研究论文|Research Article
时空智能|Spatio-Temporal Intelligence

Open Access

引用格式:张建昌, 徐伟铭, 余大文, 等. 联合全局建模与局部补偿的高分辨率遥感影像语义分割 [J]. 海南大学学报(自然科学版中英文), 2026, 44(4): 417−429.

Citation: Zhang Jianchang, Xu Weiming, Yu Dawen, et al. Semantic segmentation for remote sensing images based on joint global feature modeling and local feature compensation[J]. Natural Science of Hainan University, 2026, 44(4): 417−429.

联合全局建模与局部补偿的高分辨率遥感影像语义分割

张建昌[1,2,3], 徐伟铭[1,2,3]✉, 余大文[1,2,3], 梁昌钰[1,2,3], 陈　枭[1,2,3]

(1. 福州大学 数字中国研究院(福建), 福建 福州 350108; 2. 空间数据挖掘与信息共享教育部重点实验室, 福建 福州 350108; 3. 地理空间信息技术国家地方联合工程技术研究中心, 福建 福州, 350108)

摘　要: 为解决高分辨率遥感影像语义分割中普遍存在的局部细节刻画不足以及全局建模效率较低的问题, 本文提出了一种结合卷积神经网络与状态空间模型的分割网络模型—残差曼巴 U 型网络模型。该模型采用轻量级的 ResNet18 作为编码器, 并引入视觉状态空间模块构建解码器, 实现高效的全局上下文建模。同时, 为增强对细粒度语义信息的感知能力, 设计了局部特征补偿模块。针对深浅层特征融合过程中易出现的语义偏差问题, 进一步提出跨层级融合注意力模块, 以实现空间与语义信息的有效协同。结果表明, 该模型在 Vaihingen、Potsdam 和 LoveDA 三大遥感数据集上的平均交并比分别达到 83.78%、87.09% 和 52.85%, 均优于现有的主流模型。该模型在维持较低计算复杂度的同时, 显著提升了模型的特征表达能力与分割精度。研究结果为高分辨率遥感影像的高效智能解译提供了一种兼顾性能与效率的可行方案。

关键词: 高分辨率遥感影像; 语义分割; Mamba; 卷积神经网络

中图分类号: TP751　　**文献标志码:** A　　**文章编号:** 1004-1729(2026)04-0417-13

DOI: 10.65658/j.hndk.2025101501　　CSTR: 32403.14.hndk.2025101501

Semantic segmentation for remote sensing images based on joint global feature modeling and local feature compensation

Zhang Jianchang[1,2,3], Xu Weiming[1,2,3]✉, Yu Dawen[1,2,3], Liang Changyu[1,2,3], Chen Xiao[1,2,3]

(1. The Academy of Digital China, Fuzhou University, Fuzhou 350108, China; 2. Key Lab of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou 350108, China; 3. National Engineering Research Center of Geospatial Information Technology, Fuzhou 350108, China)

Abstract: To address the common challenges of insufficient local detail representation and low global modeling efficiency in high-resolution remote sensing image semantic segmentation, this paper proposes a segmentation network that integrates convolutional neural networks and state-space models, named RMUNet. The proposed model employs a lightweight ResNet18 as the encoder and introduces a visual state space block (VSSBlock) in the decoder to achieve efficient global context modeling. Meanwhile, a local feature compensation module (LFCM) is designed to enhance the perception of fine-grained semantic information. To mitigate the semantic bias that may arise during the fusion of shallow and deep features, a cross-level fusion

Received: 2025−10−15; **Revised:** 2026−02−03; **Accepted:** 2026−02−10.
✉ **Corresponding author(s):** Xu Weiming, E-mail: xwming2@126.com..

attention module (CFAM) is further proposed to enable effective collaboration between spatial and semantic representations. Experimental results demonstrate that RMUNet achieves mean intersection-over-union (mIoU) scores of 83.78%, 87.09%, and 52.85% on the Vaihingen, Potsdam, and LoveDA datasets, respectively, outperforming existing mainstream methods. While maintaining low computational complexity, RMUNet significantly enhances feature representation and segmentation accuracy, providing an efficient and effective solution for high-resolution remote sensing image interpretation.

Keywords: high-resolution remote sensing images; semantic segmentation; Mamba; convolutional neural networks

0 引 言

语义分割是遥感应用的核心任务之一,其目标是依据既定的类别划分,对遥感影像实施逐像素的精确分类。高分辨率遥感影像在国土规划[1]、灾害监测[2]和环境保护[3]等领域得以广泛应用。相较于中低分辨率影像,高分辨率影像在提供丰富地物细节的同时,也增加了分析的难度,对语义分割方法提出了更高要求[4]。近年来,深度学习技术凭借其出色的特征自动提取能力以及端到端学习框架,在计算机视觉领域得到广泛应用[5]。目前,基于深度学习的遥感影像语义分割方法主要分为卷积神经网络(convolutional neural networks,CNN)和 Transformer[6] 两大类。基于 CNN 的方法通常通过堆叠卷积层逐步提取影像特征,卷积操作能够有效保持像素间的局部拓扑结构,有助于捕捉空间细节信息。典型研究有 U-Net、DeepLab 和 PSPNet 等[7−9],这些方法在多尺度特征融合和空间细节保持方面取得了显著成效。然而,CNN 固定的感受野限制了其对长程依赖和复杂背景的建模能力,难以充分获取全局语义上下文。为克服 CNN 在全局建模方面的不足,Transformer 架构被引入语义分割任务。其自注意力机制能够在全局范围内捕获像素间的依赖关系,在遥感场景中展现出较强的全局感知能力。越来越多的研究将 Transformer 应用于遥感影像的编码器—解码器结构,并取得显著进展。典型的纯 Transformer 结构包括 Segmenter[10]、SegFormer[11]、Swin-unet[12] 和 Swin-UperNet[13] 等。而混合结构则将 Transformer 与 CNN 相结合,如 TransUNet[14]、DC-Swin[15]、BANet[16] 和 UnetFormer[17],这些研究有效整合了局部与全局特征提取的优势,在复杂地物识别和多尺度场景理解方面表现出良好性能。然而,Transformer 的自注意力机制具有平方级的计算复杂度,模型参数量庞大[18],难以在高分辨率遥感影像的实际应用中实现高效推理。此外,其对细节特征的表达能力仍存在局限,在小目标分割方面存在不足。

