E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS


**中图法分类号**:TP751　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)08-2866-18

**论文引用格式**:Xu S J, Liu Y R, Liu E H, Liu J, Shi Y and Li X H. 2025. Global perception and detail enhancement network for building segmentation in remote sensing images. Journal of Image and Graphics, 30(8):2866-2883(徐胜军, 刘雨芮, 刘二虎, 刘俊, 史亚, 李小晗. 2025. 引入全局感知与细节增强的非对称遥感建筑物分割网络. 中国图象图形学报, 30(8):2866-2883)[DOI:10.11834/jig.240629]

# 引入全局感知与细节增强的非对称遥感建筑物分割网络

徐胜军[1,2],刘雨芮[1,2*],刘二虎[1,2],刘俊[3],史亚[1,2],李小晗[1,2]

1. 西安建筑科技大学信息与控制工程学院,西安　710055; 2. 西安市建筑制造智动化技术重点实验室,西安　710055;
3. 西安交通大学电气工程学院,西安　710049

**摘　要**:**目的**　针对遥感图像分割的区域连续性差、边界消失和尺度变化大等导致建筑物分割精度低的问题,提出一种基于全局感知与细节增强的非对称遥感建筑物分割网络(global perception and detail enhancement asymmetric-UNet,GPDEA-UNet)。**方法**　在U-Net网络基础上,首先构建了一个基于选择性状态空间的特征编码器模块,以视觉状态空间(visual state space,VSS)作为基础单元,结合动态卷积分解(dynamic convolution decomposition,DCD)捕捉遥感图像中的复杂特征和上下文信息;其次通过引入多尺度双交叉融合注意力模块(multi-scale dual cross-attention,MDCA)解决多尺度编码器特征间的通道与空间依赖性问题,并缩小编解码器特征之间的语义差距;最后设计了一个细节增强解码器模块,使用DCD与级联上采样(cascade upsampling,CU)模块恢复更丰富的语义信息,保留特征细节与语义完整,最终确保分割结果的精确性与细腻度。**结果**　实验在WHU Aerial Imagery Dataset和Massachusetts Building Dataset数据集上与多种方法进行了比较,实验结果表明,所提出的GPDEA-UNet的交并比、精确度、召回率和F1分数在WHU Aerial Imagery Dataset数据集上分别为91.60%、95.36%、95.89%和95.62%,在Massachusetts Building Dataset数据集上分别为72.51%、79.44%、86.81%和82.53%。**结论**　所提出的基于全局感知与细节增强的非对称遥感建筑物分割网络,可以有效提高遥感影像建筑物的分割精度。

**关键词**:遥感图像;建筑物分割;视觉状态空间;动态卷积分解(DCD);交叉注意力;细节增强

# Global perception and detail enhancement network for building segmentation in remote sensing images

Xu Shengjun[1,2], Liu Yurui[1,2*], Liu Erhu[1,2], Liu Jun[3], Shi Ya[1,2], Li Xiaohan[1,2]

*1. School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an　710055, China;*
*2. Xi'an Key Laboratory of Building Manufacturing Intelligent & Automation Technology, Xi'an　710055, China;*
*3. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an　710049, China*

**Abstract**: **Objective** Remote sensing images are a type of earth observation data with wide coverage, rich spectral information, and variable target structures. The advancement of computer technology has steadily increased the demand for accurate and efficient extraction of buildings across diverse domains and industries. Meanwhile, the application prospects of

**收稿日期**:2024-10-23;**修回日期**:2025-01-05;**预印本日期**:2025-01-12

*通信作者:刘雨芮　liu18991358966@163.com

**基金项目**:国家自然科学基金项目(62476211,52278125,62276207);陕西省自然科学基础研究计划(2024JC-YBMS-483,2023-JC-YB-532);陕西省科技厅社发攻关项目(2021SF-429)

**Supported by**:National Natural Science Foundation of China(62476211,52278125,62276207)

semantic segmentation techniques in remote sensing image have progressively demonstrated substantial practical significance. By utilizing the semantic segmentation technology for remote sensing images, detailed information such as the spatial distribution and density of buildings and other infrastructures can be efficiently extracted. This information will play a crucial role in land surveying, urban planning, and post-disaster assessments. However, this advancement has simultaneously increased the complexity of semantic segmentation for buildings in remote sensing images. Consequently, the challenge of efficiently and accurately extracting building information from high-resolution imagery has emerged as a pivotal concern in the field of semantic segmentation of remote sensing images, which demands urgent attention and resolution. In recent years, deep learning has notably advanced in the field of semantic segmentation of remote sensing images. These advancements are due to its ability to learn any data distribution without requiring prior statistical knowledge of the input data, its capacity for self-learning target features, and its strong generalization capabilities. However, the process of semantic segmentation for remote sensing images of buildings faces substantial obstacles, which are primarily due to robust interferences such as varying lighting conditions, seasonal changes, and complex background information, as well as the intricate architectural structures and edge details of the buildings themselves. To address these challenges, this study proposes a global perception and detail enhancement asymmetric-UNet (GPDEA-UNet) network for building semantic segmentation in remote sensing images. **Method** First, the proposed network using UNet architecture constructs a feature encoder module based on the selective state space module. This module is specifically designed to meticulously extract the texture, boundary, and deep semantic features of buildings in remote sensing images. It leverages the visual state space as its fundamental building block and incorporates dynamic convolution decomposition (DCD) to significantly enhance the extraction of intricate features and context information in the remote sensing images while effectively reducing computational overhead. Second, a multi-scale dual cross-attention (MDCA) module is introduced to further broaden the global receptive field of the network and tackle the semantic discrepancy challenges posed by the codec during skip connections. MDCA represents an advanced attention-weighting mechanism that harmoniously integrates cross-channel attention and cross-spatial attention. This module substantially enhances the capability of the network to extract and fuse feature information pertinent to the region and boundary of the segmented target. Meanwhile, it effectively resolves the interdependencies among multi-scale encoder features in channel and spatial dimensions, which bridges the semantic gap between encoder and decoder features. Finally, a detail enhancement decoder module is designed to restore the resolution of the extracted feature maps, with the aim of addressing the issue of image detail information loss during the upsampling phase. This module builds upon the principles of DCD and incorporates a cascade upsampling (CU) module. The CU is specifically engineered to capture richer semantic information, retain feature details and semantic integrity, and ultimately ensure the high accuracy and delicate precision of the segmentation results. Our network achieves a highly specialized and nuanced segmentation of remote sensing building images by integrating these sophisticated components. **Result** Experimental results demonstrate the exceptional robustness of the GPDEA-UNet network introduced in this study across various datasets. Specifically, on the WHU Aerial Imagery Dataset (WHU), the network achieves an intersection over union (IoU) of 91.60%, precision of 95.36%, recall of 95.89%, and an F1-score of 95.62%. Similarly, on the Massachusetts Building Dataset, the network attains an IoU of 73.51%, precision of 79.44%, recall of 86.81%, and an F1-score of 82.53%. When compared with other state-of-the-art networks, the quantitative indicators reveal that the GPDEA-UNet network attains optimal performance on the WHU dataset and either optimal or near-optimal performance on the Massachusetts Building Dataset. Furthermore, qualitative analysis demonstrates that the proposed network achieves superior segmentation results on the WHU and Massachusetts Building Datasets. The network maintains high-quality segmentation even for remote sensing images with inferior imaging quality, such as those with low resolution, noise, or occlusion. **Conclusion** An asymmetric remote sensing building segmentation network with global perception and detail enhancement is proposed by combining a selective state space module and a multi-scale dual cross-attention mechanism. Experiments on two remote sensing datasets show that the proposed network can effectively improve the accuracy and visualization effects of remote sensing building segmentation. Furthermore, the network exhibits remarkable robustness and versatility. The high precision and recall rates achieved in our experiments highlight its capability to excel not only in high-quality remote sensing building segmentation but also in challenging scenarios. This study shows that the proposed network has excellent universality and application

potential in remote sensing image segmentation and provides a new research idea and method for the research and application of remote sensing image processing.

**Key words**: remote sensing images; building segmentation; visual state space; dynamic convolution decomposition (DCD); cross-attention; detail enhancement

## 0 引 言

遥感图像是一种覆盖范围广、光谱信息丰富且目标结构多变的对地观测数据，借助遥感图像提取建筑物信息在土地勘测、城市规划以及灾后评估等领域发挥着重要作用（王卓和瞿绍军，2024）。近年来，随着高分辨率遥感影像技术的飞速发展，图像中蕴含的地物信息，如植被、道路、水体、裸地和建筑物等愈发精细，但同时也极大地提升了遥感图像建筑物语义分割的难度。因此如何高效精准地从高分辨率图像中提取建筑物信息，已成为遥感图像语义分割领域亟待解决的关键问题（Xu等，2024；Aleissaee等，2023；Li等，2024）。

传统的遥感图像建筑物语义分割方法主要依赖于数学、拓扑学及图像处理技术，通过构建数学模型，并利用建筑物颜色、纹理和梯度等浅层语义信息将图像分割成独立且不重叠的区域，进而根据不同区域表现出来的特征差异，实现目标建筑物与背景的分离。常用的传统方法有：基于边缘的分割方法、基于阈值的分割方法、基于区域的分割方法和基于机器学习的分割方法。Zeng等人（2021）利用K-Mean像素聚类方法改进遥感图像的语义分割，根据未知数据与已知数据的距离确定遥感图像不同像素语义特征的相似性，并选取距离最近的$k$个像素点作为判定未知数据类别的依据。Thottolil和Kumar（2022）提出基于特征的自动化建筑物轮廓提取方法，采用多个过滤的特征轮廓捕捉建筑物的空间光谱特征向量，结合随机森林算法和形态学操作，实现高精度的建筑物轮廓提取。Li等人（2015）结合像素级和分段级的信息识别建筑物，先采用无监督的预分割对像素分类，再基于分段级的区域一致性和形状特征，使用高阶条件随机场分割实现建筑物的准确提取。尽管传统方法在提取低级语义信息时表现良好，但其高度依赖人工构造特征，易受光照、气候等因素影响，且表征能力有限，无法捕捉图像的高级语义特征，从而导致分割结果的精确度和鲁棒性受限。

