

基于 Swin Transformer 与卷积神经网络的高分遥感影像分类

何小英^{1,2,3}, 徐伟铭^{1,2,3*}, 潘凯祥^{1,2,3}, 王娟^{1,2,3}, 李紫微^{1,2,3}

¹福州大学数字中国研究院(福建), 福建 福州 350108;

²福州大学空间数据挖掘与信息共享教育部重点实验室, 福建 福州 350002;

³福州大学地理空间信息技术国家地方联合工程研究中心, 福建 福州 350002

摘要 针对现有基于深度学习的遥感智能解译方法直接获取全局信息具有挑战性,造成地物边缘模糊、相似类间分类精度低等问题,提出基于 Swin Transformer 和卷积神经网络的高分遥感图像语义分割模型(SRAU-Net)。SRAU-Net 以 Swin Transformer 编码器-解码器框架为基础,采用 U-Net 形状,提出了以下改进:构造基于 Swin Transformer 和基于卷积神经网络的双分支编码器,用不同尺度空间细节特征补充具有全局信息的上下文特征,以获得更高的地物分类精度和更清晰的地物边缘;设计特征融合模块,作为双分支编码器的桥梁从通道和空间维度对全局和局部特征进行有效融合,提升对小目标地物的分割精度;添加特征增强模块,利用注意力机制自适应融合来自编码器和解码器的特征,进一步有效聚合空间和语义特征,提升模型对特征的提取效果。结果表明,SRAU-Net 能够更好地提取地物的边缘信息,总体分类精度较原始模型提升了 2.57 百分点,提高了对小尺度地物的分类精度,有效区分如树木和低矮植被等类间相似的遥感地物,总体精度和 F1 分数分别为 92.60% 和 86.90%,总体效果优于对比模型。

关键词 高分辨率遥感影像; 卷积神经网络; Swin Transformer; 特征融合; 语义分割

中图分类号 TP751

文献标志码 A

DOI: 10.3788/LOP232003

Classification of High-Resolution Remote Sensing Image Based on Swin Transformer and Convolutional Neural Network

He Xiaoying^{1,2,3}, Xu Weiming^{1,2,3*}, Pan Kaixiang^{1,2,3}, Wang Juan^{1,2,3}, Li Ziwei^{1,2,3}

¹The Academy of Digital China, Fuzhou University, Fuzhou 350108, Fujian, China;

²Key Laboratory of Spatial Data Mining & Information Sharing, Ministry of Education, Fuzhou University, Fuzhou 350002, Fujian, China;

³National Engineering Research Center of Geospatial Information Technology, Fuzhou University, Fuzhou 350002, Fujian, China

Abstract It is challenging to directly obtain global information of existing deep learning-based remote sensing intelligent interpretation methods, resulting in blurred object edges and low classification accuracy between similar classes. This study proposes a semantic segmentation model called SRAU-Net based on Swin Transformer and convolutional neural network. SRAU-Net adopts a Swin Transformer encoder-decoder framework with a U-Net shape and introduces several improvements to address the limitations of previous methods. First, Swin Transformer and convolutional neural network are used to construct a dual-branch encoder, which effectively captures spatial details with different scales and complements the context features, resulting in higher classification accuracy and sharper object edges. Second, a feature fusion module is designed as a bridge for the dual-branch encoder. This module efficiently fuses global and local features in channel and spatial dimensions, improving the segmentation accuracy for small target objects. Moreover, the proposed SRAU-Net model incorporates a feature enhancement module that utilizes attention mechanisms to adaptively fuse features from the encoder and decoder and enhances the aggregation of spatial and semantic features, further improving the ability

收稿日期: 2023-08-29; 修回日期: 2023-09-29; 录用日期: 2023-11-21; 网络首发日期: 2023-12-12

基金项目: 国家自然科学基金(41801324)、福建省科技厅引导性项目(2017Y0055, 2022H0009)、教育部产学研合作协同育人项目(202101119001)

通信作者: *xwming2@126.com

of the model to extract features from remote sensing images. The effectiveness of the proposed SRAU-Net model is demonstrated using the ISPRS Vaihingen dataset for land cover classification. The results show that SRAU-Net outperforms other models in terms of overall accuracy and F1 score, achieving 92.06% and 86.90%, respectively. Notably, the SRAU-Net model excels in extracting object edge information and accurately classifying small-scale regions, with an improvement of 2.57 percentage points in the overall classification accuracy compared with the original model. Furthermore, it effectively distinguishes remote sensing objects with similar characteristics, such as trees and low vegetation.

Key words high-resolution remote sensing image; convolutional neural network; Swin Transformer; feature fusion; semantic segmentation

1 引言

高分辨率遥感影像分类根据地的纹理、光谱等特点识别目标地物,广泛应用于土地覆盖研究^[1-2],为国土资源调查^[3]、生态环境评估^[4]、国土空间规划^[5]等提供了数据和技术支撑。随着航空航天技术和传感器技术的发展,遥感影像的复杂性和异质性增加,遥感影像分类任务仍具有挑战性^[6]。深度学习等人工智能技术的发展为遥感智能解译提供了新的实现途径。

传统高分辨率遥感影像分类的主要依据是地物光谱,需要分类人员有丰富的先验知识和实践经验,难以满足现阶段基于遥感数据的信息提取需求^[7]。基于语义分割的遥感影像分类算法计算效率高,学习特征能力强,可实现模型特征提取、上采样和特征融合等一系列改进^[8-11]。其中,基于卷积神经网络(CNN)的方法可捕获遥感影像中的局部细节信息^[12-13]。李宏达等^[14]发现随着遥感影像分辨率的提高,CNN分类效果越好,表明CNN适用于高空间分辨率影像分类。全卷积网络(FCN)^[15]使用卷积层代替了全连接层,以克服卷积和池化对图像大小的影响,并实现了端到端的图像分割。Ronneberger等^[16]以全卷积神经网络为基础设计了U-Net,其具备捕捉上下文信息和能精确定位的编解码结构。

卷积神经网络一般通过增加深度来增加网络感受野,间接获取全局上下文信息。卷积神经网络深度增加后计算效率降低,且卷积操作的局部性获取全局上下文具有挑战性^[17]。注意力机制和多尺度特征融合策略是补充卷积神经网络全局信息的常用方法^[18-19]。Li等^[20]通过设计一种线性注意力机制,对不同位置的全局上下文依赖关系进行建模,捕获细粒度的特征。Liu等^[21]采用自联级的方式对获取的多尺度上下文信息进行聚合,并利用残差校正实现更有效的多特征融合。这些方法在CNN提取局部特征后的基础上提取全局信息,而不是直接获取全局上下文,因此,很难从地物类间相似性高、地物大小不均的复杂遥感图像中获得可靠的全局场景信息^[22]。Transformer在自然语言领域获得的成功为计算机视觉任务中的全局关系建模开辟了新的研究思路^[23-24]。Swin Transformer^[25]采用层级设计获取多尺度特征,该方法在不同的密集型预测