Gu 等[19] 提出了一种基于状态空间模型(state space models,SSMs)[20] 的 Mamba 模块,该模块用于在序列维度上递归更新状态,以实现长程信息的传播。Liu 等[21] 提出的 VMamba 通过优化 Mamba 的单向扫描机制,并引入视觉专用的状态更新策略,成功将 Mamba 架构拓展至计算机视觉领域。在遥感影像语义分割任务中,Mamba 结构同样受到了广泛关注。相较于自然场景影像,高分辨率遥感影像具有更为复杂的空间纹理、更为显著的尺度差异以及更为稠密的地物分布,经典卷积或自注意力机制在长程依赖建模方面往往存在效率上的瓶颈。基于 SSMs 的 Mamba 模块在此场景下展现出独特优势,其递归状态更新能够在维持线性复杂度的同时,有效整合跨区域、跨尺度的全局语义信息,为高分辨率遥感影像中的复杂地物模式建模提供了新的解决思路。相关研究如 Rs-mamba[22]、Samba[23]、UNetMamba[24]、RS^{3}Mamba[25]、PyramidMamba[26] 和 CM-Unet[27] 等,将基于 SSMs 的 Mamba 模块引入语义分割网络,在保持较低计算复杂度的同时,通过有效建模远程依赖关系,显著提高了模型的分割精度。

然而,这些模型主要侧重于 Mamba 高效的全局特征提取能力,对细粒度局部信息的捕捉和多层级特征的协同建模关注不足。在高分辨率遥感影像中,若忽视局部信息,模型虽可获得较好的宏观语义一致性,但在小目标和复杂纹理区域往往会出现性能退化。此外,高分辨率遥感影像中浅层特征与深层特征在空间分辨率和语义抽象程度方面存在显著差异,若缺乏有效的跨层引导机制,直接进行融合易导致语

义不对齐和特征冗余，进而限制分割性能的进一步提高。因此，提出一种残差曼巴 U 型网络（residual-mamba-u-net, RMUNet）模型，借助 Mamba 线性复杂度的全局建模优势，同时通过专门设计的局部特征补偿模块（local feature compensation module, LFCM）与跨层级融合注意力模块（cross-layer fusion attention module, CFAM），实现深层与浅层特征的有效融合以及全局与局部信息的协同表达。

1 研究方法

1.1 RMUNet 模型网络结构 Mamba 架构在保持线性计算复杂度的同时，可有效对全局语义信息进行建模，是 Transformer 的一种可行替代方案；而 CNN 能够整合相邻像素的排列关系，进而保留空间局部拓扑特征，具备卓越的局部特征提取能力。RMUNet 模型充分融合了 CNN 和 Mamba 的优势，采用经典的编码器–解码器架构。编码器以轻量的 ResNet18 作为骨干网络，由多个残差块（residual block, ResBlock）堆叠而成，高效提取多尺度语义特征；解码器则由 VMamba 的基本单元视觉状态空间模块（visual state space block, VSSBlock）构建，捕获全局上下文信息。此外，RMUNet 模型还集成了用于增强局部语义信息感知的 LFCM 和用于有效融合深浅层特征的 CFAM，如图 1 所示。

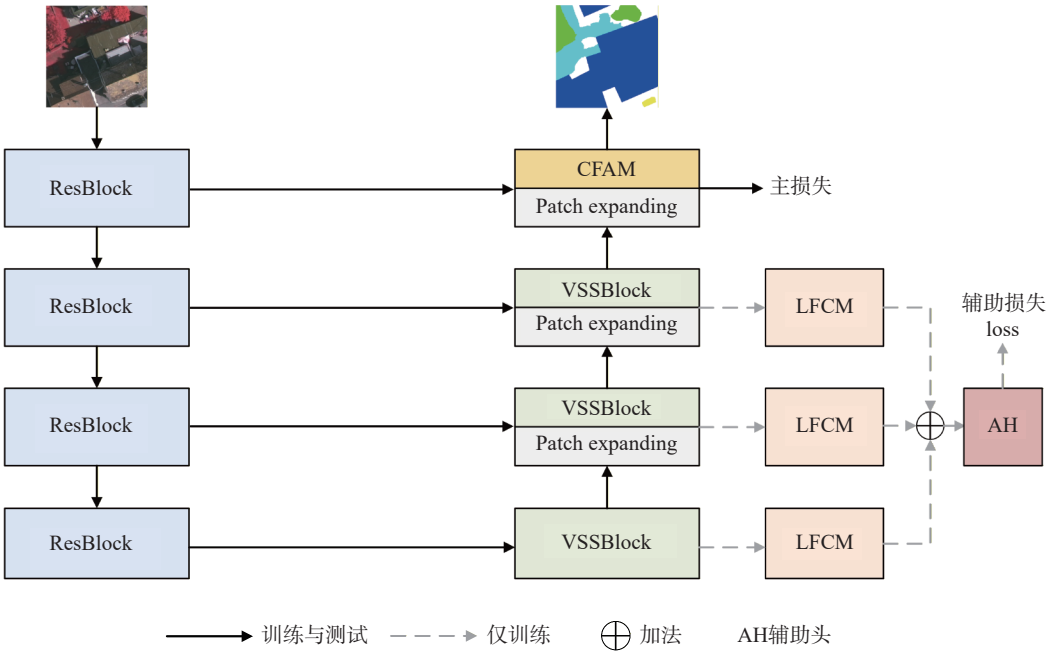

图 1 RMUNet 模型总体网络框架

1.2 视觉状态空间模块 在高分辨率遥感影像里，部分地物类别之间存在明显的光谱和形态相似性，若仅依赖局部特征进行判别，极易引发类别混淆。故而，引入全局语义信息对于达成精准分割而言至关重要。经典 CNN 因感受野有限，难以捕捉长距离像素间的依赖关系；而 Transformer 虽具备卓越的全局建模能力，在众多视觉任务中表现优异，但其自注意力机制所带来的平方级计算复杂度，限制了其在高分辨率遥感影像中的效率和扩展性。因此，如何在保留 Transformer 全局建模优势的同时降低计算成本，成为亟待攻克的研究难题。

鉴于高分辨率遥感影像展平后可被视作超长序列，相较于 Transformer 的二次方复杂度，Mamba 模型的线性计算复杂度在处理此类数据时具备天然优势。然而，在编码阶段采用预训练的 Mamba 模型作为主干网络，常常会致使模型的参数量过于庞大，难以确保效率。因此，将 VMamba 的基础单元 VSSBlock 引入解码器部分，以高效建模全局信息，从而提高复杂遥感场景的语义分割精度。

VSSBlock 的具体结构如图 2 所示。输入特征图 $X_{\text{in}} \in \mathbf{R}^{H\times W\times C}$（$H$ 表示图像在垂直方向的像素数量，即高度；W 表示图像在水平方向的像素数量，即宽度；C 表示每个像素包含的颜色或特征维度数量，即通道