随着深度学习的发展，卷积神经网络（convolutional neural network，CNN）在计算机视觉领域得到广泛应用（项伟康 等，2024）。在遥感图像建筑物语义分割任务中，CNN凭借其强大的自主学习能力与高度优化的特征提取能力，在提升分割的准确度和智能化程度方面展现出优势。Long等人（2015）提出全卷积网络（fully convolutional network，FCN），首次实现了端到端的像素级语义分割，其将深度卷积后的特征图逐级上采样恢复至原始图像大小，并通过分类器预测每个像素的标签。后续研究人员受FCN框架的启发，提出一系列优化与改进的网络，如UNet（Ronneberger 等，2015）、SegNet（segmentation network）（Badrinarayanan 等，2017）、DeepLab（Chen 等，2018）、PSPNet（pyramid scene parsing network）（Zhao 等，2017）和RefineNet（refinement network）（Lin等，2017）等。Shao等人（2020）提出一种建筑物残差细化网络，通过引入不同感受野的空洞卷积与残差细化模块，在不增加计算量的前提下，实现了对全局特征的提取与分割结果的细化预测。Chen等人（2021b）提出一种密集残差神经网络，结合DeepLabv3+Net、DCNN（deep convolutional neural network）与ResNet（residual network）的优势，实现了不同层级特征的提取和融合，并减少了网络参数。与传统方法相比，基于深度学习的遥感图像语义分割方法能够应对大规模数据集与复杂多样性样本带来的挑战。但其仍存在以下问题：一是数据依赖性强，模型性能取决于训练数据的数量和质量；二是分割精度受限，在处理高分辨率遥感图像时，无法有效捕捉图像中的大尺度上下文信息，并且多次下采样和上采样操作使模型丢失细节信息。

为了增强深度学习网络对不同遥感图像待提取目标的表征能力与关注能力，在编解码模块中增加注意力机制（Ghaffarian等，2021）是一种常用的策略（Li等，2024）。Yang等人（2021）提出基于注意力的融合网络，通过多路径注意力融合模块融合高级语义特征和低级语义特征，以提升分割效果。Li等人

(2021)提出结合空间注意力和通道注意力机制的双重注意力网络,从空间和通道维度引入双注意力机制使得网络更加重视特征之间的联系。Ding等人(2021)提出局部注意力模块和注意力嵌入模块,通过嵌入高层特征的局部焦点来丰富底层特征的语义信息,使网络更好地理解图像块之间的上下文关系,提高语义分割精度。Xu等人(2022)构造了一种多尺度区域注意力模块,通过提取空间上多尺度邻域的特征描述符,并在其上利用自注意力机制构建遥感图像局部区域内多像素的高阶空间相关性与不同局部区域的相关性,提高语义特征对不同尺度目标区域相关性特征的关注能力,也减少了网络的计算复杂度。

不同于CNN,ViT(vision Transformer)(Dosovitskiy等,2021)将图像划分成序列化大小固定的图像块,通过自注意力机制捕获图像中不同区域之间的全局特征,具有更大的感受野。Zheng等人(2021)将ViT引入语义分割中,采用纯Transformer架构作为编码器对图像进行序列化以实现全注意力特征表示,增加了模型对上下文信息以及长距离相关性表征的提取能力。Xie等人(2021)首次将Transformer结构应用于语义分割任务中,利用层次结构、重叠块投影、高效注意力机制和卷积位置嵌入等措施改进ViT,使其在保持高性能的同时,降低计算复杂度,提高运行效率。Chen等人(2021a)根据遥感影像中单个建筑只占据图像很少像素的特点,引入“Sparse Token Transformer”架构,将建筑物在特征空间中表示为一组稀疏特征向量以降低ViT的计算复杂度。然而,尽管ViT模型以其全局信息捕捉能力提升了语义分割等任务的性能,但其相比CNN缺乏局部性归纳偏置,导致在训练数据不足时模型性能下降,并且自注意力机制的计算顺序性限制了并行化的程度,使得ViT在图像处理器(graphics processing unit,GPU)上的效率受限,计算成本提高。

建立在状态空间模型(state space model,SSM)(Gu等,2022)基础上的结构化状态空间序列模型受到广泛关注,其可以实现序列长度成线性或近线性扩展,为序列模型的高效计算带来了新的视角。Gu和Dao(2023)提出首个基于选择性状态空间模型(selective,SSM)的深度学习模型Mamba,通过SSM实现了信息的有效选择与压缩,以线性时间复杂度高效地处理长序列数据中的前后依赖关系,同时在设计过程中考虑硬件特性,利用优化算法与数据结构,进一步提升计算效率。Zhu等人(2024)将SSM应用于视觉任务,建立视觉状态空间模型,实现对图像信息的动态选择与压缩,减小了运算复杂度,并且在图像分类、目标检测和语义分割等任务上超越ViT模型,达到了最优水平。以上研究表明视觉状态空间模型在视觉任务中独特的优势与巨大的潜力,有望为在计算资源有限的情况下提升遥感影像建筑物提取的精度提供新的思路。

综上所述,尽管上述算法各有其优点,但仍存在以下几个亟待解决的问题:1)受卷积网络局部感受野限制,模型难以捕捉全局上下文关系,同时跳跃连接可能引入语义差异,影响分割性能;2)遥感图像中建筑物尺寸形状多样,模型需要具备捕捉并融合多尺度特征的能力,但现有模型在特征融合方面存在不足;3)小尺度建筑物在图像中像素占比少,对应的特征信息少,受噪声和背景干扰大,导致分割精度下降;4)在图像边界分割时,上采样过程中会造成细节纹理信息损失,从而导致边界模糊或断裂。

基于此,本文提出一种基于全局感知与细节增强的非对称遥感建筑物分割网络(global perception and detail enhancement asymmetric-UNet, GPDEA-UNet)。该网络在UNet的基础上,首先构建了一个基于选择性状态空间的特征编码器模块提高特征表达能力,以视觉状态空间(visual state space,VSS)作为基础单元,并借助动态卷积分解(dynamic convolution decomposition,DCD),增强捕捉遥感图像中的复杂特征和上下文信息;其次引入多尺度双交叉融合注意力模块(multi-scale dual cross-attention,MDCA)实现特征的有效融合和增强,解决多尺度编码器特征间的通道与空间依赖性问题,缩小编解码器特征之间的语义差距;最后设计了一个细节增强解码器模块,使用DCD与级联上采样(cascade upsampling,CU)模块捕捉更丰富的语义信息,保留特征细节与语义完整,确保分割结果的精确性与细腻度。

本文贡献如下:1)构建了一种建立在状态空间模型(SSM)上的编码器结构,具有全局感受野和动态权重,有效增强了网络对复杂特征和上下文信息的捕捉能力,扩展了现有的CNN和ViT选择;2)通过多尺度双交叉融合注意力机制融合特征,解决跳跃连接引入的语义差异,以应对复杂的遥感图像建筑物特征提取;3)设计了一种保持细节信息和语义完整的解码器模块,使得网络能够在复杂场景中准确

分割出建筑物;4)融合状态空间模型与多尺度双交叉融合注意力机制,提出一种引入全局感知与细节增强的非对称遥感建筑物分割网络(GPDEA-UNet),在公开的遥感建筑物语义分割数据集WHU Aerial Imagery Datest (WHU)和Massachusetts (Building Dataset)上进行了对比实验,在不同指标上均取得了优异的性能,验证了所提算法的有效性。

## 1 研究方法

### 1.1 整体网络架构

遥感图像中光照、气候以及复杂背景等因素影响遥感建筑物语义分割的效果,导致出现分割区域不连续、边缘不平滑和误检漏检情况。为解决上述问题,本文提出一种基于全局感知与细节增强的非对称遥感建筑物分割网络GPDEA-UNet,所提网络采用基于UNet的编解码器结构,首先构建基于选择性状态空间的特征编码器模块提升特征的表达能力,其次引入多尺度双交叉融合注意力模块MDCA进行特征融合和增强,最后构建细节增强解码器模块CU保留特征细节与语义完整。GPDEA-UNet网络整体结构如图1所示,主要由基于选择性状态空间的特征编码器模块、多尺度双交叉融合注意力模块以及细节增强解码器模块组成。