任务中显示出巨大的潜力^[26-27]。

综上,本文针对现有高分辨率遥感影像分类方法难以直接获取全局上下文的问题,提出了基于Swin Transformer的遥感图像语义分割编码器-解码器框架(SRAU-Net)。SRAU-Net采用U-Net的形状,双分支编码器为基于Swin Transformer的主编码器和基于卷积神经网络的辅助编码器,充分利用Transformer模型中的注意力单元捕捉全局信息,并利用卷积神经网络获取邻域像素之间的关系进行建模,提取局部信息。特征融合模块对图像全局特征与局部特征进行有效融合,提升对相似地物的分割精度;特征增强模块抑制无效通道对地物分类的影响,提升模型对细节特征的利用。在ISPRS Vaihingen数据集和GID数据集上进行遥感影像分类实验,并与U-Net、DeeplabV3+和TransUNet等不同分割网络进行对比。

2 研究方法

2.1 SRAU-Net结构概述

Transformer模型中的注意力单元能对长距离的依赖关系进行建模,以捕捉全局信息;而卷积神经网络则对邻域像素之间的关系进行建模,获取局部信息。本研究结合有效的U-Net模型结构,充分利用Transformer和CNN的优点,设计了具有双分支编码器的SRAU-Net,总体结构如图1所示。

SRAU-Net总体架构包括主编码器、辅助编码器和解码器,遵循U形结构利用跳跃连接将编码器和解码器相连。对于主编码器,将输入的遥感图像 $\mathbf{X} \in \mathbf{R}^{H \times W \times 3}$ (其中 H 和 W 分别代表图像高度和宽度)利用序列嵌入(patch partition)转换为 $(H/4, W/4, 48)$ 的非重叠图像块,通过线性嵌入层(linear embedding)将图像块维度投影到任意维度。将投影后的图像块输入由3个特征提取阶段构成的主编码器中,每个阶段的特征输出定义为 $\mathbf{M}_n$,其中 $n=1, 2, 3$ 。基于Swin Transformer的主编码器借鉴了卷积神经网络中的层次结构,以此获得不同尺度的特征图,每个特征提取阶段包含2个Swin Transformer模块和1个图像块合并层(patch merging)。Swin Transformer模块对图像特征进行学习提取;图像块合并层通过下采样将特征尺寸减半、特征维度增倍。第 n 个阶段的输出特征图可

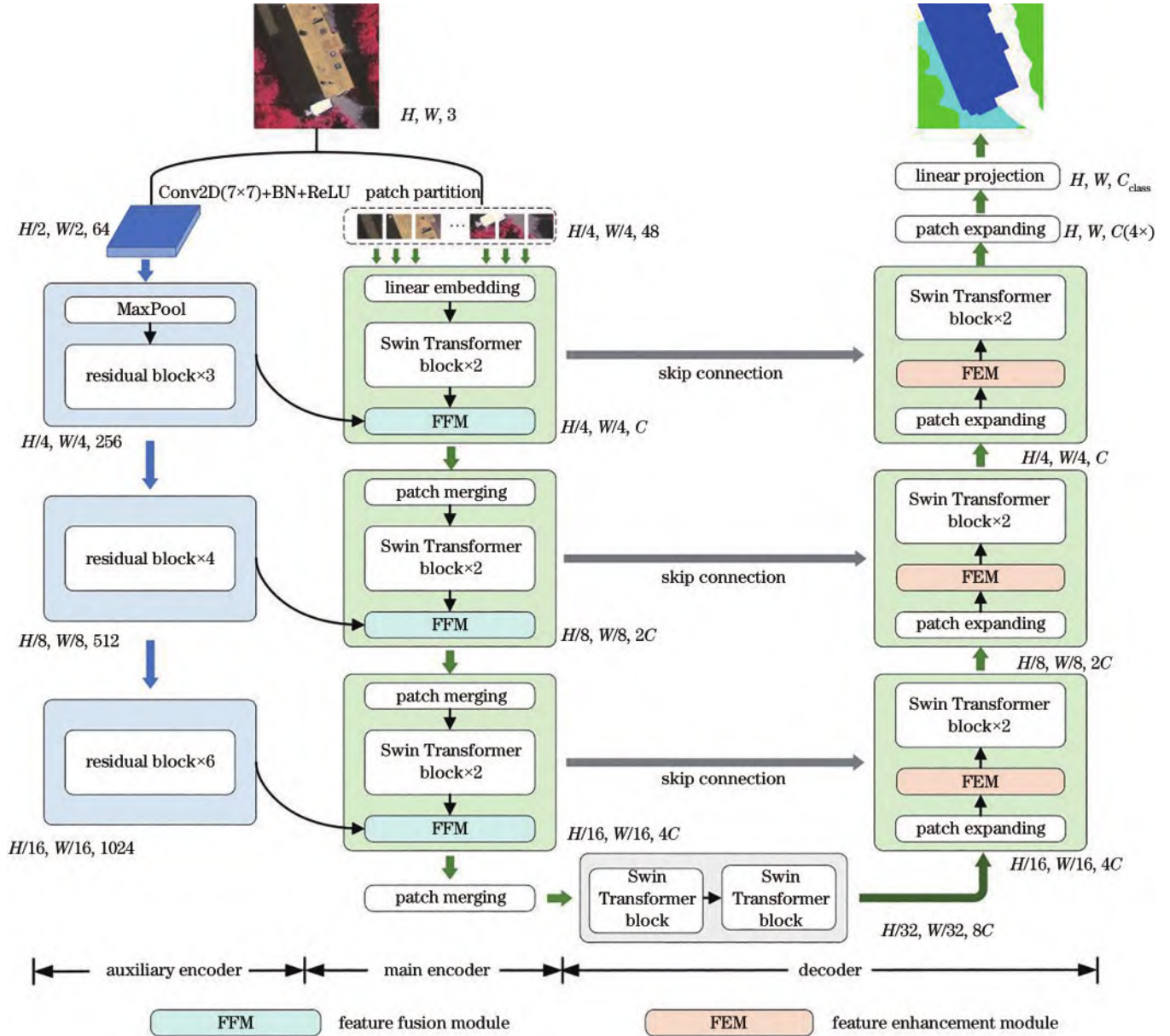


图1 SRAU-Net总体框架

Fig. 1 Overall architecture of SRAU-Net

以表示为 $M_n \in \mathbf{R}^{(H/2^{n+1}) \times (W/2^{n+1}) \times 2^{n-1}C_1}$, $C_1 = 96$ 。对于辅助编码器,原始遥感图像输入至由残差块(residual block)组成的辅助编码器中,第 n 个阶段的残差块输出特征图表示为 $A_n \in \mathbf{R}^{(H/2^{n+1}) \times (W/2^{n+1}) \times 2^{n-1}C_2}$, $n = 1, 2, 3$, $C_2 = 256$ 。将对应阶段的主编码器输出 M_n 和辅助编码器输出 A_n 输入至特征融合模块(FFM),并将融合结果返回至主编码器。解码器由 Swin Transformer 模块和图像块扩展层组成,通过跳跃连接对提取的特征与编码器不同尺度特征进行融合,以弥补降采样造成的信息丢失。进一步利用特征增强模块(FEM)减少对冗余通道的关注,图像块扩展层将相邻维度的特征图以 2 倍上采样的方式重塑为一个新的特征图。最后一个图像块扩展层对特征图进行 4 倍上采样,将特征图的分辨率恢复到输入尺寸,并利用线性投影将上采样