数)经层归一化后,并行进入两个分支,第一分支借助线性层将通道数量扩展至λC(λ为预设的通道扩展因子),随后依次通过深度可分离卷积、SiLU激活函数和SS2D模块提取特征,再经层归一化进行标准化处理;第二分支则通过线性层与SiLU激活函数完成特征变换。两分支输出的特征通过哈达玛积进行融合,之后使用线性投影将通道数恢复至C,最终生成与输入维度相同的输出$X_{\text{out}}=\mathbf{R}^{H\times W\times C}$,该过程可表示为

$$X_1=\text{LN}(\text{SS2D}(\text{SiLU}(\text{DWConv}(\text{Linear}(X_{\text{in}})))))\tag{1}$$

$$X_2=\text{SiLU}(\text{Linear}(X_{\text{in}}))\tag{2}$$

$$X_{\text{out}}=\text{Linear}(X_1\odot X_2)\tag{3}$$

式中,X_1为第一分支输出,LN(·)为层归一化层,SS2D(·)为对应二维选择性扫描模块,SiLU(·)为SiLU激活函数,DWConv(·)为深度可分离卷积层,Linear(·)为线性层,X_2为第二分支输出,$\odot$为哈达玛积。

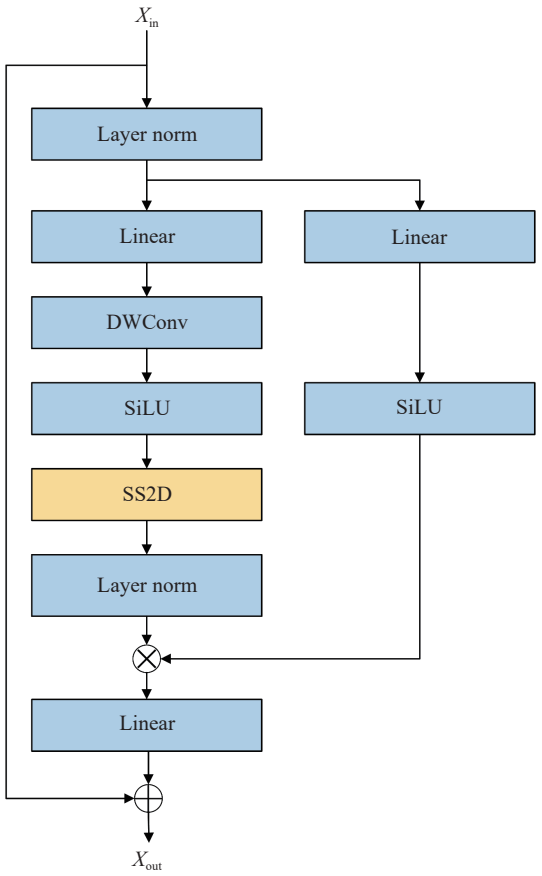

图2 视觉状态空间模块

1.3 局部特征补偿模块 尽管全局上下文信息在复杂城市场景的语义分割中具有重要作用,高分辨率影像里丰富的局部细节仍是提升分割精度的关键要素。将VSSBlock引入解码器有利于高效构建全局上下文模型,但在解码进程中,也可能造成局部语义信息出现部分缺失,进而对最终的分割效果产生影响。鉴于此,提出一种基于CNN的LFCM,旨在增强模型对局部细节的感知能力,如图3所示。值得强调的是,LFCM仅在训练阶段启用,因此不会增加推理成本,有助于提高模型的应用效率。针对高分辨率场景中尺度较小的地物目标,LFCM采用双分支并行卷积结构,在解码过程中动态补偿因上下文建模可能丢失的细节信息。每个分支之后均连接有批归一化(batch normalization, BN)层和ReLU激活函数,以增强特征表达的稳定性和非线性建模能力。

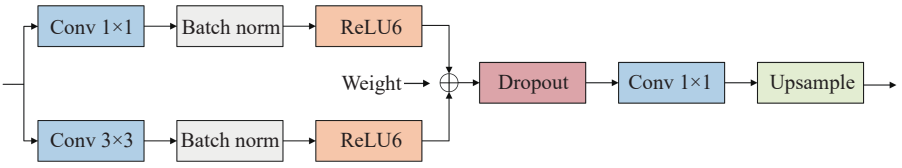

图3 局部特征补偿模块

$$X_1=\text{ReLU6}(\text{BN}(\text{Conv}_{1\times1}(X_{\text{in}})))\tag{4}$$

$$X_2=\text{ReLU6}(\text{BN}(\text{Conv}_{3\times3}(X_{\text{in}})))\tag{5}$$

$$X_{\text{out}}=\text{Upsample}(\text{Conv}_{1\times1}(\text{Dropout}(\alpha X_1+(1-\alpha)X_2)))\tag{6}$$

式中,X_1为第一分支输出,ReLU6(·)为ReLU6激活函数,BN(·)为批归一化层,Conv(·)为卷积层,X_{in}为输入特征,X_2为第二分支输出,X_{out}为输出特征,Upsample(·)为上采样层,Dropout(·)为Dropout层,α为可学习的权重参数。

1.4 跨层级融合注意力模块 首个ResBlock生成的浅层特征,虽包含丰富的局部细节信息,但语义表达能力较弱;而VSSBlock提取的深层特征具备较强的全局语义感知能力,不过其空间分辨率较低。针对直

接融合两类特征可能导致的语义偏差与精度降低问题，提出 CFAM 通过跨层级特征解耦，保留两类特征的互补优势，再利用跨维度信息交互，构建空间细节与类别语义的精确对应关系，从而提升模型的分割能力。

CFAM 的具体结构如图 4 所示。首先对浅层特征与深层特征实施加权融合，以整合空间细节与语义信息，并将融合结果作为 CFAM 的输入。具体来讲，针对输入特征图，CFAM 运用两个并行分支进行处理，分别关注通道依赖与空间上下文信息。其中一条分支沿水平方向和垂直方向分别开展自适应平均池化操作，从而有效捕获不同空间维度上的通道响应特征；随后将两个方向的池化结果在空间维度上进行拼接，并借助 1×1 卷积进行融合，进而得到用于引导通道注意力分布的空间引导特征。另一条分支则采用 3×3 卷积对原始特征进行局部建模，以增强空间结构感知能力。两个分支的输出分别经过全局平均池化与 Softmax 激活后，通过点积的方式进行交叉融合，形成注意力图。最终，利用 Sigmoid 函数对注意力图进行归一化处理，并将其作为权重与原始特征图相乘，以实现空间层面的特征增强。

