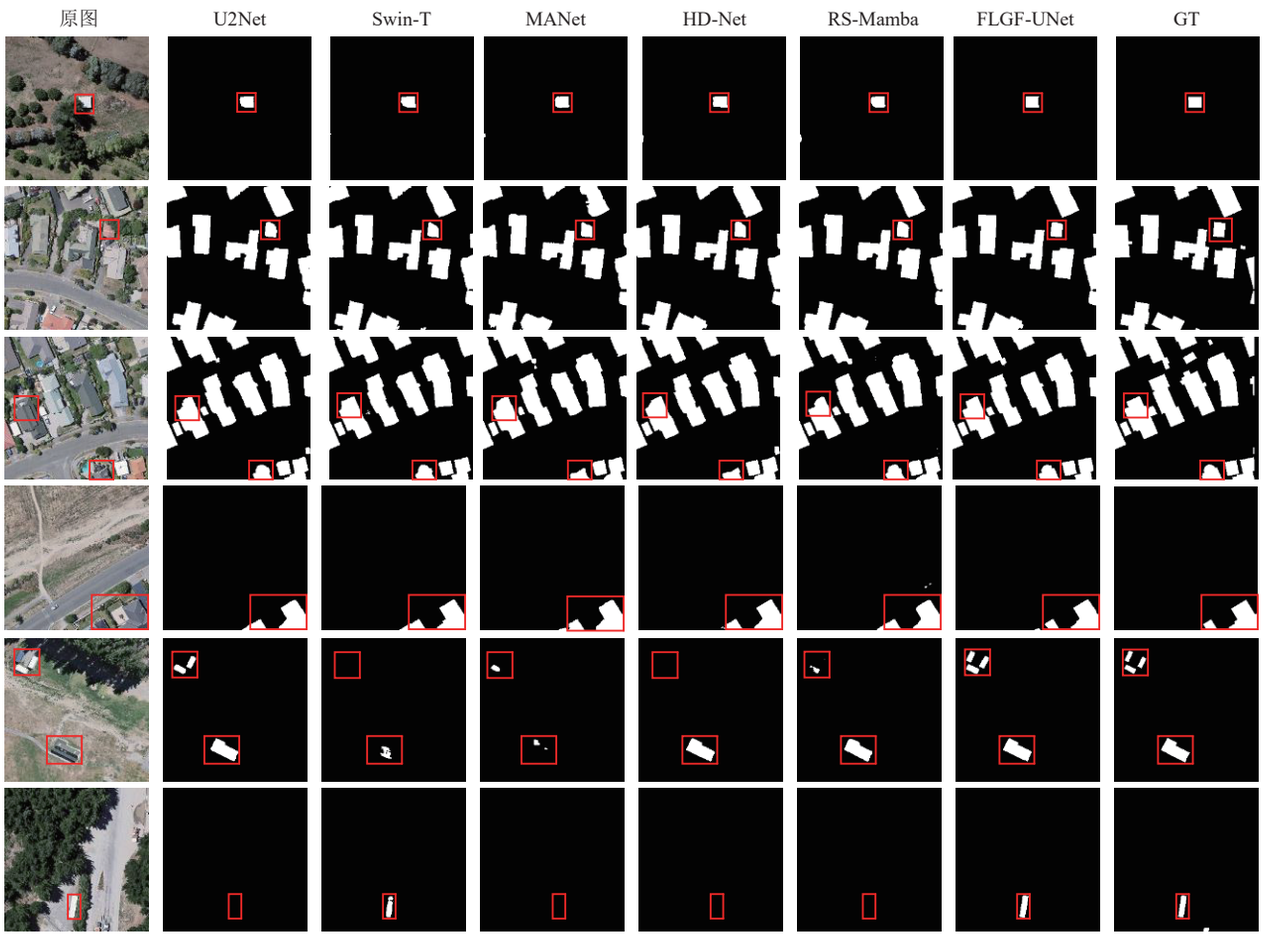

图9 在WHU数据集不同网络可视化分析

Fig.9 Visual analysis of different networks on the WHU dataset

（3）对中国典型城市建筑物实例数据集的分析。定量分析：表8展示在中国典型城市建筑物实例数据集上，FLGF-UNet相较于RS-Mamba等现有SOTA网络，IoU、F1-score、Precision及Recall分别提升0.7%、0.45%、0.31%与0.59%，4项指标全面领先，验证其通过全局—局部特征融合策略对建筑物的精准提取能力，进一步佐证方法的性能优势。