结果转换为分割类别维度。

2.2 双分支编码器

SRAU-Net 总体架构的编码器由基于 Swin Transformer 的主编码器和基于 CNN 的残差网络辅助编码器构成。基于 Swin Transformer 的主编码器通过层级结构提取不同尺度的全局信息,基于残差模块的辅助编码器能够有效获取局部细粒度信息,在编码器中添加局部空间信息,增强模型的地物边缘提取能力。

2.2.1 Swin Transformer 模块

图 2(a) 所示标准 Transformer 模块由归一化层(LN)、多头自注意力(MSA)、残差连接和包含 GeLU 激活函数的多层感知机(MLP)构成。具体的第 l 层的输出 z_l 表示为

$$\hat{z}_l = \text{MSA}[\text{LN}(z_{l-1})] + z_{l-1}, \quad (1)$$

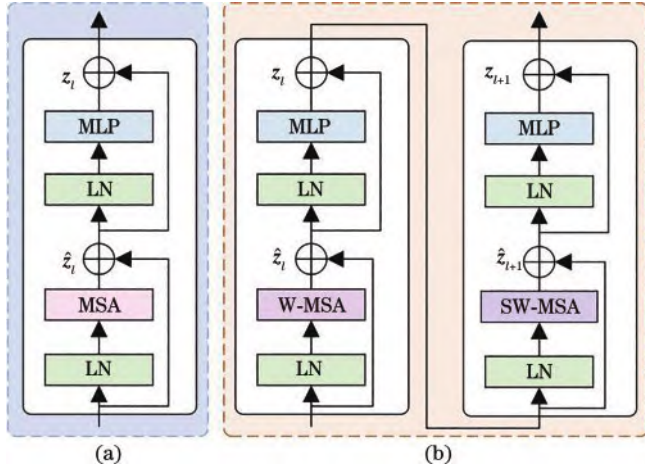


图2 标准Transformer模块和Swin Transformer模块。(a)标准Transformer模块;(b)Swin Transformer模块

Fig.2 Standard Transformer block and Swin Transformer block. (a) Standard Transformer block; (b) Swin Transformer block

$$z_l = \text{MLP}[\text{LN}(\hat{z}_l)] + \hat{z}_l. \quad (2)$$

图2(b)所示Swin Transformer模块基于移动窗口设计,并提供了基于窗口的多头自注意力(W-MSA)、基于移动窗口的多头自注意力(SW-MSA)。在标准的MSA中,计算自注意力时特征图中的每个像素需要和其他每一个像素进行计算,对于密集预测型计算机图形任务来说,计算复杂度大。Swin Transformer在窗口内计算自注意力即W-MSA,窗口大小固定,特征图的计算复杂度与特征图的尺寸成线性增长关系,并利用了图像局部性的先验知识。而W-MSA只会计算每个窗口内部的自注意力,窗口之间缺少信息交互,因此提出了SW-MSA。SW-MSA通过向右下角方向移动窗口,实现不同窗口中像素的交互,更好地获取上下文信息。对于Swin Transformer模块,第 l 层W-MSA、SW-MSA的输出分别为 z_l, z_{l+1} ,具体的表示为

$$\hat{z}_l = \text{W-MSA}[\text{LN}(z_{l-1})] + z_{l-1}, \quad (3)$$

$$z_l = \text{MLP}[\text{LN}(\hat{z}_l)] + \hat{z}_l, \quad (4)$$

$$\hat{z}_{l+1} = \text{SW-MSA}[\text{LN}(z_l)] + z_l, \quad (5)$$

$$z_{l+1} = \text{W-MSA}[\text{LN}(z_{l-1})] + z_{l-1}. \quad (6)$$

2.2.2 残差模块

对于卷积神经网络,随着网络深度的增加极易出现两类问题:一是随着网络深度增加出现的退化问题,训练误差没有持续降低反而升高,导致模型准确度饱和甚至下降;二是神经网络进行反向传播时,反向连乘的梯度小于1(或大于1),出现的梯度消失(爆炸)问题。He等^[28]和Shafiq等^[29]提出了残差模块来解决上述问题。该模块能够有效构建深度网络结构,并解决了网络层数加深出现的精度下降问题与梯度消失(爆炸)问题,计算复杂度小,计算速度快。残差模块利用卷积块学习特征,并使用 1×1 卷积核进行降维升维,

再进行原始特征矩阵 x 和学习特征矩阵 $F(x, W)$ 的相加,得到最终的学习特征即残差学习。残差模块中的残差学习能够使模型在原始特征基础上学习到新的特征,从而拥有更好的性能,具体的残差模块如图3(b)所示。残差单元可表示为

$$x_{l+1} = \text{ReLU}[x_l + F(x_l, W_l)], \quad (7)$$

式中: x_{l+1} 和 x_l 表示残差单元的输出和输入; $\text{ReLU}(\cdot)$ 表示ReLU激活函数; $F(x_l, W_l)$ 表示残差函数,其中 W_l 表示残差函数对应的权重。

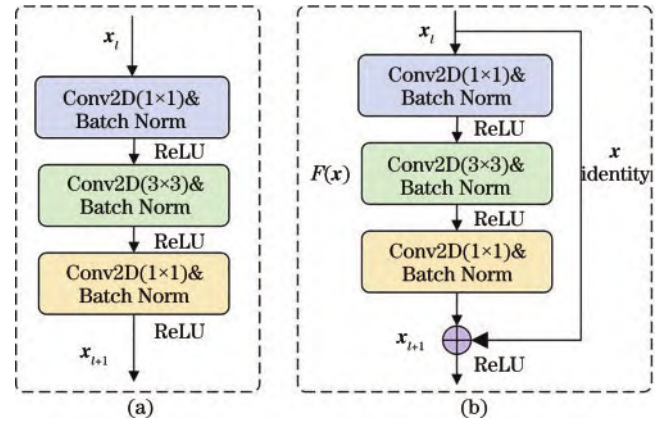


图3 卷积模块与残差模块。(a)卷积模块;(b)残差模块

Fig. 3 Convolution block and residual block. (a) Convolution block; (b) residual block

2.3 特征融合模块

基于Swin Transformer的主编码器通过窗口内的自注意力计算提取图像的多尺度全局特征,基于残差模块的辅助编码器提取受卷积核限制的局部信息,如何将提取的图像全局特征与局部特征有效融合且尽量不损失空间和通道信息至关重要。一些方法^[30-32]已经证明对通道进行空间维度编码获得的通道依赖性可以提高特征识别能力,同时对通道进行压缩在通道维度进行编码得到的空间注意力特征可以帮助模型关注更为重要的特征区域。因此,设计了FFM,其详细结构如图4所示。