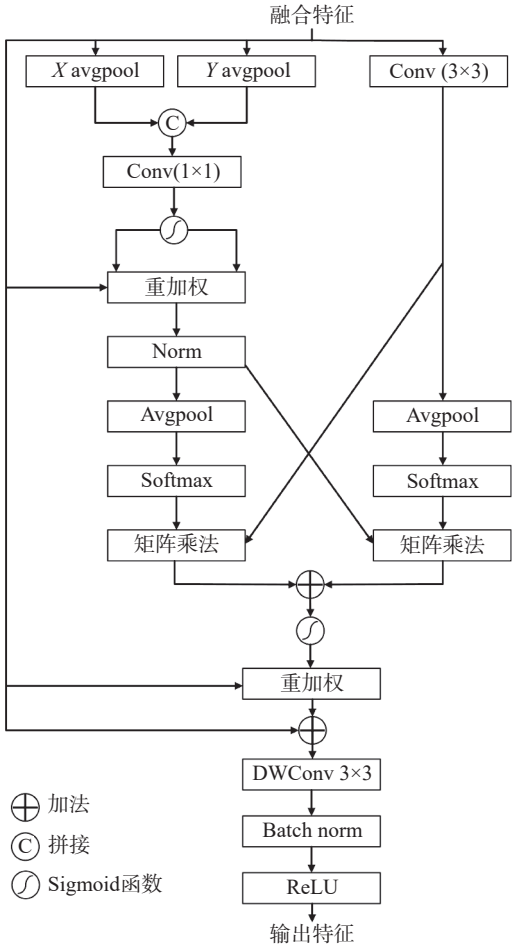

图 4　跨层级融合注意力模块

1.5　损失函数　在 RMUNet 模型的训练过程中，采用的损失函数由两部分组成，分别为用于整体优化的主损失L_{p}和用于局部监督的辅助损失L_{aux}。

为全面提升 RMUNet 模型的分割性能，选用两种经典的损失函数直接相加，以此构建主损失函数L_{p}。

$$L_{\mathrm{dice}}=1-\frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\hat{y}_k^{(n)}y_k^{(n)}}{\hat{y}_k^{(n)}+y_k^{(n)}} \tag{7}$$

$$L_{\mathrm{ce}}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\hat{y}_k^{(n)}\lg y_k^{(n)} \tag{8}$$

式中，L_{dice}为 Dice 损失[28]，N为样本总数，K为类别总数，$\hat{y}_k^{(n)}$为第 n 个样本的真实标签的独热编码（one-hot encoding）中第 k 个元素，$y_k^{(n)}$为模型预测第 n 个样本属于第 k 类别的置信度；L_{ce}为交叉熵损失。

此外，为有效实现 LFCM 的细节监督作用，额外引入了交叉熵损失函数L_{ce}作为辅助损失L_{aux}，并通过加权平衡主损失L_{p}与辅助损失L_{aux}的作用。RMUNet 模型的总体损失函数定义为

$$L=L_{\mathrm{p}}+\alpha L_{\mathrm{aux}}=(L_{\mathrm{ce}}+L_{\mathrm{dice}})+\alpha L_{\mathrm{ce}} \tag{9}$$

式中，L为总体损失函数；α为权重，设为 0.4[17, 29]，该取值可在保证核心任务性能的同时，充分发挥辅助损失的监督补偿作用，避免单一损失占比过高导致模型训练偏置。

2　实验与分析

2.1　数据集　（1）Vaihingen 数据集由 33 幅高分辨率正射影像构成，影像平均大小为 2 494 像素×2 064 像

素。每幅影像包含3个多光谱波段(近红外、红、绿波段),地面采样距离为9 cm。数据内容涵盖5种前景类别(不透水面、建筑物、低矮植被、树木及汽车)和一种背景类别(混合区域)。实验中将ID为2、4、6、8、10、12、14、16、20、22、24、27、29、31、33、35和38的影像设为测试集,其余16幅影像作为训练集。

(2)Potsdam数据集由38幅高分辨率正射影像组成,每幅影像的尺寸为6 000像素×6 000像素,地面分辨率为5 cm。其类别设置与Vaihingen数据集相同,影像包含红、绿、蓝和近红外4个多光谱通道。实验选取ID为2_13、2_14、3_13、3_14、4_13、4_14、4_15、5_13、5_14、5_15、6_13、6_14、6_15和7_13的影像作为测试集(共14幅),其余24幅影像作为训练集。

(3)LoveDA数据集包含5 987幅高分辨率光学遥感影像,地面采样距离为0.3 m,影像尺寸为1 024像素×1 024像素,涵盖建筑、道路、水体、裸地、森林、农田及背景类共7类土地覆盖类型。该数据采集于中国南京、常州和武汉3座城市,覆盖城市与农村两种典型场景,因存在多尺度物体、复杂背景和不一致的类别分布,具有较高的挑战性。实验遵循官方划分标准:2 522幅用于训练,1 669幅用于验证,1 796幅用于测试。

2.2 实验环境与参数 为保障实验条件的一致性,所有模型均依托PyTorch框架在单个NVIDIA GTX 3090 GPU上实现。采用AdamW优化器以加速收敛进程,基础学习率设定为0.000 6,并结合余弦退火策略实施动态调整。训练过程中,将输入影像随机裁剪为512像素×512像素的图块,同时采用多尺度随机缩放(缩放因子为0.50、0.75、1.00、1.25、1.50)垂直/水平翻转以及旋转等数据增强策略。训练周期设定为100个epoch,批量大小设为8。测试阶段采用测试时增强策略(test-time augmentation, TTA),包括垂直翻转和水平翻转等操作。

2.3 评价指标 为验证RMUNet模型的有效性,选取在遥感影像语义分割领域被广泛应用的交并比(intersection over union, IoU)、平均F_1分数(mean F_1, mF_1)、总体准确率(overall accuracy, OA)和平均交并比(mean intersection over union, mIoU)作为评价指标,其表达式分别为

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{11}$$

$$IoU = \frac{TP}{TP+FP+FN} \times 100\% \tag{12}$$

$$mF_1 = \frac{1}{K}\sum_{i=1}^{K} \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{14}$$

$$mIoU = \frac{1}{K}\sum_{i=1}^{K} \frac{TP}{TP+FP+FN} \tag{15}$$

式中,*TP*、*FP*、*TN*和*FN*分别表示实际为正类且被正确预测为正类的样本数、实际为负类但被错误预测为正类的样本数、实际为负类且被正确预测为负类的样本数和实际为正类但被错误预测为负类的样本数,*Precision*为精确率,*Recall*为召回率。

2.4 对比实验 为评估RMUNet模型的性能与泛化能力,在3个遥感数据集上,将其与当前主流的语义分割模型进行对比分析,这些模型包括DeepLabV3+[8]、BANet[16]、Segmenter[10]、UNetformer[17]、RS^{3}Mamba[25]和CM-UNet[27]。

2.4.1 Vaihingen数据集上的对比分析 RMUNet模型在5个类别上均展现出卓越表现,显著优于其他模型,如表1所示。其中,在汽车类别上,RMUNet的*IoU*达到82.10%,相较于次优模型提升了3.12百分点,验证了其在小目标识别任务中的鲁棒性。此外,在mF_1、*OA*和*mIoU*这3项综合指标上,RMUNet模