表8 在中国典型城市建筑物实例数据集不同网络定量分析

Table 8 Quantitative analysis of different networks in typical urban building instance datasets in China

/%

Network	IoU	F1	Precision	Recall
U2Net	76.07	86.41	86.02	86.80
Swin-T	76.01	86.37	85.95	86.68
MANet	74.88	85.63	85.51	85.75
HD-Net	75.69	86.17	85.95	86.35
RS-Mamba	76.23	86.51	86.11	86.92
FLGF-UNet	**76.93**	**86.96**	**86.42**	**87.51**

注：黑体数值表示该指标的最好结果。

可视化分析：在中国典型城市建筑物实例数据集上，图10给出了该数据集上的可视化对比结果。现有方法（如U2Net、RS-Mamba等）对树木遮挡区域的建筑轮廓提取存在误检，第一行案例清晰地展示了建筑物边缘遭受树木遮挡的情况，树木明显地覆盖了建筑物的部分轮廓，第2、3行案例则着重体现了建筑物受到广场、集装箱以及汽车等不同元素干扰的情形，对建筑物的准确识别产生了一定程度的影响，FLGF-UNet凭借全局语义关联与局部细节增强机制，显著抑制干扰并还原完整建筑边界，其预测结果与真值标注高度吻合，印证全局上下文建模在复杂遮挡场景中的关键作用。

（4）树木遮挡与类似物体干扰场景下的建筑物提取分析。为评估模型在复杂环境中的鲁棒性，本研究从WHU（地貌干扰）、Massachusetts（植被遮挡）及中国典型城市数据集（邻近建筑干扰）中分别选取一组典型场景进行对比分析。实验表

明如图11所示，RS-Mamba在各类干扰场景下均存在显著局限性：地貌起伏导致建筑轮廓模糊时，其提取结果出现不完整；影像中树木遮挡区域漏检现象突出；高密度城区内相似屋顶结构则引发误检。相比之下，FLGF-UNet通过全局—局部协同建模机制，有效克服上述挑战。

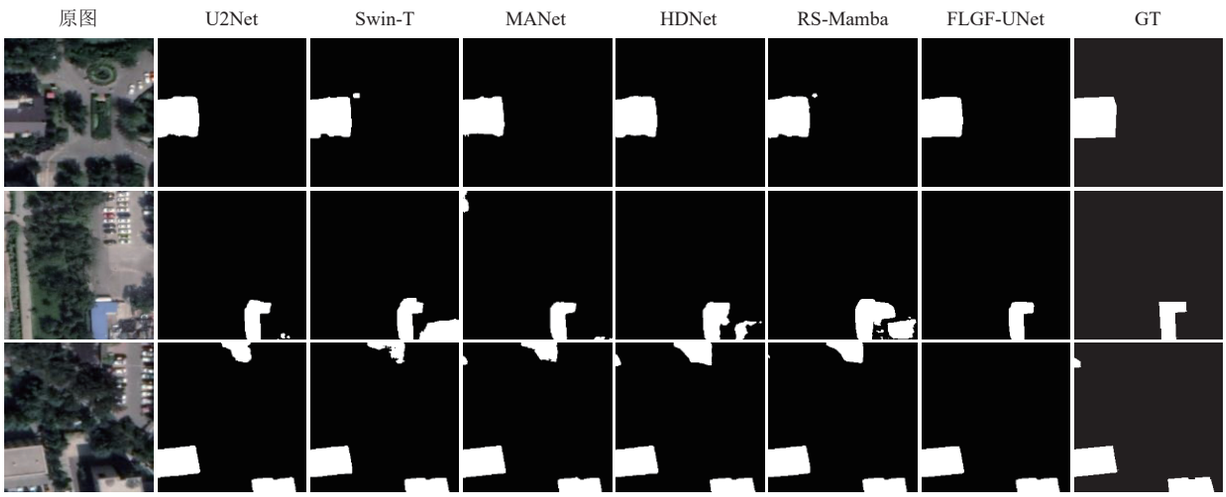

图10 在中国典型城市建筑物实例数据集不同网络可视化分析

Fig. 10 Visualization analysis of different networks in typical urban building instance datasets in China