为了从不同特征图中强调重要且更具代表性的通道和空间信息,将主编码器的全局特征和辅助编码器的局部特征拼接后依次对通道和空间信息进行编码,获得最终的融合特征。在FFM中,将 M_n 和 A_n 分别表示第 n 阶段主编码器和辅助编码器的输出,通过拼接获得特征图 $F_n \in \mathbf{R}^{(H/2^{n+1}) \times (W/2^{n+1}) \times C_3}$, $C_3 = 2^{n-1}C_1 + 2^{n-1}C_2$,其中 C_1 和 C_2 分别表示主编码器和辅助编码器第 n 个阶段的输出通道数。采用了两种池化策略来获得更全面的通道重要性,利用平均池化层和最大池化层来计算通道上特征图的统计特征,并将其传送至共享全连接层,通过将两者相加得到通道重要性权值 $W_c \in \mathbf{R}^{1 \times 1 \times C_3}$,具体表示为

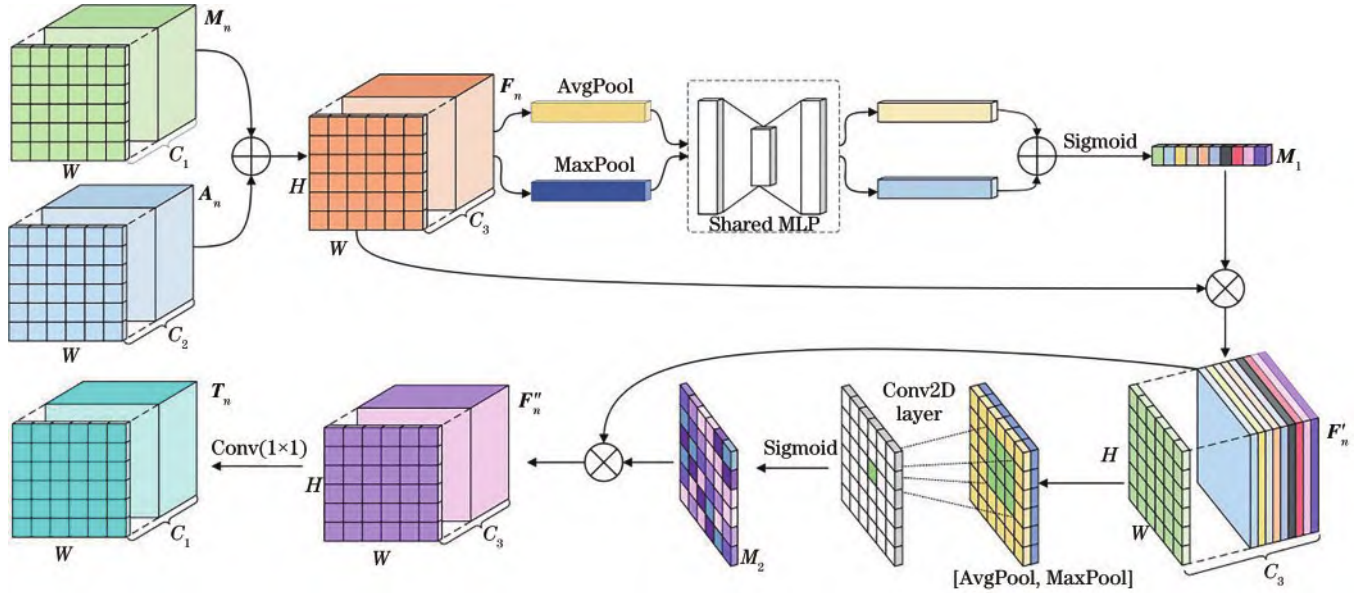


图4 特征融合模块

Fig.4 FFM

$$W_c = \text{Sigmoid} \left\{ \text{MLP} \left[\text{AvgPool}(F_n) \right] + \text{MLP} \left[\text{MaxPool}(F_n) \right] \right\}, \quad (8)$$

式中: Sigmoid 为激活函数; MLP 为一个多层感知机, 包含 2 层神经网络, 激活函数为 ReLU。

将通道重要性权值 W_c 与 F_n 相乘, 获得具有通道依赖的特征 F'_n , 为进一步对空间区域进行自适应选择, 在通道维度分别进行平均池化和最大池化, 拼接后得到空间重要性权值 $W_s \in \mathbf{R}^{H \times W \times 1}$, 具体表示为

$$W_s = \text{Sigmoid} \left\{ f^{3 \times 3} \left[\text{AvgPool}(F'_n) + \text{MaxPool}(F'_n) \right] \right\}, \quad (9)$$

式中: $f^{3 \times 3}$ 表示卷积核为 3×3 的卷积层。

将空间重要性权值 W_s 与具有通道依赖的 F'_n 相乘, 获得特征 F''_n , 最后通过 1×1 卷积对 F''_n 进行降维以

减少计算量, 获得 FFM 的输出特征 T_n , 表示为

$$T_n = f^{1 \times 1} \left[W_s \otimes (W_c \otimes F_n) \right], \quad (10)$$

式中: $f^{1 \times 1}$ 表示卷积核为 1×1 的二维卷积层。

2.4 特征增强模块

SRAU-Net 利用跳跃连接对高维特征与低维特征进行融合, 能够保留更多高分辨率细节信息, 但在融合的过程中简单的拼接易导致对丰富信息的利用不足。因此, 在跳跃连接之后加入图 5 所示的 FEM, 强调模型所需的重要特征, 抑制无效通道对地物分类的影响, 有利于提升模型对细节特征的利用。FEM 利用一维卷积替换传统全连接层, 可避免降维对通道注意力学习效果的影响。首先对输入特征图进行通道全局平均池化和全局最大池化, 从而得到不同通道特征图中的重要信息; 进一步根据特征向量的通道维数自适应确