型分别达到 91.05%、91.48% 和 83.78%，均为所有模型中最高值，充分体现了该模型在整体语义理解和细节还原方面的强大能力。

表 1　Vaihingen 数据集上的性能比较

类别	*IoU*/%					mF_1/%	*OA*/%	*mIoU*/%
	不透水面	建筑物	低矮植被	树木	汽车			
DeepLabV3+	85.08	89.75	71.77	80.88	72.21	88.68	90.01	79.94
BANet	85.53	90.88	72.14	81.69	76.71	89.61	90.52	81.39
Segmenter	81.49	86.92	68.35	80.02	51.06	84.13	87.42	73.57
UNetformer	86.95	91.34	73.80	82.09	78.60	90.32	91.07	82.56
RS^3Mamba	87.54	92.27	73.87	82.17	78.98	90.55	91.33	82.96
CM-UNet	86.85	91.37	74.09	82.47	78.89	90.43	91.16	82.73
RMUNet	87.72	92.45	74.13	82.50	82.10	91.05	91.48	83.78

RMUNet 模型在 Vaihingen 数据集上若干预测结果的可视化如图 5 所示。图中，Ground truth（真值标签）是人工精细标注的真实地物边界。在屋顶和街区交界处等复杂结构区域，其分割结果更贴合实际情况，边缘更为清晰明确。与其他模型在细节处存在的模糊、粘连和漏分等问题相比，RMUNet 模型展现出良好的空间结构保持能力。

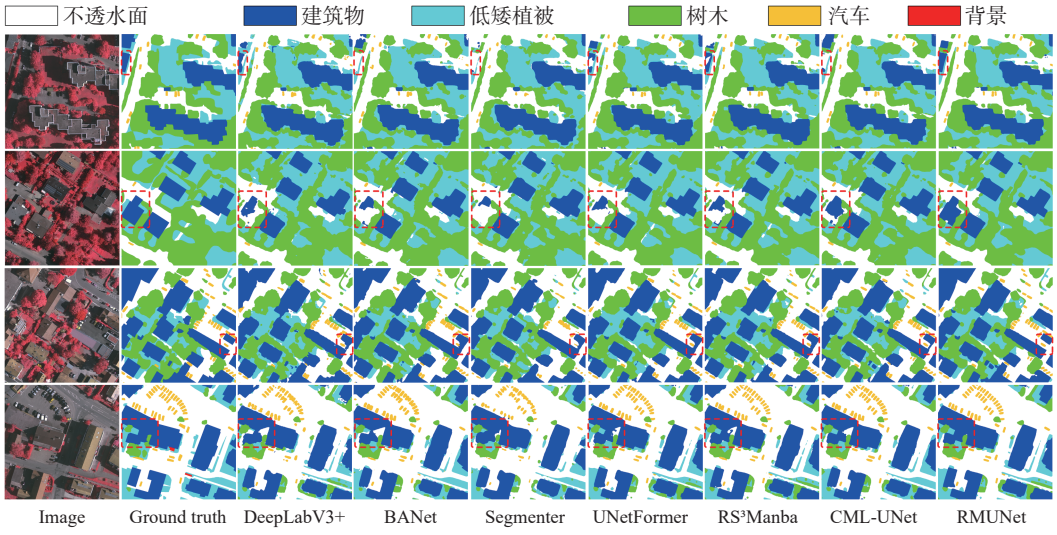

图 5　Vaihingen 数据集的可视化结果

2.4.2　Potsdam 数据集上的对比分析　与 Vaihingen 相比，Potsdam 数据集更为复杂，具有更大的影像尺寸和更丰富的场景变化，如表 2 所示。RMUNet 模型在多个类别上展现出显著的性能优势，例如在不透水面、建筑物、低矮植被和汽车 4 个类别上，*IoU* 分别达到 88.38%、94.38%、78.76% 和 93.67%。在综合评价指标方面，RMUNet 模型的 mF_1、*OA* 和 *mIoU* 分别达到 92.97%、91.57% 和 87.09%，均优于现有的主流模型。尽管 RMUNet 模型在树木类别上的表现略低于 BANet，但整体精度依然保持稳定，充分展现了该模型在复杂地物分割任务中的适应性和泛化能力。

在面对更高的分辨率、更复杂的街道布局和密集的车辆目标时，RMUNet 模型仍展现出良好的分割稳定性，如图 6 所示。特别是在建筑物区域，RMUNet 模型能够有效区分不同类别，规避了诸多模型中常见的将建筑物误分为低矮植被的问题，彰显了该模型的精细化表达能力。

2.4.3　LoveDA 数据集上的对比分析　LoveDA 数据集涵盖城乡混合区域，其类别分布呈现高度不均衡性，且场景差异性显著，从而对模型的跨场景泛化能力提出了更高要求。尽管 RMUNet 模型在建筑物、水体和耕地 3 个类别上取得了较高的交并比，但在背景、道路、裸土和林地类别上的性能略低于部分模型，

如表3所示。

表2 Potsdam数据集上的性能比较

类别	*IoU*/%					mF_1/%	*OA*/%	*mIoU*/%
	不透水面	建筑物	低矮植被	树木	汽车			
DeepLabV3+	85.26	91.55	75.55	77.49	90.18	91.17	89.73	84.00
BANet	87.12	93.35	77.86	80.30	92.72	92.51	91.02	86.27
Segmenter	84.50	91.11	74.59	73.99	79.50	89.23	88.74	80.74
UNetformer	87.25	93.73	77.77	79.45	92.64	92.44	90.98	86.17
RS^{3}Mamba	87.23	94.05	78.09	79.73	93.20	92.60	91.08	86.46
CM-UNet	87.55	93.78	77.70	79.55	92.72	92.49	91.01	86.26
RMUNet	88.38	94.38	78.76	80.25	93.67	92.97	91.57	87.09