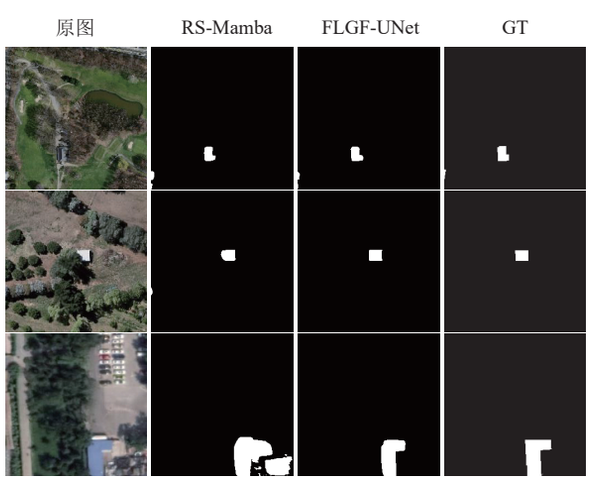

图11 树木遮挡与类似物体干扰场景下的建筑物提取可视化对比

Fig. 11 Visual comparison of building extraction under tree occlusion and similar object interference scenes

（5）网络复杂性分析。表10展示FLGF-UNet与其他网络在WHU数据集上的参数量（Parameters）、浮点运算量（FLOPs）、IoU和F1指标的比较。尽管FLGF-UNet的精度高于U2Net（2020）、Swin-T（2021）、MANet（2021）、HD-Net（2024）和RS-Mamba（2024）但其参数量和浮点较高。这一现象归因于其核心模块设计：线性融合（LF）模块通过多尺度特征提取（MSF）模块与全局注意力（GA）模块，虽提升复杂建筑结构的表征能力，却引入额外的计算负载。实验表明，FLGF-UNet在精度与复杂度间的权衡验证全局上下文建模的必要性，同时也为后续研究指明方向——通过引入动态卷积、通道剪枝等轻量化策略，在维持高精度的前提下优化模型效率，以满足遥感边缘计算设备的实际部署需求。

表10 在WHU数据集不同网络复杂性分析

Table 10 Analysis of different network complexities in the WHU dataset

Model	FLOPs/G	Parameters/M	IoU/%	F1%
U2Net	37.8	44.0	90.46	94.99
Swin-T	66.8	66.0	90.07	94.78
MANet	**13.5**	35.9	89.54	94.48
HD-Net	50.3	**14.7**	90.43	94.97
RS-Mamba	94.2	27.9	90.77	95.16
FLGF-UNet	105.9	54.4	**91.33**	**95.47**

注：黑体数值表示该指标的最好结果。

4 结 论

本文提出一种新的遥感图像语义分割网络FLGF-UNet。通过将残差注意力模块应用于编码器和解码器，有效防止梯度消失，并利用注意力模块实现高效的特征提取；同时，采用LF模块既可捕获全局背景信息，也可捕获局部空间细节。LF模块通过MSF模块和GA模块进行局部和全局特征提取，从而增强遥感建筑物特征并抑制噪声。