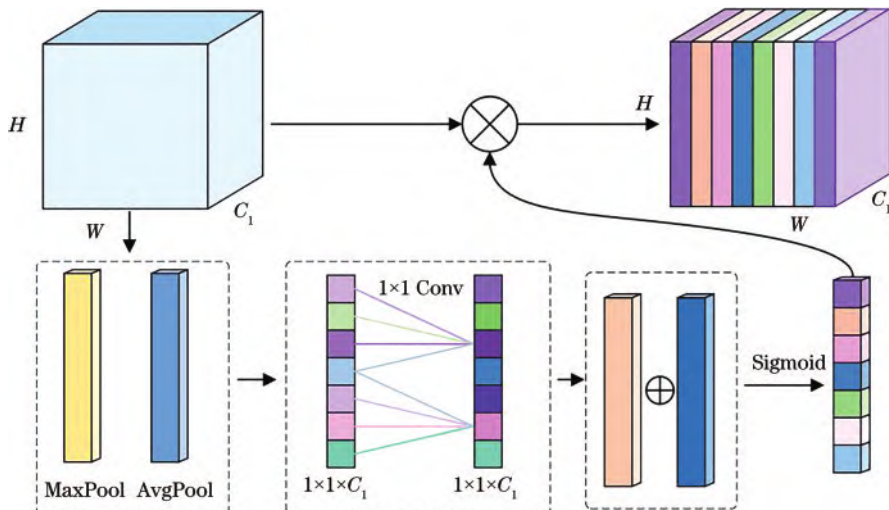


图5 特征增强模块

Fig.5 FEM

定 k 值,再使用卷积核为 k 的一维卷积和激活函数计算提取通道间的依赖关系;最后将各通道权重与输入特征图相乘,获得最终的重要特征图。这过程具体表示为

$$\mathbf{X}_n' = \mathbf{X}_n \times \text{Sigmoid}\left\{f^{1 \times k}\left[\text{AvgPool}(\mathbf{X}_n)\right] + f^{1 \times k}\left[\text{MaxPool}(\mathbf{X}_n)\right]\right\}, \quad (11)$$

式中: $\mathbf{X}_n'$ 和 $\mathbf{X}_n$ 表示第 n 阶段 FEM 的输出和输入特征; $f^{1 \times k}$ 表示卷积核为 $1 \times k$ 的一维卷积层。

2.5 损失函数

遥感数据集的类别比例通常不平衡,导致模型训练侧重于占比较大的类别,会忽略占比较小的类别。为了缓解这一问题,采用带权重的交叉熵(WCE)损失函数和 Dice 损失函数联合的损失函数对模型进行监督。WCE 损失函数针对不同的类别,使用相应的权重衡量样本标签和预测结果之间的像素级相似性;Dice 损失函数用于评估真实样本和预测结果在整体上的相似度。WCE 损失函数的表达式为

$$l_{\text{WCE}} = -\frac{1}{N} \sum_i \sum_{c=1}^M w_c y_{ic} \log P_{ic}, \quad (12)$$

$$w_c = \text{Median}(F_s)/F_c, \quad (13)$$

式中: N 表示输入图像的像素总数; M 代表类别数量; P_{ic} 表示样本 i 属于第 C 类的概率; y_{ic} 表示当样本 i 属于第 C 类时取值; w_c 表示第 C 类的权重,将在实验数据部分统计计算; F_c 表示第 C 类的频率; F_s 表示全部类别频率的集合。

Dice 损失函数常使用混淆矩阵计算,公式为

$$l_{\text{Dice}} = 1 - \frac{2N_{\text{TP}}}{2N_{\text{TP}} + N_{\text{FN}} + N_{\text{FP}}}, \quad (14)$$

式中: N_{TP} 、 N_{FN} 、 N_{FP} 分别表示预测正确的正样本像素个数、预测错误的负样本像素个数、预测错误的正样本像素个数。最终的联合损失函数表示为

$$l = l_{\text{WCE}} + l_{\text{Dice}} \quad (15)$$

3 数据及评价指标

3.1 实验数据

Vaihingen 数据集是由 ISPRS 组织提供的包含 33 张先进机载传感器收集的真实地表图像,主要用于城市分类,地面采样距离约为 0.09 m。每幅图像都有红外、红色和绿色通道,其中 16 幅图像准确分类为 6 种类别,分别为不透水面、建筑物、低矮植被、树木、汽车和背景。从表 1 可以看出 Vaihingen 数据集中类别分布极不平衡,不透水面和建筑物占比超过 50%,而汽车和杂乱/背景仅占 2%。选择 Vaihingen 数据集完全注释图像中的 11 张图像进行训练,5 张图像用于测试,裁剪尺寸为 256×256 像素。为确保模型在不同批次中看到数据集的不同方面,通过图像随机旋转、镜像、翻转和添加噪声等方式进行数据增强,最后训练集、验证集和测试集比例为 7:2:1。

表 1 Vaihingen 数据集类别分布及权重计算

Table 1 Class distribution and weight calculation of Vaihingen dataset

Class name	Number of pixels	Class frequency	Weight
Impervious surface	46792757	0.2780	0.7939
Building	43779851	0.2601	0.8486
Low vegetation	35766767	0.2125	1.0387
Tree	38534748	0.2290	0.9641
Car	2096078	0.0125	17.7239
Clutter/background	1317670	0.0078	28.1943

GID 数据集包含大比例尺分类集和精细土地覆盖分类集,本文使用大比例尺分类集。GID 大比例尺分类集由 150 张分辨率为 1 m 的 GF-2 影像组成,包含蓝色、绿色、红色和近红外波段,图像尺寸为 6800×7200 像素,其中训练集图像有 120 张,验证集图像有 30 张。土地覆盖分类为 6 个主要类别:建筑、水体、森林、农田、草地及未标记区域。从表 2 可以看出 GID 大比例尺数据集中类别分布同样极不平衡,特别是未标记区域占比超过 40%,草地占比仅为 2%。数据裁剪和增强方式与 Vaihingen 数据集相同。

表 2 GID 数据集类别分布及权重计算

Table 2 Class distribution and weight calculation of GID dataset

Class name	Number of pixels	Class frequency	Weight
Building	651080927	0.0899	1.0996
Water	780799058	0.1079	0.9169
Forest	277330405	0.0383	2.5815
Farmland	2222929336	0.3071	0.3221
Meadow	144943831	0.0200	4.9394
Others	3162486239	0.4368	0.2264

3.2 实验环境与策略

本研究实验环境为 Windows 10 操作系统,CPU 为 Intel(R)Xeon(R)Platinum 8255C CPU@2.50 GHz, GPU 为 NVIDIA RTX 3080,采用 Python 编程语言,相关模型基于 PyTorch 深度学习框架实现。超参数配置:SGD 作为模型优化器,其动量项为 0.9,权重衰减为 1×10^{-4} ;设置批次大小为 8;初始学习率为 1×10^{-4} ,并使用余弦退火学习率策略来设置每个参数组的学习率,该策略通过余弦函数来降低学习率;训练迭代次数为 150,保证充分拟合。

3.3 评价指标

选取在图像语义分割任务中常用的 F1 分数(F1-score)、总体精度(OA)和平均交并比(MIoU)作为评价指标,评价本研究模型的可靠性并与现有主流方法进行比较。F1-score 是精确率和召回率的组合度量,同时评估模型的准确率和召回率;OA 计算正确预测的像素个数占比;IoU 是各类别预测正确值与真实标

签值的交集与并集的比值,用于评价各类别的分类精度,MIoU为所有类别IoU的平均值。通过混淆矩阵来计算各评价指标,其中列代表预测类别,行代表真实类别,具体如表3所示。

表3 混淆矩阵
Table 3 Confusion matrix

Actual	Predict	
	Positive	Negative
Positive	True positive(N_{TP})	False positive(N_{FP})
Negative	False positive(N_{FP})	True negative(N_{TN})