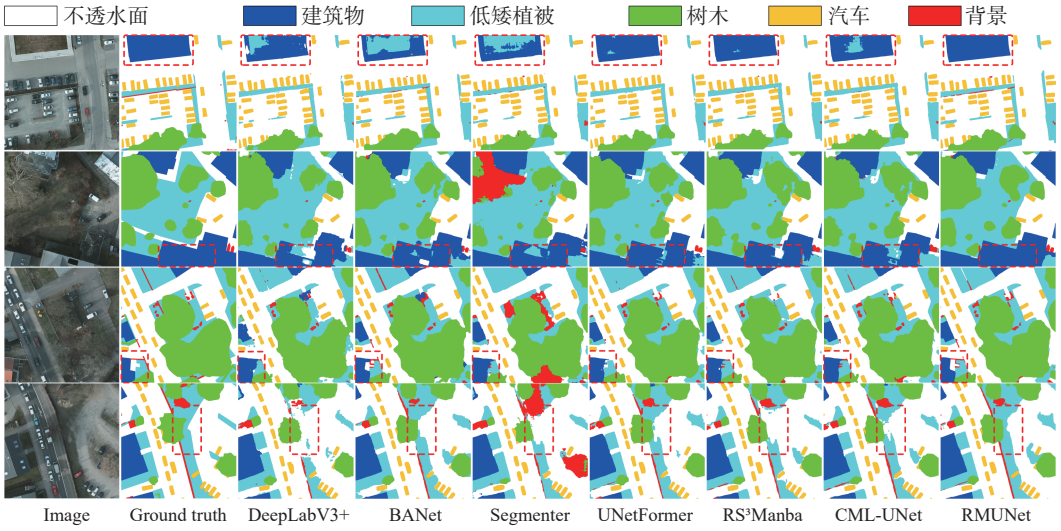

图6 Potsdam数据集的可视化结果

表3 LoveDA数据集上的性能比较

类别	*IoU*/%							*mIoU*/%
	背景	建筑物	道路	水体	裸土	林地	耕地	
DeepLabV3+	42.15	55.42	52.39	76.01	16.89	44.41	62.45	49.96
BANet	46.29	58.43	51.75	80.60	16.51	45.25	61.69	51.50
Segmenter	38.02	50.73	48.73	77.44	13.31	43.52	58.23	47.14
UNetformer	44.68	57.97	54.46	79.04	19.56	47.45	61.64	52.11
RS^{3}Mamba	44.89	58.63	55.04	80.19	18.43	47.12	62.59	52.42
CM-UNet	45.41	57.66	56.48	80.55	17.95	46.10	61.89	52.29
RMUNet	45.93	58.65	55.54	81.30	17.92	46.24	64.34	52.85

对上述弱势类别的进一步分析表明,这些类别普遍表现出边界模糊性、纹理高度相似性以及跨城郊场景的显著差异性等特征,易于在高分辨率遥感影像中与相邻类别发生混淆。其中,道路与背景在城郊区域常呈现细长形态和破碎化分布;裸土与背景在光谱和纹理特征上高度相似;而林地在城乡过渡区域则具有复杂的内部结构和显著的尺度变化。这些因素均对模型在细节表征和区域一致性建模方面构成挑战。

此外,LoveDA数据集中城乡场景的类别分布呈现显著偏差,例如背景与裸土在城郊区域表现出大尺度、非均匀分布,这在一定程度上提高了模型对少样本或高变异类别的学习难度。该现象表明,尽管RMUNet模型在多尺度特征融合与语义建模方面具有优势,但在处理类别不平衡和复杂地表背景时仍存

在局限性，尤其在对弱势类别进行精细区分的能力方面，尚有提升空间。

然而，从整体性能来看，RMUNet 模型依然优于其他模型，显示出良好的综合泛化能力。尤其在耕地类别上，其 *IoU* 达到了 64.34%，显著优于其他对比模型，表明该模型在处理尺度差异显著的目标时，具备卓越的建模能力。

在城乡接合部的复杂地形条件下，RMUNet 模型同样展现出较强的适应性，如图 7 所示。在耕地、水体、林地等多种地物类型交错分布的场景中，RMUNet 模型能够稳定地识别并区分各类别边界，表现出优异的类别识别能力和结构保持性能。其分割结果与真实场景分布更接近，整体影像呈现出更高的一致性。

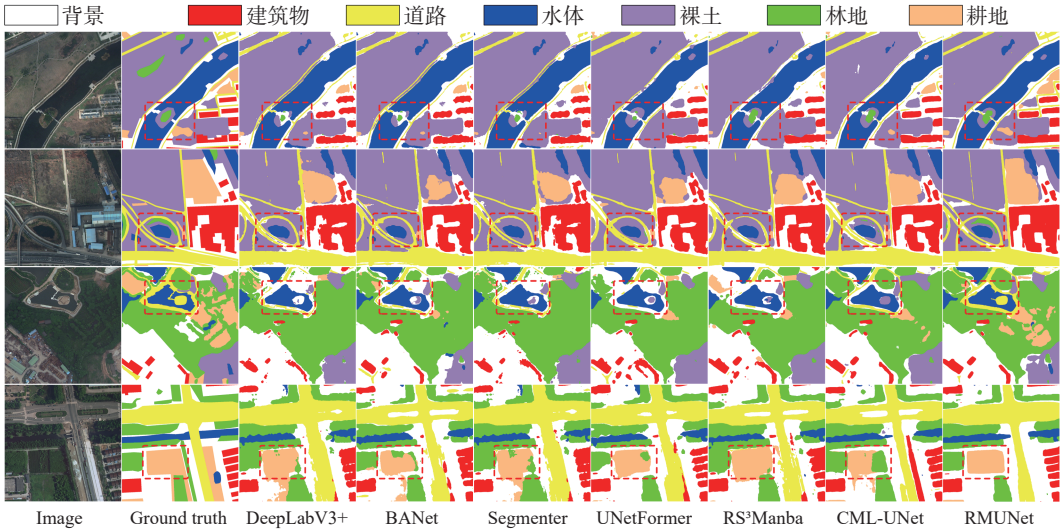

图 7　LoveDA 数据集的可视化结果

2.4.4　计算复杂性分析　为全面评估模型的计算效率与部署友好性，对比不同语义分割方法的模型参数量（Parameters）与计算量（FLOPs），结果如表 4 所示。FLOPs 指的是每次前向传播所需的浮点运算次数，能够较好地衡量模型在推理阶段的计算开销。

表 4　参数量和计算量对比

分割方法	参数量/×10⁶个	计算量/×10⁹次
DeepLabV3+	5.81	105.80
BANet	12.73	52.71
Segmenter	6.70	26.80
UNetformer	11.72	46.97
RS³Mamba	43.32	253.20
CM-UNet	13.11	48.04
RMUNet	28.91	90.01

RMUNet 模型的计算复杂度主要源于编码器和由 VSSBlock、LFCM 和 CFAM 构成的解码器。编码器采用轻量级 ResNet18 进行特征提取，其计算成本较低。解码器中的 VSSBlock 通过线性递归形式实现序列信息的动态传播，相较于自注意力机制的二次复杂度，该模块的复杂度降至线性级别，从而显著降低了计算负担。LFCM 采用局部卷积运算执行细粒度特征补偿，其计算量与卷积核尺寸的平方及通道数成正比。CFAM 引入跨层级特征交互机制，并采用轻量化注意力策略，以确保额外开销可控。整体而言，RMUNet 模型在保持线性全局建模效率的同时，仅引入有限的局部计算成本。

从表中可以看出，虽然 RMUNet 模型的参数量和计算量相较于一些轻量级模型有所增加，但其计算复杂度仍显著低于 RS³Mamba。这表明 RMUNet 模型在实现性能提升的同时，维持相对合理的计算开销，从而平衡模型的准确性与其在实际部署中的可行性。在计算量仅轻微增加的基础上，该模型显著增强了