此外，为弥补编码器和解码器之间的语义鸿沟，设计交互融合模块IF，在低级和高级表征之间编码语义特征，进一步提高上下文信息的对比度。

为验证FLGF-UNet的有效性和通用性，本文在WHU建筑数据集、Massachusetts建筑数据集和中国典型城市建筑物实例数据集3个公开数据集上进行实验。实验结果表明，训练后的FLGF-UNet能够准确区分前景和背景，并在IoU、F1、Recall和Precision等指标上表现出较高的性能。此外，与其他SOTA网络进行对比；结果显示，FLGF-UNet在性能指标上显著优于现有网络。同时，通过基于Massachusetts建筑数据集的消融实验，验证FLGF-UNet各分支结构和关键模块的作用和重要性。最后，本文讨论FLGF-UNet的效率。值得注意的是，尽管本文的分析表明FLGF-UNet在当前任务中是有效的，但仍存在以下局限性：FLGF-UNet仅在城市和山区场景中的遥感建筑语义分割任务中进行测试，尚未在道路分割和地块分割等其他任务中进行验证。

参考文献（References）

Azad R, Heidari M, Wu Y L and Merhof D. 2022. Contextual attention network: transformer meets U-net//Proceedings of the 13th International Workshopon Machine Learning in Medical Imaging. Singapore: Springer: 377-386 [DOI: 10.1007/978-3-031-21014-3_39]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chen C F R, Fan Q F and Panda R. 2021. CrossVIT: cross-attention multi-scale vision transformer for image classification//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 347-356 [DOI: 10.1109/ICCV48922.2021.00041]

Ding L, Tang H and Bruzzone L. 2021. LANet: local attention embedding to improve the semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 59 (1): 426-435 [DOI: 10.1109/TGRS.2020.2994150]

Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W and Lu H Q. 2019. Dual attention network for scene segmentation//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 3141-3149 [DOI: 10.1109/CVPR.2019.00326]

He X, Zhou Y, Zhao J Q, Zhang D, Yao R and Xue Y. 2022. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 60: 4408715 [DOI: 10.1109/TGRS.2022.3144165]

Ji S P, Wei S Q and Lu M. 2019. 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 G, Zhao S L, Li M, Zhou M L and Ying Z B. 2024a. IDP-Net: industrial defect perception network based on cross-layer semantic information guidance and context concentration enhancement. Engineering Applications of Artificial Intelligence, 130: 107677 [DOI: 10.1016/j.engappai.2023.107677]

Li R, Zheng S Y, Zhang C, Duan C X, Su J L, Wang L B and Atkinson P M. 2022. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 60: 5607713 [DOI: 10.1109/TGRS.2021.3093977]

Li Y X, Hong D F, Li C Y, Yao J and Chanussot J. 2024b. HD-Net: high-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition. ISPRS Journal of Photogrammetry and Remote Sensing, 209: 51-65 [DOI: 10.1016/j.isprsjprs.2024.01.022]

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//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Mnih V. 2013. Machine Learning for Aerial Image Labeling. Toronto: University of Toronto (Canada).

Mou L C, Hua Y S and Zhu X X. 2020. Relation matters: relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. IEEE Transactions on Geoscience and Remote Sensing, 58(11): 7557-7569 [DOI: 10.1109/TGRS.2020.2979552]

Qin X B, Zhang Z C, Huang C Y, Dehghan M, Zaiane O R and Jagersand M. 2020. U^2-Net: going deeper with nested U-structure for salient object detection. Pattern Recognition, 106: 107404 [DOI: 10.1016/j.patcog.2020.107404]

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Munich: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Song P F, Li J J, An Z Y, Fan H and Fan L W. 2023. CTMFNet: CNN and transformer multiscale fusion network of remote sensing urban scene imagery. IEEE Transactions on Geoscience and Remote Sensing, 61: 5900314 [DOI: 10.1109/tgrs.2022.3232143]

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: 5625711 [DOI: 10.1109/TGRS. 2022.3186634]

Wang P Q, Chen P F, Yuan Y, Liu D, Huang Z H, Hou X D and Cottrell G. 2018. Understanding convolution for semantic segmentation//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE: 1451-

1460 [DOI: 10.1109/WACV.2018.00163]

Zhang S, Jiang Y H, Wang C J, Tan M L, Du B and Tian F. 2024. S^2PNet: an interactive learning framework for addressing spatial-spectral heterogeneity in H_2 imagery classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 18456-18473 [DOI: 10.1109/jstars.2024.3464758]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