各评价指标的具体计算过程分别为

$$P_{\text{precision}} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad (16)$$

$$R_{\text{recall}} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad (17)$$

$$F_1 = 2 \times \frac{P_{\text{precision}} \times R_{\text{recall}}}{P_{\text{precision}} + R_{\text{recall}}}, \quad (18)$$

$$P_{OA} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \quad (19)$$

$$R_{MIoU} = \frac{1}{N} \sum_{i=1}^N \frac{N_{TP}}{N_{FP} + N_{TP} + N_{FN}}. \quad (20)$$

4 实验结果与分析

4.1 同类模型结果对比分析

将所提SRAU-Net与现有的同类模型进行比较,包括FCN^[15]、U-Net^[16]、DeeplabV3+^[33]和TransUNet^[17],前3种模型都是传统的CNN,TransUNet是U-Net和

ViT的组合。其中TransUNet采用了Transformer和CNN的串行结构,所提SRAU-Net则是Transformer与CNN的并行结构。在Vaihingen数据集和GID数据集上开展实验,并对不同模型进行了定量和定性的比较。

Vaihingen数据集上不同语义分割模型的定量结果如表4所示,在分析结果时忽略了占比不足1%的杂乱/背景类别。从表4可以看出:所提SRAU-Net的各类评价指标均优于其他模型,MIoU为81.35%、OA为92.60%、F1分数为86.90%,表明具有全局建模能力和局部信息提取能力的SRAU-Net能够较好地完成复杂遥感语义分割任务。从同类模型的评价结果中可以发现:将传统CNN中的全连接层替换为卷积层的FCN没有充分考虑像素与像素之间的关系,忽略了在遥感影像中至关重要的空间一致性信息,总体精度仅为85.67%;U-Net通过跳跃连接不断融合来自低级特征的空间信息,其产生的分割结果较好,三类评价指标值均较高,总体精度为89.57%;DeeplabV3+充分利用多尺度特征,其结果相较于FCN有较大的提升,总体精度为89.39%;TransUNet在U-Net架构编码器的底层串联了一个Transformer模块,同时学习来自CNN的高分辨率空间信息和来自Transformer的全局上下文信息,与U-Net相比MIoU增加了0.95个百分点,F1分数增加了1.05个百分点,表明CNN与Transformer的组合是有效的。与TransUNet模型相比,SRAU-Net的MIoU提升了2.25个百分点,OA提升了2.58个百分点,F1提升了3.26个百分点。

表4 不同模型在Vaihingen数据集上的分割结果比较

Table 4 Comparison of segmentation results of different models on Vaihingen dataset

Method	IoU / %					MIoU / %	OA / %	F1 / %
	Impervious surface	Building	Low vegetation	Tree	Car			
FCN	65.96	86.34	67.09	75.63	67.17	72.44	85.67	76.56
U-Net	80.40	88.07	70.32	77.78	74.16	78.15	89.57	82.59
DeeplabV3+	81.15	88.62	69.97	78.32	73.95	78.40	89.39	82.93
TransUNet	81.26	89.79	71.01	78.49	74.95	79.10	90.02	83.64
SRAU-Net	83.88	90.97	73.22	80.95	77.71	81.35	92.60	86.90

为验证所提SRAU-Net识别遥感图像中不同地物的能力,对不同模型地物识别结果进行了定性分析,不同模型对从Vaihingen数据集中随机选取的4幅图像进行可视化,结果如图6所示。可以看出:由于FCN没有考虑到遥感图像中像素之间的相关性,整体分割效果较差,如第二幅图像低矮植被和树木两个类别的错分情况较多;U-Net整体分类结果较好,但缺乏全局信息建模,仍会出现相似地物的错分现象;DeeplabV3+采用扩张卷积,从而获得更大的感受野,相较于FCN,分类结果有大幅提升,但没有较好地解决地物阴影遮挡问题,如第四幅图像的右下角房屋阴

影遮挡的不透水面被错分为建筑;TransUNet获得比DeeplabV3+更好的分割效果,但无法准确分割遥感图像地物边缘信息,如第三幅图像中没有准确地分割出三辆小目标汽车;基于CNN的分割模型能够更好地捕捉密集分布的建筑局部特征,分割指标表现更好,与其他模型相比,SRAU-Net减小了分割误差,特别是对Vaihingen数据集中相似度高的地物,CNN与Transformer的并行结构可充分利用不同尺度的空间信息增强了地物边缘信息提取能力。