特征提取与融合能力,进而说明其在计算效率与表达能力之间取得良好的平衡。

2.5 消融实验 为评估各模块的实际效用与协同效应,在 Vaihingen、Potsdam 及 LoveDA3 个遥感语义分割数据集上开展系统性的消融实验。以 U-Net 为基础网络,依次引入 VSSBlock、LFCM 和 CFAM,采用渐进叠加策略,以剖析不同模块对整体性能的贡献及其互补关系。RMUNet 在 3 个数据集上的消融实验结果如表 5 所示。Vaihingen 数据集上的消融实验可视化结果如图 8 所示,Baseline(基准网络)指未引入 VSSBlock、LFCM 和 CFAM 的原始 U-Net。

表 5 RMUNet 模型的消融实验结果

数据集	VSSBlock	LFCM	CFAM	mF_1/%	*mIoU*/%
Vaihingen	×	×	×	89.66	81.49
	√	×	×	90.43	82.73
	√	√	×	90.72	83.23
	√	√	√	91.05	83.78
Potsdam	×	×	×	92.16	85.73
	√	×	×	92.62	86.50
	√	√	×	92.80	86.79
	√	√	√	92.97	87.09
LoveDA	×	×	×	—	51.37
	√	×	×	—	52.12
	√	√	×	—	52.54
	√	√	√	—	52.85

注:表中“√”和“×”分别表示该模块是否被纳入网络,即“√”表示该模块被启用,“×”表示该模块被移除。

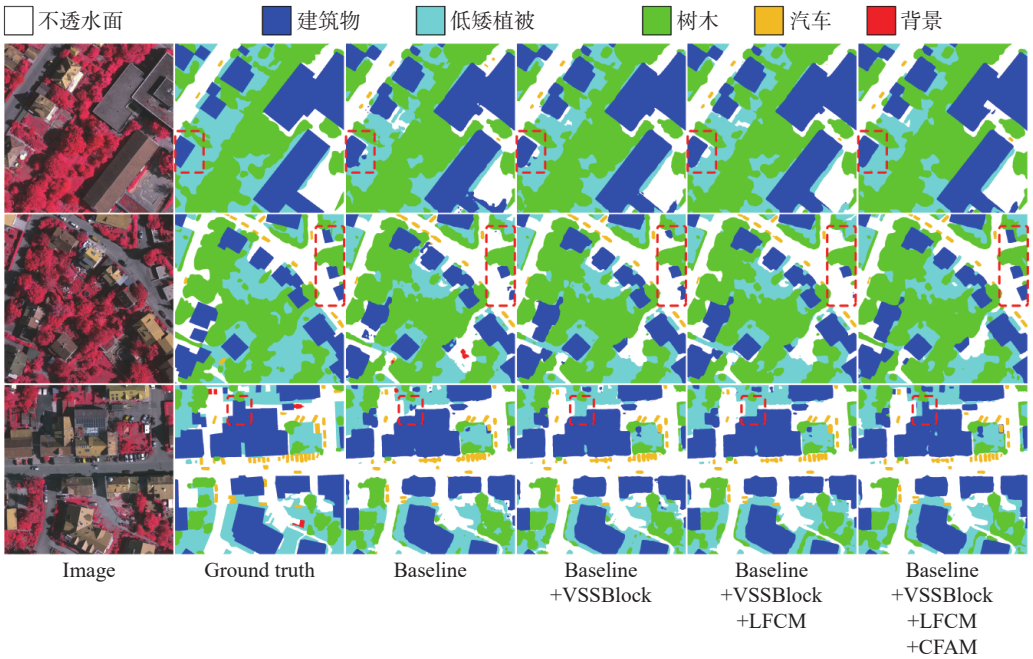

图 8 Vaihingen 数据集上的消融实验可视化结果

在 Vaihingen 数据集上,将 VSSBlock 引入基础模型后,*mIoU* 和 mF_1 分数分别提升了 1.24% 和 0.77%。该性能提升归因于 VSSBlock 依赖 Mamba 的选择性扫描机制,能够以线性计算复杂度有效建模全局上下文信息,从而增强模型对复杂场景的解析能力,并减少了类别混淆问题。进一步集成 LFCM 后,*mIoU* 和 mF_1 相较于基线分别提升至 1.74% 和 1.06%。图 8 显示,LFCM 有效强化了解码阶段的局部细节

建模能力，弥补了 VSSBlock 在细粒度特征表达方面的不足。最终，在集成 CFAM 后，模型整体性能达到最优，*mIoU* 和 mF_1 相较于 Baseline 分别提升了 2.29% 和 1.39%。值得注意的是，LFCM 与 CFAM 在结构上形成了互补协同关系：前者通过增强浅层特征的细粒度表征能力，为跨层融合提供了更高质量的空间细节信息；后者则通过跨层级注意力机制引导深浅特征进行语义对齐与选择性融合，从而实现局部纹理与全局语义的动态互补，使得模型在分割边界和小目标识别方面表现更为稳健，边界模糊与阴影干扰现象明显减少。

在 Potsdam 数据集上，Baseline 模型的 *mIoU* 和 mF_1 分别为 85.73% 和 92.16%。引入 VSSBlock 模块后，*mIoU* 提升至 86.50%，mF_1 分数增加至 92.62%；进一步集成 LFCM 后，两项指标分别提升至 86.79% 和 92.80%；最终，在整合三大模块后，模型性能达到极优，充分验证了模块间存在层次互补与协同增强效应。

在 LoveDA 数据集上，Baseline 模型的 *mIoU* 为 51.37%。引入 VSSBlock 模块后，*mIoU* 提升至 52.12%；进一步整合 LFCM 后，该指标提升至 52.54%；当全部 3 项改进均被整合时，RMUNet 模型在该数据集上达到了极佳性能，说明各模块在跨域场景与类别分布不均衡条件下展现出良好的互补性与协同增益效果。

综上所述，3 个数据集的消融实验结果表明，各模块在提升语义分割性能方面均具有显著有效性。VSSBlock 模块负责全局上下文建模，LFCM 注重细节补偿，而 CFAM 则实现跨层信息融合。三者的协同机制保证了全局语义一致性与局部细节保真度之间的平衡，从而使得整体网络在分割精度和鲁棒性方面均获得显著提升。