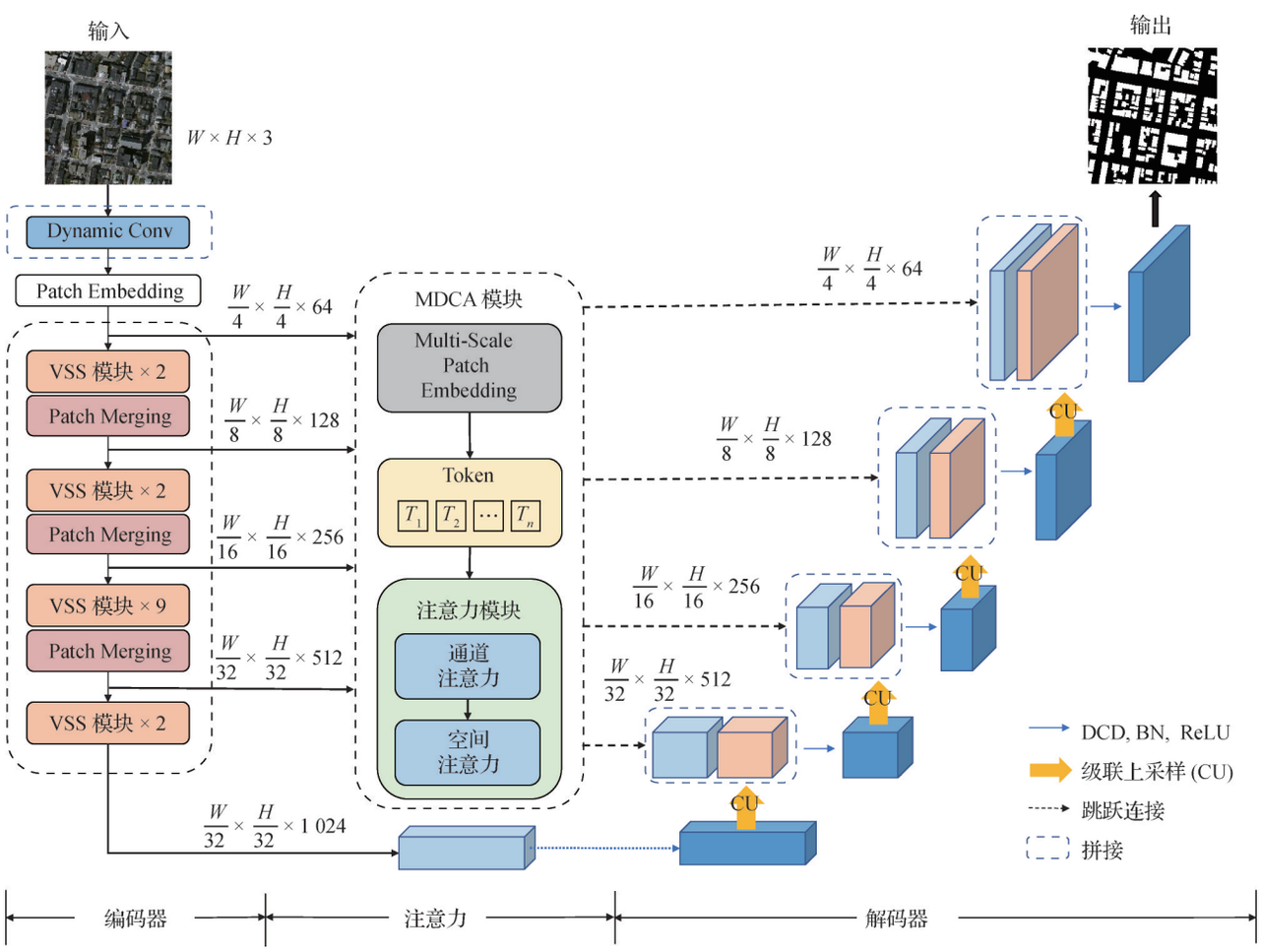



图1 GPDEA-UNet框架

Fig. 1 Schematic of GPDEA-UNet

### 1.2 基于选择性状态空间的特征编码器模块

为了有效捕捉全局上下文信息,降低计算复杂度,提出一种基于选择性状态空间的特征编码器模块。该模块以VSS为基本单元,可以在线性时间复杂度内有效捕捉图像中的关键信息,并借助DCD进一步提升网络的特征表达能力。其结构如图2所示,由Dynamic Conv层、Patch Embedding层以及4个Stage构成。

Dynamic Conv层是动态卷积分解层,主要负责在初始阶段对输入图像进行预处理,通过输入图像

的特征动态调整卷积核参数，使得模型能够更灵活地捕捉不同尺度、不同形态的特征。Patch Embedding层将输入图像划分为多个Patch，不同于ViT，没有将Patch进一步展平成1D序列，保留了图像的2D结构。具体来说，对于输入图像 $\boldsymbol{x} \in \mathbf{R}^{H \times W \times 3}$ 通过Dynamic Conv层提取浅层特征，然后通过Patch Embedding层将其划分为4 × 4的不重叠图像块后，将其图像维度映射到 $C$，得到 $\boldsymbol{x}^{*} \in \mathbf{R}^{\frac{H}{4} \times \frac{W}{4} \times 64}$，并进行归一化。Dynamic Conv层和Patch Embedding层的结合为模型提供了灵活的初始特征表示能力。

4个Stage主要用于特征提取，其核心在于VSS。前3个Stage都由多个VSS和Patch Merging组成，在VSS后应用Patch Merging操作，可以在减少输入特征高度和宽度的同时增加通道数。VSS由选择性扫描状态空间序列模型（2D selective scan，SS2D）、层归一化（layer norm，LN）、线性层（Linear）、深度可分离卷积（depthwise separable convolution，DWConv）和SiLU激活函数组成，其结构如图3所示。具体来说，当输入经过LN后，被拆分为两个分支。第1个分支经过Linear后直接通过Silu激活函数。第2个分支先经过Linear和3 × 3的深度卷积层DWConv，随后，通过SiLU激活函数进入核心SS2D模块进行进一步特征提取。深度可分离卷积可以实现高效的特征提取，接着，SS2D的输出通过LN进行归一化后与第1个分支的输出执行逐元素生成，以合并两个路径。最后，使用Linear混合特征，并与残差连接相结合以形成VSS的输出。

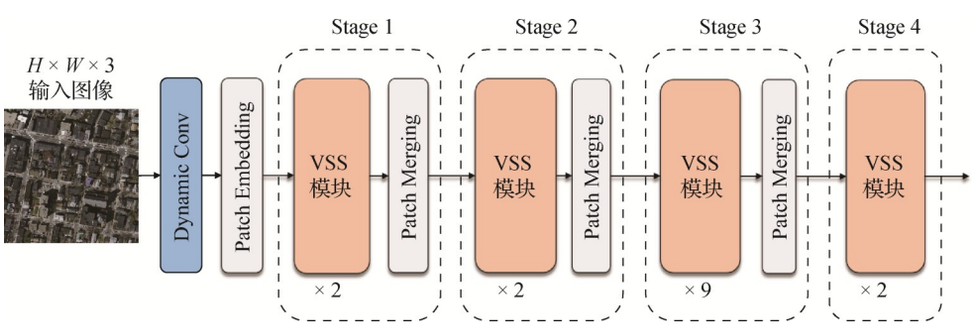



图2　基于选择性状态空间的特征编码器模块结构

Fig. 2　Schematic of feature encoder module based on selective state space

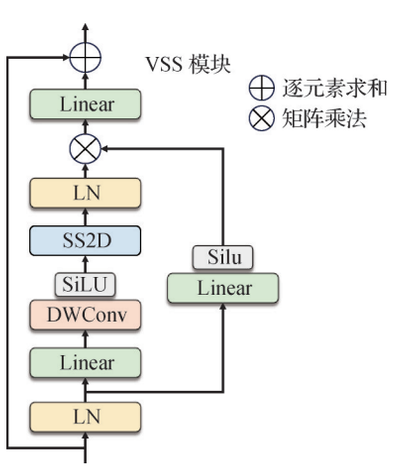



图3　VSS结构

Fig. 3　Schematic of VSS

SS2D是VSS的核心计算单元，由扫描扩展操作（scan expanding）、选择性扫描空间状态序列模型（selective scan space state sequential model，S6）和扫描合并操作（scan merging）3个部分组成。其能够沿着多个扫描路径动态地选择与图像Patch相关的上下文信息，通过跨路径融合将不同路径上的信息整合到一起，形成一个全局的上下文表示，从而在保持全局感受野的同时显著降低了计算成本。具体来说，先使用扫描扩展操作沿4个不同方向（左上到右下、右下到左上、右上到左下、左下到右上）将Patch展开为序列；再由S6进行特征提取，确保来自各个方向的信息得到彻底扫描，形成全局感受野，从而捕获不同的特征；最后扫描合并操作对来自4个方向的序列求和后合并，将输出图像恢复到与输入相同的大小。其中，S6可以根据输入调整SSM的参数，在结构化序列空间模型基础上引入选择机制使模型能够在区分并保留相关信息的同时过滤不相关信息，并且S6使一维数组中的每个元素通过压缩隐藏状态与先前扫描的任何样本进行交互，将二次复杂度降为线性。

### 1.3 多尺度双交叉融合注意力模块

为进一步拓宽网络的全局视野，同时解决跳跃连接时在编解码器引起的语义差异问题，提出一种多尺度双交叉融合注意力模块MDCA，其结构如图4所示，由Multi-scale Patch Embedding、Tokens和双交叉融合注意力机制（dual cross attention，DCA）组成。

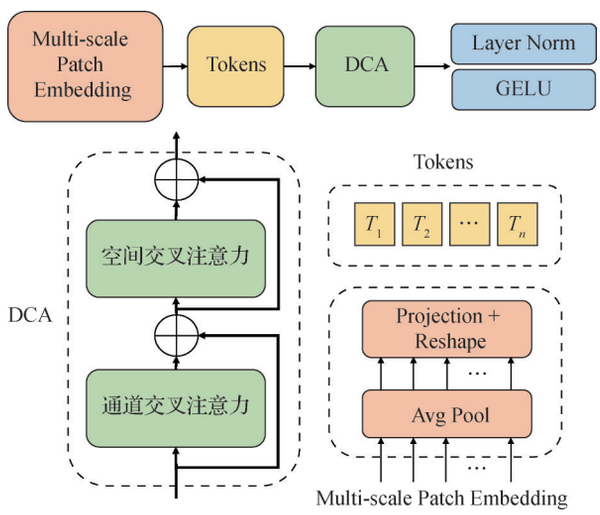



图4 多尺度双交叉融合注意力模块结构

Fig. 4 Schematic of MDCA

Multi-scale Patch Embedding将输入特征图划分为一系列多尺度Patch，并将每个Patch转换为特征向量。多尺度表示同时考虑不同大小和分辨率的Patch，这有助于捕捉图像中的不同细节和上下文信息。Tokens是指从Multi-scale Patch Embedding中获得的特征向量集合，其作为后续注意力机制处理的输入，携带了图像的关键信息。DCA是整个模块的核心，由通道交叉注意力（channel cross attention，CCA）和空间交叉注意力（spatial cross attention，SCA）两部分组成。CCA利用多尺度编码器特征的跨通道Token进行交叉注意力，通过计算不同通道之间的相关性来增强这些特征之间的交互，从而提高模型对通道间依赖关系的理解能力，实现全局信息的整合。SCA利用跨空间Token计算图像不同位置之间的相似性或相关性来增强模型对空间上下文信息的捕捉能力。最后将DCA的输出进行LN和GELU后，通过跳跃连接将它们连接到解码器对应部分。

具体来说，Multi-scale Patch Embedding从$n$个多尺度编码器Stage中提取Patch，设$n$个多尺度的编码器Stage为$E_i \in \mathbf{R}^{C_i \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$，块大小为$P_i^S = \frac{P^S}{2^{i-1}}$，其中，$i = 1, 2, 3, \cdots, n$，使用大小和步长为$P_i^S$的平均池化提取Patch，并在展平的二维Patch上使用$1 \times 1$的深度可分离卷积进行映射，表达为

$$T_i = DConv1D_{E_i}\left(Reshape\left(AvgPool2D_{E_i}\left(E_i\right)\right)\right) \tag{1}$$

式中，$T_i \in \mathbf{R}^{p \times C_i}$表示第$i$个编码器Stage展平的Patch；$p$代表Patch的数量。

通道交叉注意力模块结构如图5（a）所示，使用CCA对每个Token$T_i$进行处理。首先对每个$T_i$进行层归一化，接着沿通道维度对$T_i$进行拼接，得到$T_c$，从而生成Key和Value，同时使用$T_i$作为Query。对Query、Key、Value使用$1 \times 1$深度可分离卷积进行投影映射，分别为

$$\boldsymbol{Q}_i = DConv1D_{Q_i}\left(T_i\right) \tag{2}$$

$$\boldsymbol{K} = DConv1D_{K}\left(T_c\right) \tag{3}$$

$$\boldsymbol{V} = DConv1D_{V}\left(T_c\right) \tag{4}$$

式中，$\boldsymbol{Q}_i \in \mathbf{R}^{p \times C_i}$，$\boldsymbol{K} \in \mathbf{R}^{p \times C_c}$，$\boldsymbol{V} \in \mathbf{R}^{p \times C_c}$，分别映射Queries、Keys和Values。使用深度可分离卷积可以捕获局部信息并降低计算复杂度。即CCA可表示为

$$CCA\left(\boldsymbol{Q}_i, \boldsymbol{K}, \boldsymbol{V}\right) = softmax\left(\frac{\boldsymbol{Q}_i^{\mathrm{T}}\boldsymbol{K}}{\sqrt{C_c}}\right)\boldsymbol{V}^{\mathrm{T}} \tag{5}$$