为进一步证明所提SRAU-Net的有效性,在GID大比例尺数据集上进行实验,具体分割结果如表5所

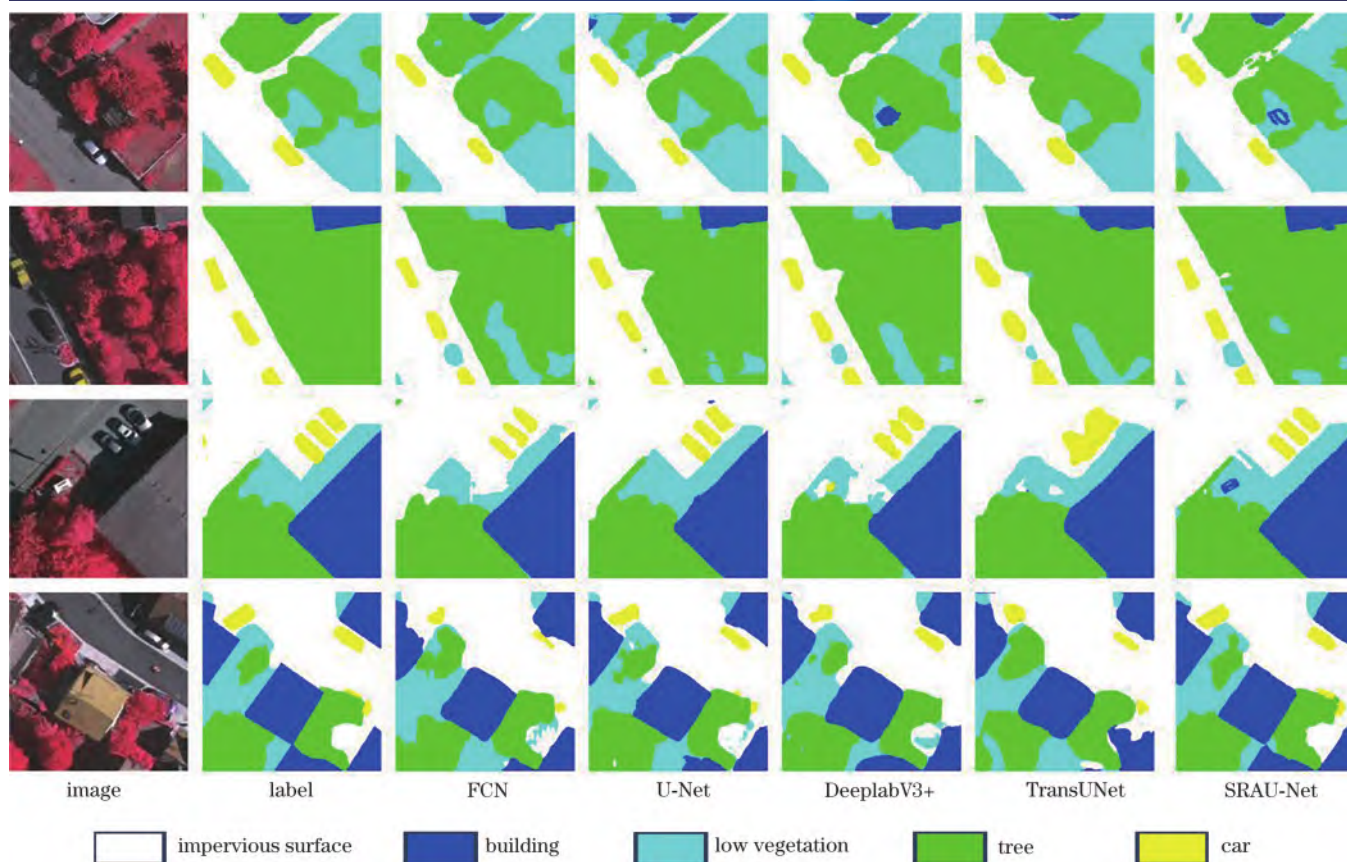


图6 不同模型在 Vaihingen 数据集上的分割结果可视化

Fig. 6 Visualization of segmentation results of different models on Vaihingen dataset

示。由于 GID 大比例尺数据集中背景类别是未知的,在评估时忽略该类别。可以看出 SRAU-Net 的 MIoU 达到 67.35%、OA 达到 84.26%、F1 达到 71.84%,均优于其他方法。由于数据集的质量和数据类型不同,GID 大比例尺数据集的分割精度一般低于 ISPRS 提供的 Vaihingen 数据集。从表 5 还可以发现,混合结构的

TransUNet 的分割精度超过了表 4 中其余基于 CNN 的模型,表明基于 CNN 的模型在提取全局信息特征时存在一定的局限性。与 TransUNet 相比,SRAU-Net 的 MIoU、OA 和 F1 分数分别提高了 1.04 百分点、1.35 百分点、2.72 百分点。与其他分割模型比较,SRAU-Net 提高了对建筑、农田和草地 3 类的分割精度。

表 5 不同模型在 GID 数据集上的分割结果比较

Table 5 Comparison of segmentation results of different models on GID dataset

Method	IoU / %					MIoU / %	OA / %	F1 / %
	Building	Water	Forest	Farmland	Meadow			
FCN	47.03	72.50	46.72	49.11	43.08	51.69	70.85	65.11
U-Net	62.21	81.22	51.07	68.79	64.08	65.47	82.80	69.05
DeeplabV3+	63.22	79.49	53.23	67.53	60.93	64.88	80.66	67.14
TransUNet	63.03	80.35	55.02	68.93	64.24	66.31	82.91	69.12
SRAU-Net	64.58	80.82	55.73	70.11	65.52	67.35	84.26	71.84

GID 大比例尺数据集分割结果可视化如图 7 所示。从第 1 组图像可以发现,建筑和农田边缘较难识别区分,其他方法处理的结果出现地物边缘不清且建筑群内部易空洞的现象,SRAU-Net 分割结果更为准确。在第 2 组图像和第 4 组图像中,其他方法不能准确地区分大面积相似度高的林地、农田和草地,SRAU-Net 从全局信息和局部特征中聚合出更多有效特征,可以获得相对准确的分割结果。在第 3 组图像中,

SRAU-Net 更好地识别了水体中的其他类别和大规模建筑的边缘。

4.2 模型消融实验结果分析

为了评估 SRAU-Net 和 3 个模块的性能,通过逐步添加模块,在 Vaihingen 数据集上进行了综合消融实验,详细实验设置如表 6 所示。具体实验结果如表 7 所示。模型(a)表示 Swin Transformer 模块组成的 U-Net,作为基线模型(baseline)。模型(b)表示在模