3 结 论

本文针对高分辨率遥感影像语义分割中局部细节刻画不足与全局上下文建模效率低下的问题，提出了一种基于 Mamba 架构的遥感影像语义分割模型即 RMUNet 模型。该模型在解码器中引入 VSSBlock 模块，以线性计算复杂度实现全局上下文信息的高效建模；同时，结合 LFCM 与 CFAM，显著提升了模型在复杂场景中的细节表征能力与语义一致性。实验结果表明，RMUNet 模型在 Vaihingen、Potsdam 和 LoveDA 3 个典型数据集上均取得了优于主流模型的分割性能，验证了所提方法在精度与效率之间的平衡性及其实用价值。然而，RMUNet 模型仍存在一定的局限性。首先，模型的参数量与计算成本仍高于部分轻量化网络，限制了其在资源受限平台上的应用潜力。其次，本次实验主要基于城市地物数据集进行验证，模型在多源和多尺度任务中的泛化能力仍有待进一步评估。此外，在应对跨城乡场景差异、类别不均衡及高纹理相似度地物时，模型的特征分离能力与区域一致性建模仍有改进空间。未来的研究工作将重点围绕模型的轻量化设计与结构优化展开，通过参数共享、动态卷积及网络剪枝等技术进一步降低计算复杂度，同时探索光学、SAR 和高光谱等多源异构数据的融合策略，以提升模型在多模态场景下的适应性与鲁棒性。

参考文献：

[1] Liu Y C, Fan B, Wang L F, et al. Semantic labeling in very high resolution images *via* a self-cascaded convolutional neural network[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 145: 78−95.

[2] Schumann G J P, Brakenridge G R, Kettner A J, et al. Assisting flood disaster response with earth observation data and products: a critical assessment[J]. Remote Sensing, 2018, 10(8): 1230.

[3] Chen S W. SAR image speckle filtering with context covariance matrix formulation and similarity test[J]. IEEE Transactions on Image Processing, 2020, 29: 6641−6654.

[4] Zhao T J, Wang S, Ouyang C J, et al. Artificial intelligence for geoscience: progress, challenges, and perspectives[J]. The Innovation, 2024, 5(5): 100691.

[5] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: a brief review[J]. Computational Intelligence and Neuroscience, 2018, 2018(1): 7068349.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on

Neural Information Processing Systems. Long Beach: Curran Associates Inc. , 2017: 6000−6010.

[7] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015: 234−241.

[8] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 833−851.

[9] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6230−6239.

[10] Strudel R, Garcia R, Laptev I, et al. Segmenter: transformer for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 7242−7252.

[11] Xie E Z, Wang W H, Yu Z D, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[C]//Advances in Neural Information Processing Systems 34 (NeurIPS 2021).New York: Curran Associates, 2021: 12077−12090.

[12] Cao H, Wang Y Y, Chen J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation[C]//Proceedings of the European Conference on Computer Vision. Tel Aviv: Springer, 2022: 205−218.

[13] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[14] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: transformers make strong encoders for medical image segmentation[C]// Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part II. Cham: Springer International Publishing, 2021: 234−243.

[15] Wang L B, Li R, Duan C X, et al. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6506105.

[16] Wang L B, Li R, Wang D Z, et al. Transformer meets convolution: a bilateral awareness network for semantic segmentation of very fine resolution urban scene images[J]. Remote Sensing, 2021, 13(16): 3065.

[17] Wang L B, Li R, Zhang C, et al. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196−214.

[18] Zhang H W, Zhu Y, Wang D, et al. A survey on visual Mamba[J]. Applied Sciences, 2024, 14(13): 5683.

[19] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces[C]//Proceedings of the 41st International Conference on Machine Learning (ICML 2024). Vienna, Austria: PMLR, 2024, 235: 16237−16280.

[20] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces[C]//Proceedings of the 10th International Conference on Learning Representations (ICLR 2022). Virtual Conference: OpenReview.net, 2022: uYLFoz1vlAC.

[21] Liu Y, Tian Y J, Zhao Y Z, et al. VMamba: visual state space model[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. , 2024: 3273.

[22] Zhao S J, Chen H, Zhang X L, et al. RS-Mamba for large remote sensing image dense prediction[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5633314.

[23] Zhu Q F, Cai Y Z, Fang Y, et al. Samba: semantic segmentation of remotely sensed images with state space model[J]. Heliyon, 2024, 10(19): e38495.

[24] Zhu E Z, Chen Z, Wang D K, et al. UNetMamba: an efficient UNet-like Mamba for semantic segmentation of high-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 6001205.

[25] Ma X P, Zhang X K, Pun M O. RS^{3}Mamba: visual state space model for remote sensing image semantic segmentation[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 6011405.

[26] Wang L B, Li D X, Dong S J, et al. PyramidMamba: rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery[J]. International Journal of Applied Earth Observation and Geoinformation, 2024, 144: 104884.

[27] Liu M S, Dan J, Lu Z Q, et al. CM-UNet: hybrid CNN-Mamba UNet for remote sensing image semantic

segmentation[PP/OL]. arXiv (2024-05-16)[2025-07-01]. https://arxiv.org/abs/2405.10530.

[28] Milletari F, Navab N, Ahmadi S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C]//Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford: IEEE, 2016: 565−571.

[29] Zhu Z, Xu M D, Bai S, et al. Asymmetric non-local neural networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 593−602.

作者贡献声明

张建昌负责论文的整体构思与实验方案设计,撰写论文初稿。徐伟铭提供研究经费。余大文负责全过程的论文指导。梁昌钰负责实验数据的分析与解释。陈枭负责实验数据的收集与处理。

利益冲突声明

作者声明,不存在已知的可能影响本论文所报告工作的竞争性经济利益或个人关系。

AI 使用声明

本文的英文题名、摘要和关键词采用腾讯元宝翻译生成,并进行了部分修改。

伦理声明

本研究不涉及人类或动物的生物学研究。

电子补充材料

补充材料可在本文的在线版本中获取 https://doi.org/10.65658/j.hndk.2025101501。

数据可用性

支持论文结论所需的所有数据均为公开数据集。

致谢/Acknowledgements

本文得到国家自然科学基金项目 (41801324) 的支持。

This work was supported by the National Natural Science Foundation of China (41801324).

作者简介

张建昌(2001—),男,江西南昌人,福州大学数字中国研究院地理学专业 2023 级硕士研究生。E-mail: 1641876521@qq.com。

徐伟铭(1986—),男,福建福州人,副研究员,研究方向:城市变化检测和目标识别、智慧城市应用研究。E-mail: xwming2@126.com。

余大文(1995—),男,福建南平人,助理研究员,研究方向:高分辨率遥感影像智能解译的理论与方法。E-mail: yudawen@fzu.edu.cn。

梁昌钰(2000—),男,海南三亚人,福州大学数字中国研究院测绘工程专业 2023 级硕士研究生。E-mail: 239254391@qq.com。

陈　枭(2003—),男,江西吉安人,福州大学数字中国研究院地理学专业 2024 级硕士研究生。E-mail: 2370549719@qq.com。

(责任编辑:张秀梅,英文校对:王玖琪)