式中，$\frac{1}{\sqrt{C_c}}$为比例因子。CCA的输出是Values的加权和，权重由Queries和Keys之间相似性决定。

空间交叉注意力模块结构如图5（b）所示。将CCA的输出沿通道维度进行层归一化和拼接，拼接后的Token$T_i^*$作为Query和Key，每个$T_i^*$作为Value，对Query、Key、Value使用$1 \times 1$深度可分离卷积进行投影映射，表达式分别为

$$\boldsymbol{Q}_i = DConv1D_{Q}\left(T_i^*\right) \tag{6}$$

$$\boldsymbol{K} = DConv1D_{K}\left(T_i^*\right) \tag{7}$$

$$\boldsymbol{V} = DConv1D_{V_i}\left(T_i^*\right) \tag{8}$$

即SCA可以表示为

$$SCA\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}_i\right) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_i \tag{9}$$

式中，$\frac{1}{\sqrt{d_k}}$为比例因子，当为多头时$d_k = \frac{c_c}{h_c}$，$h_c$是头（head）的数目。

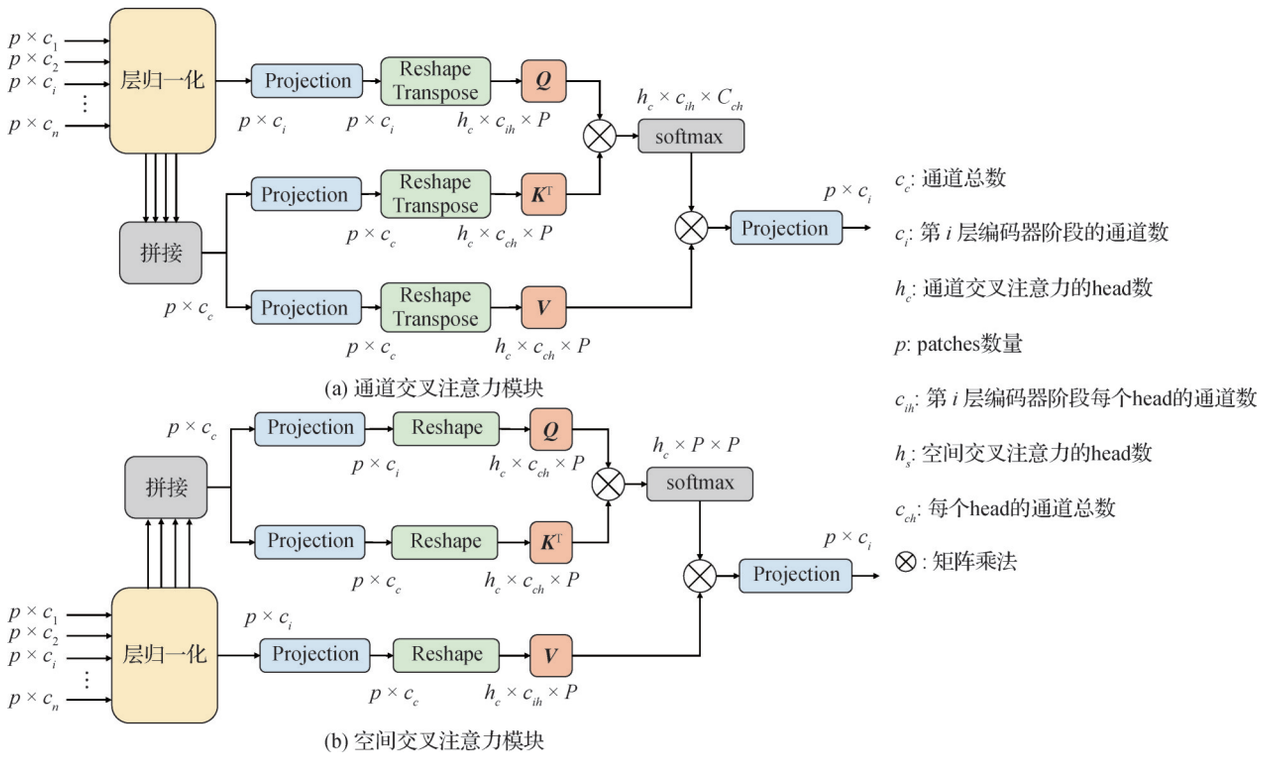



图 5　通道交叉注意力和空间交叉注意力模块结构

Fig. 5　Schematic of CCA and SCA ((a) channel cross attention; (b) spatial cross attention)

### 1.4　细节增强解码器模块

为了解决上采样过程中图像细节信息丢失的问题，提出一个细节增强解码器模块，以确保分割结果的精确性与细腻度。在 DCD 的基础上，设计了级联上采样模块 CU，结构如图 6 所示。

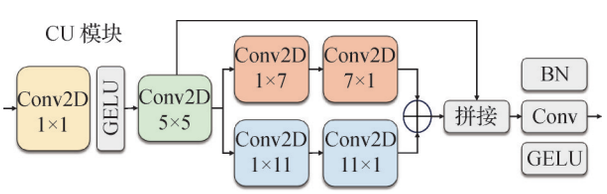



图 6　级联上采样模块结构

Fig. 6　Schematic of CU

级联上采样模块采用了级联的结构进行衔接，具体步骤如下：

1）使用不同尺寸的卷积核进行分层次学习。通过引入 1 × 1 卷积进行升维，随后使用 5 × 5 的卷积核以及一系列分组卷积（如 1 × 7, 7 × 1, 1 × 11, 11 × 1），使模型能够同时捕捉图像的局部细节和全局结构。这种多尺度策略有助于模型理解和识别图像中的复杂目标，并且不同尺寸的卷积核为模型提供了灵活性，使其能够适应不同大小和形状的特征，从而提高了模型对图像内容的适应性。

2）进行分组卷积后结果的级联与融合。将两次分组卷积的结果建立级联操作，并与 5 × 5 卷积的结果进行融合。这一过程可以将不同尺度和不同角度的特征结合起来，增强其特征表示的丰富性和鲁棒性，并且多尺度卷积的结合可以使模型捕捉到更广泛的上下文信息，提升模型对图像中建筑物的理解能力。特别地，不同分组卷积捕获的信息具有互补性，融合这些信息可以更全面地描述图像内容。

3）使用 1 × 1 卷积层进行降维。1 × 1 的卷积层主要用于调整通道数，起到降低后续层参数数量和计算复杂度的作用。虽然 1 × 1 的卷积在空间上不改变特征图的大小，但它通过跨通道的点积操作可以实现通道间的信息融合和交互，提高学习通道间依赖关系的能力。同时通过降维操作，模型可以去除冗余信息，保留对目标建筑物最有用的特征，从而提高模型的泛化能力和识别能力。此外，这种压缩和凝练过程也可以防止出现过拟合现象。

## 2　实验结果与分析

### 2.1　实验环境及参数设置

实验平台配置 Inter Xeon E52650 处理器，50 GB

内存和NVIDIA GeForce RTX 2080Ti 12 G的显卡；采用Python3.6以及PyTorch 1.7的深度学习框架，使用CUDA11.2以及cuDNN8.0的深度学习GPU加速库。

在GPDEA-UNet网络的训练过程中，设置输入遥感图像尺寸为512 × 512像素与256 × 256像素。采用随机翻转与归一化技术增强数据多样性和模型泛化能力。训练参数方面，设置批量大小为12、训练周期为200。优化器使用AdamW，并设置初始学习率为$1 \times 10^{-3}$，还使用余弦学习率调整策略。实验损失函数采用二进制交叉熵损失函数。

### 2.2 实验数据集和评价指标

实验使用WHU Aerial Imagery Dataset和Massachusetts Building Dataset两个公开遥感建筑物数据集。

WHU Aerial Imagery Dataset由新西兰基督城的高分辨航空图像组成。为了方便模型的训练和验证，数据集被裁剪成尺寸为512 × 512像素的8 188幅遥感图像，其中训练集4 736幅、验证集1 036幅、测试集2 416幅。WHU Aerial Imagery Dataset中部分样本及其对应标签示例如图7所示。

Massachusetts Building Dataset由波士顿地区的航拍图像组成。为了方便训练与验证，采用边缘重叠方法将原始图像裁剪成尺寸为256 × 256像素的子图，其中训练集、验证集、测试集分别为4 412、144、360幅。Massachusetts Building Dataset中部分样本及其对应标签示例如图8所示。

为了验证所提网络的有效性，采用精确度(precision)、召回率(recall)、F1分数(F1-score)和交并比(intersection over union，IoU)作为评估指标，对网络进行有效性评价。

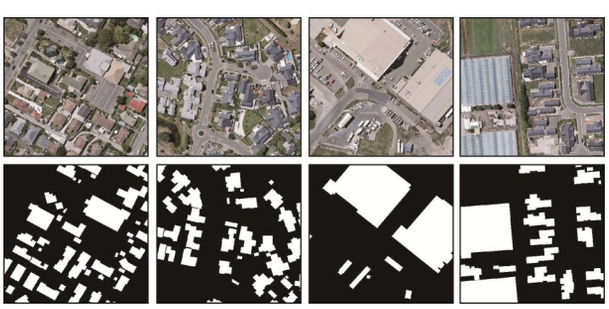


图7 WHU Aerial Imagery Dataset样本及其标签

Fig. 7 Sample images and labels of WHU Aerial Imagery Dataset

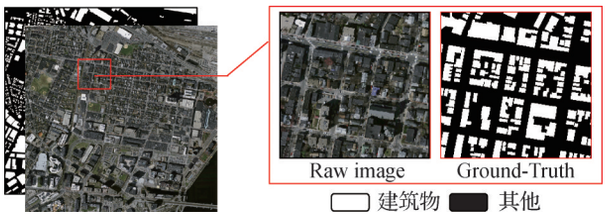



图8 Massachusetts Building Dataset样本及其标签

Fig. 8 Sample images and labels of Massachusetts Building Dataset

### 2.3 消融实验

#### 2.3.1 定量评价指标对比

为了验证所提出的GPDEA-UNet网络各模块的有效性，以UNet作为基线模型(baseline)，在WHU Aerial Imagery Dataset上开展消融实验，并对变种模型的精确度、召回率、F1分数、IoU指标进行对比。各变种模型的模块配置如表1所示，在消融实验中，除了模块搭配，网络均采用相同的参数设置和训练策略。消融实验结果如表2所示。