Zhao S J, Chen H, Zhang X L, Xiao P F, Bai L and Ouyang W L. 2024. RS-mamba for large remote sensing image dense prediction. IEEE Transactions on Geoscience and Remote Sensing, 62: 5633314 [DOI: 10.1109/TGRS.2024.3425540]

Wu K, Zheng D Y, Chen Y L, Zeng L Y, Zhang J H, Chai S H, Xu W J, Yang Y L, Li S W, Liu Y Y and Fang F. 2021. A dataset of building instances of typical cities in China. China Scientific Data, 6 (1): 187-195 (吴开顺, 郑道远, 陈妍伶, 曾林芸, 张嘉辉, 柴生华, 徐文杰, 杨永亮, 李圣文, 刘袁缘, 方芳. 2021. 中国典型城市建筑物实例数据集. 中国科学数据(中英文网络版), 6(1): 187-195) [DOI: 10.11922/noda.2021.0013.zh]

FLGF UNet：Remote sensing building extraction network for optical remote sensing images that integrates local−global features

LI Guoyan，LIU Tao，WANG Li，LIU Yi

School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China

Abstract：Semantic segmentation of remote sensing images plays a crucial role in fields such as urban change detection, environmental protection, and geological disaster identification. Addressing issues in current remote sensing building extraction, such as missed detections, false positives, and incomplete extractions due to tree obstruction or similar object interference, this paper proposes an improved building extraction network based on the U-Net architecture—the Fusion of Local Global Features Network (FLGF-UNet). This model integrates local and global information, acquiring features in parallel through multiscale local and global modeling. This step ensures that the feature representation at each stage of the network incorporates both local and global information, enabling a more natural and comprehensive integration of information across different scales and levels during the feature extraction process.

The parallel feature fusion method of FLGF-UNet ensures that the features of each stage contain fine-grained local information and global dependencies so that the network has both local and global information in the feature representation of each stage. As a result, it effectively overcomes the shortcomings of transformer in local information exchange and at the same time outperforms traditional CNN in global information modeling. The LF module is introduced to extract local information and global context information of different scales to ensure the integrity of feature dependencies, thereby making up for the shortcomings of a single module in feature extraction. In addition, the semantic gap between the encoder and decoder is bridged by adding an interactive fusion module between the encoder and decoder to enhance the fusion effect of spatial details, global context, and semantic features. The superiority and versatility of FLGF-UNet are verified by comparing it with networks such as U2Net, Swin transformer, MA-Net, HD-Net, and RS-Mamba on the WHU dataset, the Massachusetts dataset, and typical urban building instance datasets in China. Results show that FLGF-UNet outperforms other SOTA networks in performance and has high practical application value.

In conclusion, FLGF-UNet is an innovative network for extracting buildings from high-resolution remote sensing images. It takes parallel multiscale local-global modeling as its core, so that features at all levels can simultaneously have local details and long-range semantics, effectively bridging the gap between spatial dependence and local details and significantly improving extraction accuracy. Extensive experiments across datasets verified that its performance is significantly better than that of existing methods, providing a reliable solution for high-precision building extraction from high-resolution remote sensing images. FLGF-UNet brings together cutting-edge methods and technological innovations, marking an important leap forward in remote sensing image analysis, and will continue to drive the in-depth development of this field in terms of accuracy improvement, scene expansion, and practical applications.

In the future, the breakthrough results of FLGF-UNet will inspire the academic community to continue to delve deeper into the local-global structure. Migrating and expanding this fusion strategy to more remote sensing interpretation tasks has great potential and promising prospects. At the same time, in view of the diverse scale and complexity of urban scenes, further optimizing the model architecture and improving adaptive capabilities are natural evolutionary paths. Subsequent research should focus on the implementation of technology and effectively transform the advantages of the algorithm into perceptible, quantifiable, and sustainable socioeconomic benefits for urban planners and disaster emergency decision-makers.

Key words：remote sensing images, building extraction, fusion of local-global feature networks, feature fusion, interactive fusion module

Supported by National Natural Science Foundation of China (No. 52178295)