分析表1和表2可知：

1)对比baseline，网络加入动态卷积、多尺度双交叉注意力网络、选择性状态空间模型和级联上采样后，性能有较为明显的提升。由表2可知，baseline+DCD比baseline分别提高0.49%(IoU)、1.66%(precision)和0.27%(F1-score)；baseline+MDCA比baseline分别提高1.13%(IoU)、0.93%(precision)和0.63%(F1-score)；baseline+SSM比baseline分别提高1.70%(IoU)、1.22%(precision)、0.21%(recall)和0.94%(F1-Score)；baseline+CU比baseline分别提高2.18%(IoU)、0.37%(precision)、1.58%(recall)和1.21%(F1-score)。数据显示所加入的模块都能提高网络对于遥感图像建筑物的提取精度，证明了所提模块的有效性。

**表1 GPDEA-UNet模型变种结构**

**Table 1 Variant structure of GPDEA-UNet**

| 网络变体模块 | 描述 |
|---|---|
| baseline | 基线模型UNet |
| DCD | 动态卷积 |
| MDCA | 多尺度双交叉注意力网络 |
| SSM | 选择性状态空间模型 |
| CU | 级联上采样模块 |

2）对比 baseline+SSM 和 baseline+DCD+SSM 的性能，前者是在baseline上直接引入选择性状态空间模型，后者相较于前者又引入动态卷积，提出基于选择性状态空间模型与动态卷积改进的特征编码模型的策略，分别在 IoU 上提升 0.98%、recall 上提升 1.28%、F1-score上提升0.54%。数据证明引入基于选择性状态空间的特征编码器模型策略可有效增强捕捉遥感图像中的时序依赖性和空间上下文信息，提升模型对动态场景的理解能力，且可动态调整卷积核参数，提高模型的灵活性和泛化能力。

3）对比 baseline 和 baseline+MDCA、baseline+DCD 和 baseline+DCD+MDCA、baseline+SSM 和 baseline+MDCA+SSM 3种组合的性能，后者都是在前者的基础上引入多尺度双交叉注意力网络，即交叉通道注意力和交叉空间注意力的注意力加权模块。baseline+MDCA 比 baseline 分别提高 1.13%（IoU）、0.93%（precision）和 0.63%（F1-score）；baseline+DCD+MDCA 比 baseline+DCD 分别提高 1.96%（IoU）、2.92%（recall）和 1.09%（F1-score）；baseline+MDCA+SSM 比baseline+SSM 分别提高 1.42%（IoU）、1.86%（recall）和 0.78%（F1-score）。数据证明加入多尺度双交叉注意力网络能够增强捕捉遥感图像中的细节信息和上下文依赖的能力，从而在分割过程中减少误判和漏判。

4）对比 baseline 和 baseline+CU、baseline+DCD+MDCA+SSM 和baseline+DCD+MDCA+SSM+CU（本文方法）两组的性能，后者都是在前者的基础上加入了细节增强特征解码器模块。baseline+CU 比 baseline 分别提高 2.18%（IoU）、0.37%（precision）、1.58%（recall）和 1.21%（F1-score）；baseline+DCD+MDCA+SSM+CU 比 baseline+DCD+MDCA+SSM 分别提高 0.04%（IoU）、0.10%（recall）和 0.03%（F1-score）。数据表明加入了细节增强特征解码器模块能够对低层次特征进行自适应调整，使其与高层次特征更好地融合，且能够增强上采样过程中图像的细节信息，使得分割结果更加精细和准确。

**表 2　消融实验结果**

**Table 2　Ablation experimental results**

/%

| 方法 | baseline | DCD | MDCA | SSM | CU | IoU | precision | recall | F1-score |
|---|---|---|---|---|---|---|---|---|---|
| baseline | √ | – | – | – | – | 88.26 | 94.35 | 93.66 | 93.77 |
| baseline + DCD | √ | √ | – | – | – | 88.75 | **96.01** | 92.14 | 94.04 |
| baseline + MDCA | √ | – | √ | – | – | 89.39 | 95.28 | 93.53 | 94.40 |
| baseline + SSM | √ | – | – | √ | – | 89.96 | 95.57 | 93.87 | 94.71 |
| baseline + CU | √ | – | – | – | √ | 90.44 | 94.72 | 95.24 | 94.98 |
| baseline + DCD + MDCA | √ | √ | √ | – | – | 90.71 | 95.19 | 95.06 | 95.13 |
| baseline + DCD + SSM | √ | √ | – | √ | – | 90.94 | 95.36 | 95.15 | 95.25 |
| baseline + MDCA + SSM | √ | – | √ | √ | – | 91.38 | 95.25 | 95.73 | 95.49 |
| baseline + DCD + MDCA + SSM | √ | √ | √ | √ | – | 91.56 | 95.39 | 95.79 | 95.59 |
| baseline + DCD + MDCA + SSM + CU（本文） | √ | √ | √ | √ | √ | **91.60** | 95.36 | **95.89** | **95.62** |

注：加粗字体表示各列最优结果。“√”表示使用对应模块，“–”表示不使用对应模块。

### 2.3.2　可视化结果对比

为进一步验证提出的GPDEA-UNet的有效性，在WHU Aerial Imagery Dataset中选择了有代表性的建筑图像对GPDEA-UNet进行可视化，可视化结果如图9所示。

GPDEA-UNet由DCD、SSM、MDCA和CU模块组成。使用UNet作为baseline，baseline+DCD、baseline+SSM、baseline+MDCA、baseline+CU分别为仅加入动态卷积、选择性状态空间模型、多尺度双交叉注意力网络和级联上采样模块，本文方法为DCD、SSM、MDCA和CU 4个模块的组合。从baseline+DCD、baseline+SSM的可视化图中可以看出，当网络仅有动态卷积和选择性状态空间模型时，由于缺乏捕捉全局上下文信息的能力，难以有效区分前景中

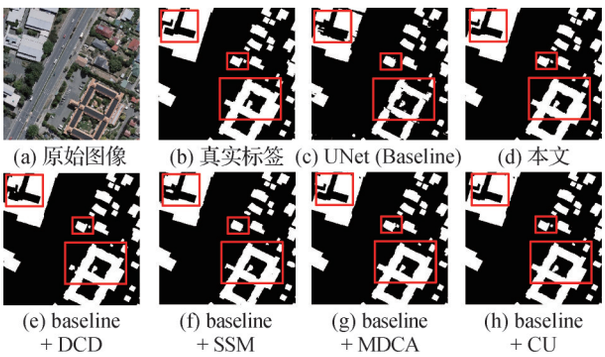



图9　GPDEA-UNet模型各模块可视化结果

Fig. 9　Visualization results of modules of GPDEA-UNet ((a)original images; (b)ground truth; (c)UNet(baselinc); (d)ours; (e)baseline+DCD; (f)baseline+SSM; (g)baseline+MDCA; (h)baseline+CU)

的建筑物与背景噪声，进而导致了建筑物整体轮廓的误识别与检测偏差。从baseline+SSM、baseline+MDCA的可视化图中可以看出，当网络仅有多尺度双交叉注意力和选择性状态空间模型时，由于对细节信息的敏感度不足，导致分割中常出现小目标的欠分割问题。从baseline+CU的可视化图中可以看出，当网络仅有级联上采样模块时，由于对局部细节信息的敏感度不足，所提取的建筑物形状边缘显得较为模糊和粗糙。在baseline基础上同时添加DCD、SSM、MDCA和CU模块后，网络的性能显著提升，可以准确提取到建筑物相互粘连、模糊细小和前景背景相似等区域。综上所述，所提出的网络在不同典型建筑物分布情况下，不仅能够强化模型对待分割区域的关注度，而且抑制了遥感图像背景区域的显著程度，从而有效地提取待分割建筑物的特征和建筑物边缘特征信息，提升了建筑物分割的精度与效果。

为了深入探索GPDEA-UNet网络对建筑物的关注程度，在解码器上采样过程中对Stage1和Stage3两个阶段的特征图进行可视化，可视化结果如图10所示。选择了4种代表性的建筑物地图：1)复杂结构、2)小型密集建筑群、3)有阴影的建筑物、4)材质变换和屋顶阴影。在较大尺度的遥感图像中，小型密集建筑群易出现对小型目标区域及其边缘的关注度不足，导致小型密集区域出现分割不足的问题。而对于背景与待分割区域颜色近似的建筑物，则易发生过度分割现象。此外，由于建筑物易受噪声、纹理等复杂因素的干扰，常会引发分割区域出现斑块误判及边缘界定模糊等问题。

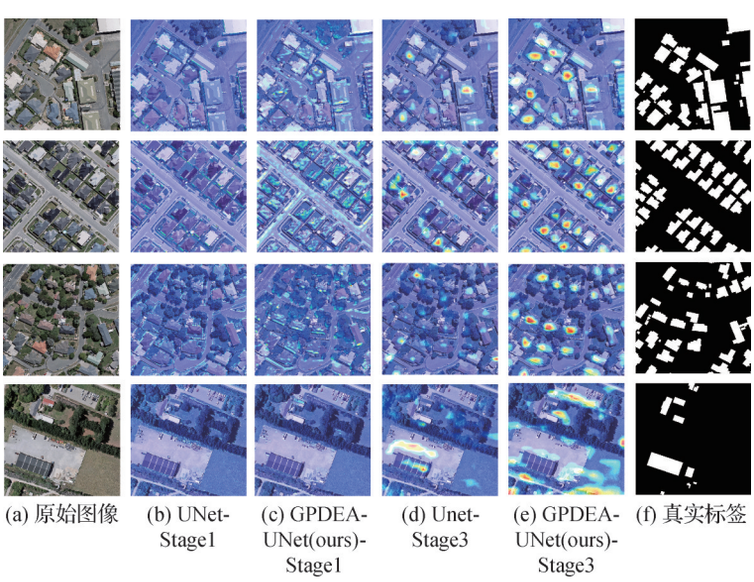



图10　GPDEA-UNet模型区域注意力可视化结果

Fig. 10　Visualization results of region attention module of GPDEA-UNet((a)original images; (b)UNet-Stage1; (c)GPDEA-UNet(ours)-Stage1; (d)UNet-Stage3; (e)GPDEA-UNet(ours)-Stage3; (f)ground truth)

如图10所示，GPDEA-UNet中解码器模块的Stage1直接对深层特征处理，由于其具有较大的局部感受野，故其能提取到更为丰富的遥感建筑物语义信息。相较于原始UNet的特征图，GPDEA-UNet中解码器模块的Stage1可有效提高待分割建筑物群区域的显著性，增强网络在区分遥感图像不同区域间相关性的表征能力和对待分割区域的关注能力，有效减少深层特征图在分辨率恢复过程中背景对待

分割区域特征的干扰，实现对背景的抑制与对待分割区域的增强。

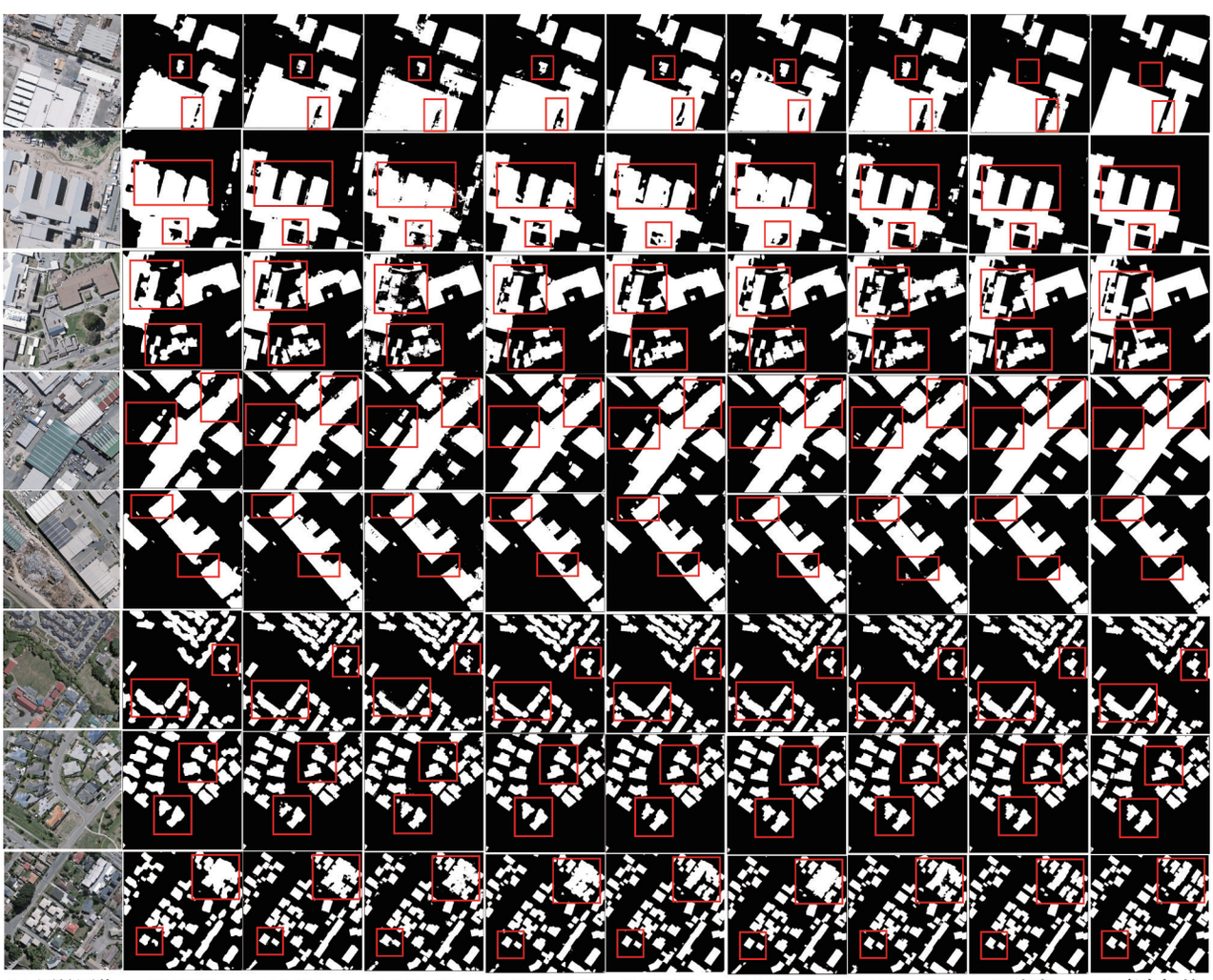
GPDEA-UNet 中解码器的 Stage3 对浅层特征进行处理，由于其具有较小局部感受野，故其对遥感建筑物空间细节信息的提取能力更强。相较于原始 UNet 的特征图，GPDEA-UNet 的 Stage3 能够精确捕捉到待分割建筑物的形状轮廓与边缘特征，同时有效削弱背景的干扰。可视化实验结果显示，在各种复杂的建筑物分布场景中，所提出的双交叉注意力机制能更好地关注并提取出建筑物及其边缘信息。

## 2.4 对比实验

为了验证所提 GPDEA-UNet 网络的有效性，与主流语义分割网络在 WHU Aerial Imagery Dataset 和 Massachusetts Building Dataset 上进行了建筑物分割对比实验。对比方法包括 UNet（Ronneberger 等，2015）、PSPNet（pyramid scene parsing network）（Zhao 等，2017）、HR-Net（high-resolution network）（Sun 等，2019）、Res-UNet（Diakogiannis 等，2020）、DR-Net（dense residual neural network）（Chen 等，2021b）、UNetFormer（Wang 等，2022）、MSL-Net（multi-scale level network）（Qiu 等，2022）、PPANet（parallel path attention net）（杨坚华 等，2023）、RH-CUNet（ridge and corner-embedded hierarchical convolutional U-Net）（朱梓萌 等，2024）和 EAMFNet（edge-attention guided multi-scale fusion network）（董杰 等，2024）。

### 2.4.1 WHU Aerial Imagery Dataset

图 11 为各类对比方法在 WHU Aerial Imagery Dataset 上部分图像的分割结果，图 11（a）为输入测



(a) 原始图像 (b) FCN (c) UNet (d) SegNet (e) DeeplabV3 (f) PSPNet (g) HR-Net (h) Res-UNet (i) 本文 (j) 真实标签

图 11 WHU Aerial Imagery Dataset 定性对比实验结果

Fig. 11 Qualitative comparison results of WHU Aerial Imagery Dataset ((a)original images; (b)FCN; (c)UNet; (d)SegNet; (e)DeeplabV3; (f)PSPNet; (g)HR-Net; (h)Res-UNet; (i)ours; (f)ground truth)

试的遥感建筑物图像，图11(b)-(h)分别为FCN、UNet、SegNet、DeeplabV3、PSPNet、HR-Net和Res-UNet网络的语义分割结果，图11(i)为本文所提GPDEA-UNet的语义分割结果，图11(j)为真实标签图像。

1)定性分析。遥感建筑物在语义分割过程中面临多重挑战，特别是建筑物小、密集、相互粘连、前景复杂与背景相似等问题，这些问题均显著影响分割的准确性和质量。如图11第1、3、5行所示，图像中存在建筑物颜色与背景颜色较为相似的问题；如图11第2、3行所示，存在建筑物相互粘连情况，粘连部分相对于建筑物整体大小较为细小，且粘连处出现颜色与背景颜色相似的问题；如图11第6、7、8行所示，存在建筑物较小、分布密集连续且小建筑物模糊的问题；如图11第4行所示，存在前景颜色较为复杂、与背景颜色较为相似的问题。光照与阴影同样也会影响建筑物的分割性能，如图11第2行所示，图像中建筑物的边界受到光照所产生阴影的影响。

相较于FCN、UNet、SegNet、DeeplabV3、PSPNet、HR-Net和Res-UNet等方法，所提GPDEA-UNet通过构建基于选择性状态空间的特征编码器模型，引入SSM和DCD，更好地提升了特征编码的鲁棒性，进一步提高了区分相似物体的能力。如图11第1、3、5行所示，使用GPDEA-UNet网络能够有效地解决建筑物颜色与背景颜色相似问题，提高分割的准确性。对于建筑物相互粘连问题，如图11第2、3行所示，与其他网络相比，GPDEA-UNet通过引入多尺度双交叉注意力模块MDCA，不仅能更好地捕捉遥感图像中的细节信息和上下文依赖的能力，而且能够有效解决由建筑物粘连带来的分割不精确问题，更清晰地分割出了图中粘连的建筑物，而且其分割边缘更为清晰。遥感建筑物图像中由于建筑物的大小不同，导致分割中常出现小目标的欠分割问题，如图11第6、7、8行所示，本网络的细节增强解码器模块，通过DCD自适应调整卷积核参数，使得模型更好地适应不同尺度、不同形态的目标特征，又利用级联上采样逐步细化特征图，保留了更多细节信息，避免了直接上采样带来的信息损失。相较于其他网络能进一步提高小目标建筑物的分割效果，从而解决建筑物较小、分布密集连续且小建筑物模糊的挑战。由于遥感建筑物图像中常出现边界受到光照、阴影与复杂前景颜色的干扰，如图11第2、4行所示，相较于其他网络，所提网络能有效注意到各类干扰下的建筑物边界，得到了较好的建筑物分割结果，因此所提网络对于光照、阴影与复杂前景颜色影响具有较强的鲁棒性。

从WHU Aerial Imagery Dataset的对比实验结果可以看出，与FCN、UNet、SegNet、DeeplabV3、PSPNet、HR-Net和Res-UNet等方法相比，所提出的网络在复杂场景下的遥感建筑物图像语义分割问题上，不仅能够获得最好的分割结果，并且对于具有光照阴影干扰、建筑物粘连、目标前景与背景相似且颜色干扰以及小建筑物密集模糊等问题干扰具有较强的鲁棒性。

2)定量分析。不同方法在WHU Aerial Imagery Dataset数据集上分割结果的定量分析表如表3所示，定量分析所采用的指标主要为交并比(IoU)、精确度(precision)、召回率(recall)和F1分数(F1-score)。由表3可知，提出的GPDEA-UNet在Aerial Imagery Dataset数据集上IoU为91.60%，precision为95.36%，recall为95.89%。第1～8行分别为UNet、HR-Net、Res-Unet、DR-Net、UNetFormer、MSL-Net、PPANet、RH-CUNet等对比方法的语义分割量化指标。实验表明，相较于UNet、HR-Net、Res-UNet、DR-Net、UNetFormer、MSL-Net、PPANet、RC-CUNet等对比方法，IoU分别提升3.34%、5.09%、3.55%、

**表3 WHU Aerial Imagery Dataset数据集定量对比实验结果**

**Table 3 Quantitative comparison results of WHU Aerial Imagery Dataset**

/%

| 方法 | IoU↑ | precision↑ | recall↑ | F1-score↑ |
|---|---|---|---|---|
| UNet | 88.26 | 94.35 | 93.66 | 93.77 |
| HR-Net | 86.51 | 92.67 | 93.66 | 93.16 |
| Res-UNet | 88.05 | 94.48 | 92.84 | 93.65 |
| DR-Net | 88.30 | 94.30 | 94.30 | 93.80 |
| UNetFormer | 89.66 | 94.82 | 94.34 | 94.57 |
| MSL-Net | 90.40 | 95.00 | 95.10 | 94.80 |
| PPANet | 90.45 | 94.90 | 95.07 | 94.98 |
| RH-CUNet | 90.52 | 90.56 | 90.63 | 90.83 |
| 本文 | **91.60** | **95.36** | **95.89** | **95.62** |

注：加粗字体表示各列最优结果。“↑”表示值越大越好。

3. 30%、1. 94%、1. 20%、1. 15%、1. 08%；precision 分别提升 1. 01%、2. 69%、0. 88%、1. 06%、0. 54%、0. 36%、0. 46%、4. 80%；recall 分别提升 2. 23%、2. 23%、3. 05%、1. 59%、1. 55%、0. 79%、0. 82%、5. 26%；F1-Score 分别提升 1. 85%、2. 46%、1. 97%、1. 82%、1. 05%、0. 82%、0. 64%、4. 79%。

WHU Aerial Imagery Dataset 的分割实验对比结果表明，所提 GPDEA-UNet 网络在有效性指标 precision、recall、IoU 和 F1-score 上均达到最优。因此，与先进的网络相比，所提网络能有效实现复杂场景下的遥感建筑物图像的准确分割。

2. 4. 2　Massachusetts Building Dataset

为进一步验证所提网络的普适性，在 Massachusetts Building Dataset 上对 GPDEA-UNet 与主流分割网络进行了对比实验。图 12 为各类对比方法的部分分割结果对比。

1）定性分析。由于 Massachusetts Building Dataset 整体图像质量较差，因而在此数据集上进行可视化实验更能直观表现各网络的效果。如图 12 第 1、2 行所示，由于光照产生的阴影造成建筑物边界与颜色等特征不显著；如图 12 第 3、5、8 行所示，图像前景与背景颜色较为相似，导致建筑物未被识别，存在漏检现象；如图 12 第 4 行所示，由于建筑物被遮挡，识别不完全；如图 12 第 6、7 行所示，小型建筑物分布较为密集且存在成像模糊的情况，导致建筑物粘连且小建筑物难识别。相较于 FCN、UNet、SegNet、DeeplabV3、PSPNet、Res-UNet 等对比网络，所提出的网络 GPDEA-UNet 获得最好的分割结果。图 12 中红框所

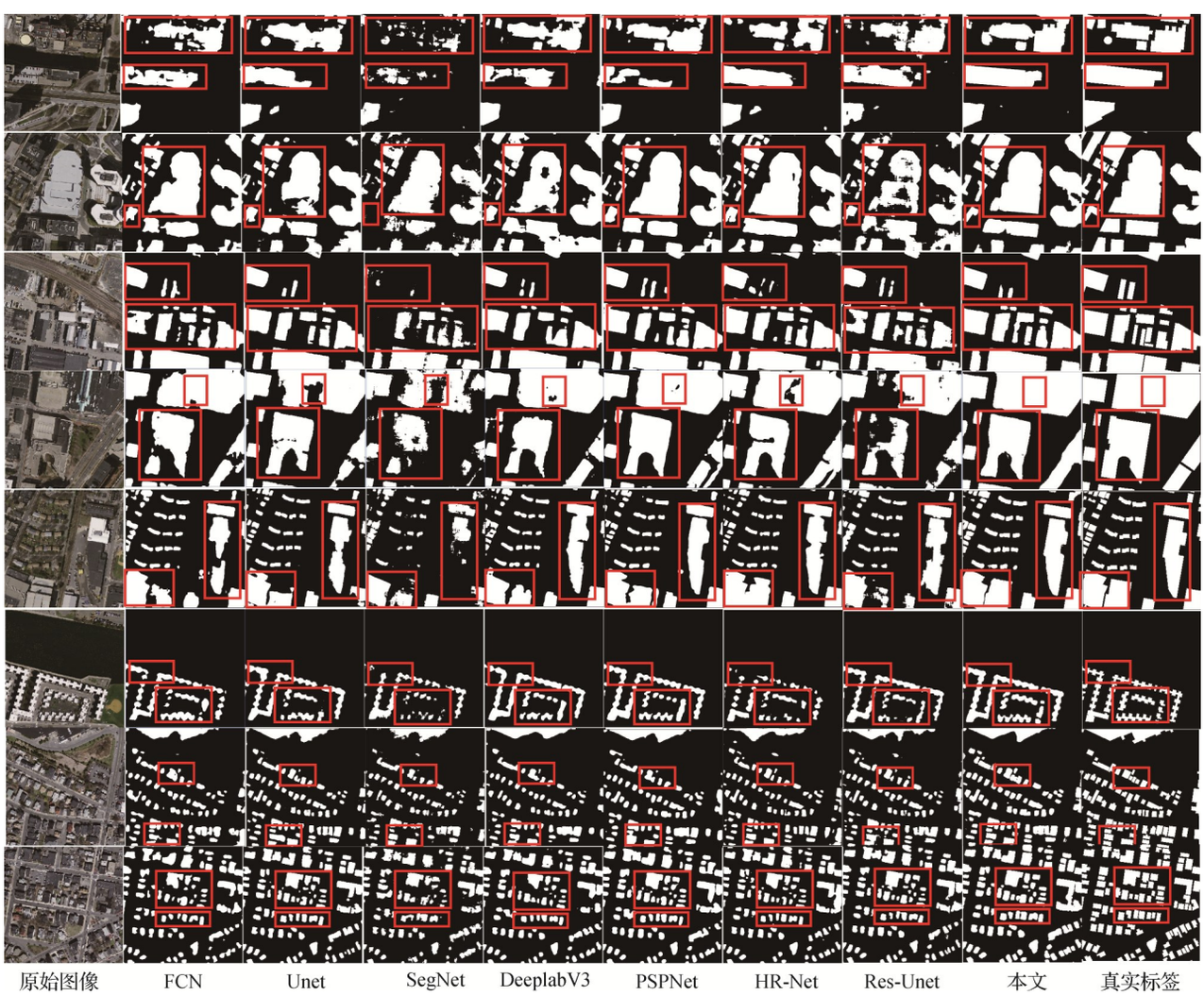



图 12　Massachusetts Building Dataset 定性对比实验结果

Fig. 12　Qualitative comparison results of Massachusetts Building Dataset ((a) original images; (b) FCN; (c) UNet; (d) SegNet; (e) DeeplabV3; (f) PSPNet; (g) HR-Net; (h) Res-UNet; (i) ours; (f) ground truth)

标记的分割对比结果显示，所提网络不仅获得了更准确的建筑物分割区域，而且建筑物分割的轮廓更为清晰，更接近于真实标签图像。这是由于所提的基于选择性状态空间的特征编码器模块能更好地增强网络对建筑物图像语义特征的提取能力，并且通过加入多尺度双交叉注意力模块，增强了模型建筑物局部区域和轮廓的关注程度，提升了模型的分割效果，还通过细节增强的解码器模块保留了特征细节，确保了语义的完整性。

从定性对比可知，相比于对比网络，所提网络在Massachusetts Building Dataset上，具有较好的分割效果，并且对于成像质量较差的遥感图像能进行较高质量的分割。

2）定量分析。对比方法在Massachusetts Building Dataset上分割结果如表4所示，定量分析所采用的指标主要为交并比（IoU）、精确度（precision）、召回率（recall）和F1分数（F1-score）。

**表4　Massachusetts Building Dataset定量对比实验结果**

**Table 4　Quantitative comparison results of Massachusetts Building Dataset**

/%

| 方法 | IoU | precision | recall | F1-score |
|---|---|---|---|---|
| UNet | 69.98 | 80.36 | 84.40 | 82.34 |
| PSPNet | 68.04 | 79.76 | 82.24 | 81.00 |
| HR-Net | 67.89 | 79.51 | 82.28 | 80.87 |
| Res-UNet | 66.21 | 76.97 | 82.58 | 79.67 |
| UNetFormer | 71.83 | 83.60 | 85.33 | 81.95 |
| MSL-Net | 70.90 | 83.00 | 81.90 | 84.10 |
| PPANet | 72.84 | **84.81** | 81.91 | **84.29** |
| EAMFNet | 71.60 | 80.13 | 80.61 | 82.45 |
| 本文 | **73.51** | 79.44 | **86.81** | 82.53 |

注：加粗字体表示各列最优结果。

由表4可知，相较于UNet、PSPNet、HR-Net、Res-UNet、UNetFormer、MSL-Net、PPANet、EAMFNet等对比方法，本文网络对比指标IoU分别提升3.53%、5.47%、5.62%、7.30%、1.68%、2.61%、0.67%、1.91%；指标recall分别提升2.41%、4.57%、4.53%、4.23%、1.48%、4.91%、4.90%、6.20%；相较于UNet、PSPNet、HR-Net、Res-UNet、UNetFormer、EAMFNet等对比方法，本文网络指标F1-score分别提升0.19%、1.53%、1.66%、2.86%、0.58%、0.08%。

从定量指标可知，本文所提GPDEA-UNet在Massachusetts Building Dataset上，虽然在precision上略有欠佳，但在IoU、recall和F1-score均达到了最优或次优水平。

2.4.3　模型复杂度与效率分析

为了定量评估模型的复杂度与效率，本文采用每秒帧数（frames per second，FPS）和模型参数量（Params，单位M）指标衡量模型复杂度与效率。速度代表每秒帧数，用于衡量模型的实时性能；参数量代表模型可训练的总参数的数量，用于衡量模型的复杂度。不同方法的速度与参数量见表5。

**表5　不同网络速度与参数量比较**

**Table 5　Comparison of speed and computational amount of different models**

| 方法 | FPS/（帧/s） | Params/M |
|---|---|---|
| UNet | **41.08** | 17.26 |
| SegNet | 39.10 | 29.44 |
| DeeplabV3 | – | 25.31 |
| PSPNet | 25.45 | 53.51 |
| HR-Net | 30.50 | 29.53 |
| Res-UNet | 20.71 | **13.04** |
| PPANet | – | 26.43 |
| 本文 | 26.22 | 24.23 |

注：加粗字体表示各列最优结果。“–”表示缺少数据。

如表5所示，所提模型GPDEA-UNet的FPS为26.22帧/s，参数量为24.23M，FPS在8种模型中适中，参数量较少。GPDEA-UNet仅采用UNet作为编解码器，并且所设计的模块中都没有引入非常复杂的操作，因此整体参数量并不是非常大，但由于在编码过程多尺度模型和注意力机制的使用，解码全过程中高分辨率特征的引入，模型的计算复杂度与参数量会增加，模型的速度会降低。虽然牺牲了一定的效率，但是在更重要的有效性指标IoU、precision、recall和F1-score上，所提算法均得到了最好的指标，取得了更好的建筑物提取效果。

## 3 结 论

针对遥感图像分割的区域连续性差、边界模糊及尺度变化大导致的建筑物分割精度低的问题，本文提出基于全局感知与细节增强的非对称遥感建筑物分割网络GPDEA-UNet。所提网络设计了基于选择性状态空间的特征编码器模块、多尺度双交叉融合注意力模块以及细节增强解码器模块，有效提升了建筑物特征的提取、融合及细节保留能力，解决了通道与空间依赖性问题，并缩小了编解码器间的语义差距，最终确保了分割结果的精确性与细腻度。

通过在公开数据集WHU Aerial Imagery Dataset和Massachusetts Building Dataset上的定量实验、定性实验与消融实验，并与现有主流方法进行对比，验证所提GPDEA-UNet的有效性和鲁棒性。

尽管所提方法在处理复杂场景和多变尺度时表现出色，但在极端复杂或噪声干扰严重的遥感图像中，其分割性能和泛化能力仍有待提升。因此，未来研究将致力于优化网络结构、探索更高效的训练策略，并考虑在网络中引入更多遥感建筑物先验信息，以进一步提升GPDEA-Unet的分割性能和实用性。

## 参考文献(References)

Aleissaee A A, Kumar A, Anwer R M, Khan S, Cholakkal H, Xia G S and Khan F S. 2023. Transformers in remote sensing: a survey. Remote Sensing, 15(7): #1860 [DOI: 10.3390/rs15071860]

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. 2021a. Building extraction from remote sensing images with sparse token Transformers. Remote Sensing, 13(21): #4441 [DOI: 10.3390/rs13214441]

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/TPAMI.2017.2699184]

Chen M, Wu J J, Liu L Z, Zhao W H, Tian F, Shen Q, Zhao B Y and Du R H. 2021b. Dr-Net: an improved network for building extraction from high resolution remote sensing image. Remote Sensing, 13(2): #294 [DOI: 10.3390/rs13020294]

Diakogiannis F I, Waldner F, Caccetta P and Wu C. 2020. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162: 94-114 [DOI: 10.1016/j.isprsjprs.2020.01.013]

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]

Dong J, Mao J K, Liu K and Cheng L Y. 2024. Remote sensing image building segmentation based on edge attention. Journal of Tianjin University of Technology, (3):92-97(董杰, 毛经坤, 刘坤, 程良勇. 2024. 基于边缘注意的遥感图像建筑物分割. 天津理工大学学报, (3):92-97)

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. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2024-10-23]. https://arxiv.org/pdf/2010.11929v1.pdf

Ghaffarian S, Valente J, Van Der Voort M and Tekinerdogan B. 2021. Effect of attention mechanism in deep learning-based remote sensing image processing: a systematic literature review. Remote Sensing, 13(15): #2965 [DOI: 10.3390/rs13152965]

Gu A and Dao T. 2023. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. [2024-10-23]. https://arxiv.org/pdf/2312.00752.pdf

Gu A, Goel K and Ré C. 2022. Efficiently modeling long sequences with structured state spaces//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Li E, Femiani J, Xu S B, Zhang X P and Wonka P. 2015. Robust rooftop extraction from visible band images using higher order CRF. IEEE Transactions on Geoscience and Remote Sensing, 53(8): 4483-4495 [DOI: 10.1109/TGRS.2015.2400462]

Li H F, Qiu K J, Chen L, Mei X M, Hong L and Tao C. 2021. SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters, 18(5): 905-909 [DOI: 10.1109/LGRS.2020.2988294]

Li Q Y, Mou L C, Sun Y, Hua Y S, Shi Y L and Zhu X X. 2024. A review of building extraction from remote sensing imagery: geometrical structures and semantic attributes. IEEE Transactions on Geoscience and Remote Sensing, 62: #4702315 [DOI: 10.1109/TGRS.2024.3369723]

Lin G S, Milan A, Shen C H and Reid I. 2017. RefineNet: multi-path refinement networks for high-resolution semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5168-5177 [DOI: 10.1109/CVPR.2017.549]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks

for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Qiu Y, Wu F, Yin J C, Liu C Y, Gong X Y and Wang A D. 2022. MSL-Net: an efficient network for building extraction from aerial imagery. Remote Sensing, 14(16): #3914 [DOI: 10.3390/rs14163914]

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Shao Z F, Tang P H, Wang Z Y, Saleem N, Yam M and Sommai C. 2020. BRRNet: a fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. Remote Sensing, 12(6): #1050 [DOI: 10.3390/rs12061050]

Sun K, Zhao Y, Jiang B R, Cheng T H, Xiao B, Liu D, Mu Y D, Wang X G, Liu W Y and Wang J D. 2019. High-resolution representations for labeling pixels and regions [EB/OL]. [2024-10-23]. https://arxiv.org/pdf/1904.04514.pdf

Thottolil R and Kumar U. 2022. Automatic building footprint extraction using random forest algorithm from high resolution Google earth images: a feature-based approach//Proceedings of 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). Bangalore, India: IEEE: 1-6 [DOI: 10.1109/CONECCT55679.2022.9865829]

Wang L B, Li R, Zhang C, Fang S H, Duan C X, Meng X L and Atkinson P M. 2022. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 190: 196-214 [DOI: 10.1016/j.isprsjprs.2022.06.008]

Wang Z and Qu S J. 2024. Research progress and challenges in real-time semantic segmentation for deep learning. Journal of Image and Graphics, 29(5): 1188-1220 (王卓, 瞿绍军. 2024. 深度学习实时语义分割研究进展和挑战. 中国图象图形学报, 29(5): 1188-1220) [DOI: 10.11834/jig.230605]

Xiang W K, Zhou Q, Cui J C, Mo Z Y, Wu X F, Ou W H, Wang J D and Liu W Y. 2024. Weakly supervised semantic segmentation based on deep learning. Journal of Image and Graphics, 29(5): 1146-1168 (项伟康, 周全, 崔景程, 莫智懿, 吴晓富, 欧卫华, 王井东, 刘文予. 2024. 基于深度学习的弱监督语义分割方法综述. 中国图象图形学报, 29(5): 1146-1168) [DOI: 10.11834/jig.230628]

Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #924

Xu S J, Deng B W, Meng Y B, Liu G H and Han J Q. 2022. ReA-Net: a multiscale region attention network with neighborhood consistency supervision for building extraction from remote sensing image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15: 9033-9047 [DOI: 10.1109/JSTARS.2022.3204576]

Xu S J, Du M, Meng Y B, Liu G H, Han J Q and Zhan B H. 2024. MDBES-Net: building extraction from remote sensing images based on multiscale decoupled body and edge supervision network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 519-534 [DOI: 10.1109/jstars.2023.3331444]

Yang J H, Zhang H and Hua H Y. 2023. Parallel path and strong attention mechanism for building segmentation in remote sensing images. Optics and Precision Engineering, 31(2): 234-245 (杨坚华, 张浩, 花海洋. 2023. 并行路径与强注意力机制遥感图像建筑物分割. 光学精密工程, 31(2): 234-245) [DOI: 10.37188/OPE.20233102.0234]

Yang X, Li S S, Chen Z C, Chanussot J, Jia X P, Zhang B, Li B P and Chen P. 2021. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 177: 238-262 [DOI: 10.1016/j.isprsjprs.2021.05.004]

Zeng X H, Chen I and Liu P. 2021. Improve semantic segmentation of remote sensing images with K-mean pixel clustering: a semantic segmentation post-processing method based on K-means clustering// Proceedings of 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE). SC, USA: IEEE: 231-235 [DOI: 10.1109/CSAIEE54046.2021.9543336]

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

Zheng S X, Lu J C, Zhao H S, Zhu X T, Luo Z K, Wang Y B, Fu Y W, Feng J F, Xiang T, Torr P H S and Zhang L. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6877-6886 [DOI: 10.1109/CVPR46437.2021.00681]

Zhu L H, Liao B C, Zhang Q, Wang X L, Liu W Y and Wang X G. 2024. Vision Mamba: efficient visual representation learning with bidirectional state space model//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: OpenReview.net

Zhu Z M, Li S D, Zheng D B and Xue P S. 2024. RH-CUnet: extracting traditional village buildings by embedding edges and corners. Remote Sensing Information, 39(4): 166-173 (朱梓萌, 李少丹, 郑东博, 薛彭帅. 2024. RH-CUnet: 嵌入边缘和角点的传统村落建筑物提取. 遥感信息, 39(4): 166-173) [DOI: 10.20091/j.

cnki.1000-3177.2024.04.018］

## 作者简介

徐胜军，男，教授，主要研究方向为图像处理、模式识别、人工智能与智动化系统。E-mail：duplin@sina.com

刘雨芮，通信作者，女，硕士研究生，主要研究方向为计算机视觉和遥感影像解译。E-mail：liu18991358966@163.com

刘二虎，男，副教授，主要研究方向为机器学习和计算机视觉。E-mail：liuerhu@xauat.edu.cn

刘俊，男，教授，主要研究方向为机器学习、人工智能在电力系统中的应用。E-mail：eeliujun@mail.xjtu.edu.cn

史亚，女，副教授，主要研究方向为机器学习。E-mail：shiyaworld@163.com

李小晗，女，讲师，主要研究方向为智能机器人控制、3D场景理解和人工智能。E-mail：lixiaohan1993@stu.xjtu.edu.cn