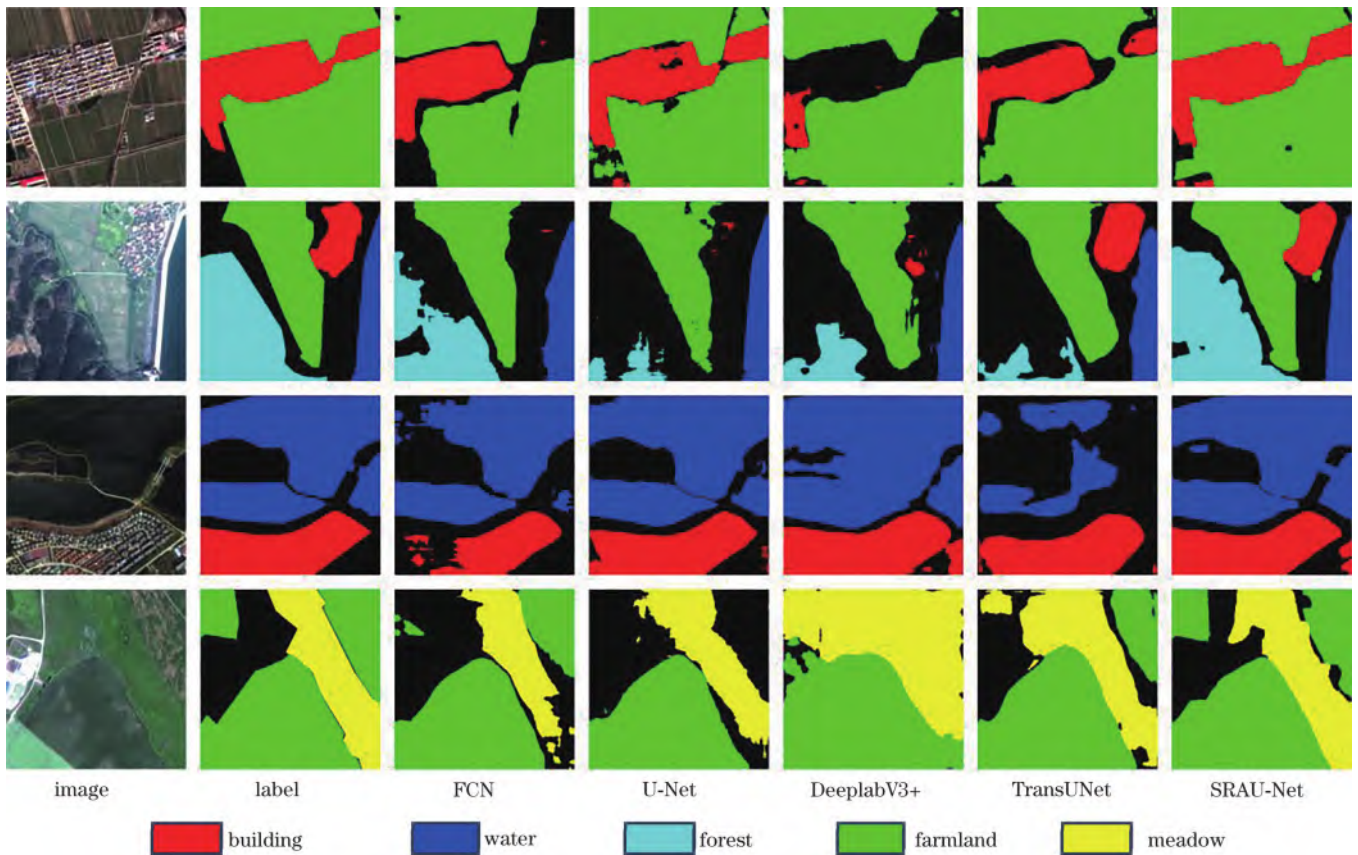


图 7 不同模型在 GID 数据集上的分割结果可视化

Fig. 7 Visualization of segmentation results of different models on GID dataset

表 6 SRAU-Net 消融实验设置

Table 6 Ablation experiment setting of SRAU-Net

Model name	Baseline	Residual block	FFM	FEM
(a)	✓			
(b)	✓	✓		
(c)	✓	✓	✓	
(d)	✓	✓		✓
(e)	✓	✓	✓	✓

型(a)的基础上添加残差模块组成的辅助编码器,添加了多尺度空间信息,与模型(a)相比 MIoU 提升了 1.91 百分点。模型(c)在模型(b)的基础上添加了 FFM, FFM 对主编码器与辅助编码器提取到的全局信息和局部空间特征进行有效融合,与没有特征融合模块的模型(b)相比 MIoU 增加了 0.18 百分点。当在 SRAU-

Net 框架中考虑 FFM 时,对建筑类别的分割精度提高最大, IoU 提高了 0.73 百分点,其次是不透水表面类别,提高了 0.12 百分点。模型(d)在模型(b)的基础上添加了 FEM, 验证 FEM 的有效性,结果表明添加 FEM 后与模型(b)相比, MIoU 增加了 0.22 百分点。模型(e)进一步添加了 FFM 和 FEM, 为最终完整的 SRAU-Net 模型,从表 7 可以看出, FEM 对 Vaihingen 数据集中的汽车类别影响最为明显。以上结果和分析证明了所提模块是有效的。

所提 SRAU-Net 模型不同模块在 Vaihingen 数据集上的可视化结果如图 8 所示。从图 8 可以看出,添加残差模块的辅助编码器捕获了远程依赖关系的同时添加局部空间信息,使对象的边缘变得更加清晰,如第 1 幅图像中左下角低矮植被的边缘。模型(c)将 FFM 应用

表 7 在 Vaihingen 数据集上的消融实验结果

Table 7 Result of the ablation experiment on Vaihingen dataset

Method	IoU / %					MIoU / %	OA / %	F1 / %
	Impervious surface	Building	Low vegetation	Tree	Car			
(a)	82.66	89.37	70.07	77.03	75.12	78.85	90.03	85.55
(b)	83.68	90.09	73.02	80.67	76.33	80.76	91.25	86.38
(c)	83.80	90.82	73.05	80.69	76.36	80.94	92.35	86.62
(d)	83.78	90.87	73.14	80.72	76.43	80.98	92.37	86.68
(e)	83.88	90.97	73.22	80.95	77.71	81.35	92.60	86.90

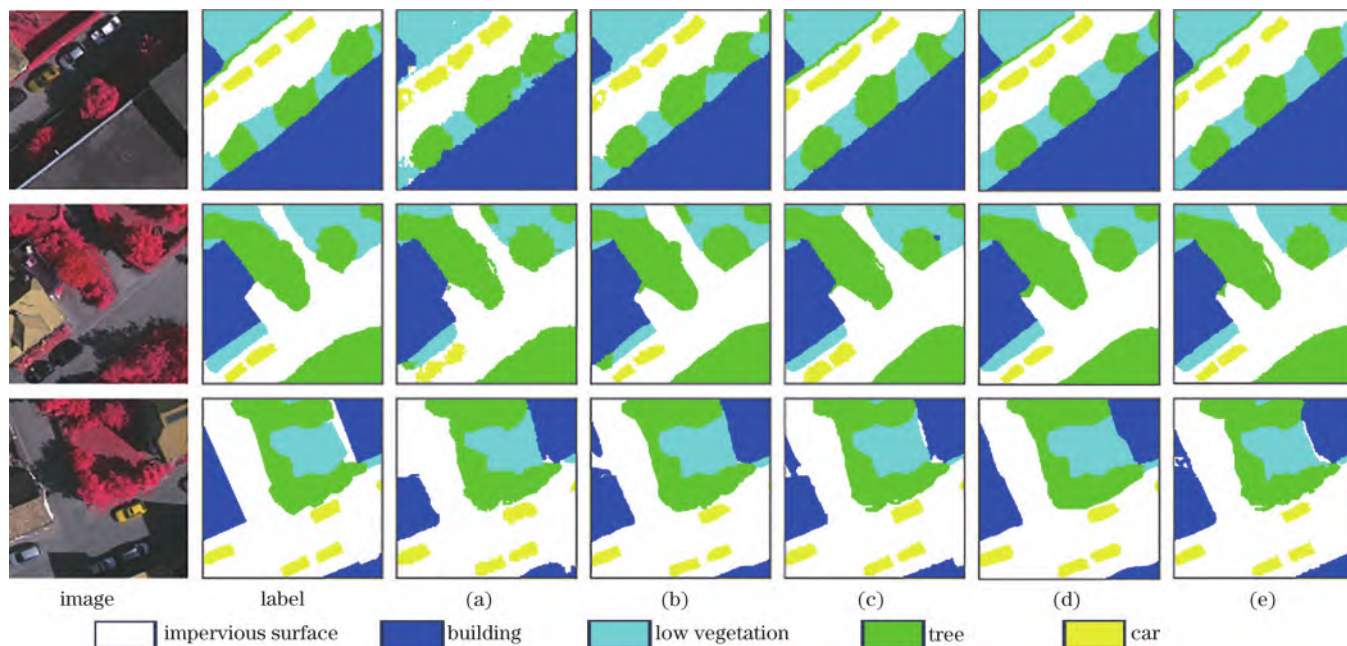


图8 消融实验结果可视化。(a) baseline; (b) baseline+residual block; (c) baseline+residual block+FFM; (d) baseline+residual block+FEM; (e) baseline+residual block+FFM+FEM

Fig. 8 Visualization of the ablation experiment results. (a) baseline; (b) baseline+residual block; (c) baseline+residual block+FFM; (d) baseline+residual block+FEM; (e) baseline+residual block+FFM+FEM

于模型(b)中关注较小的目标,消除分类错误的地物,如第3幅图像中最左侧的小面积建筑,表明FFM聚合有效信息后能够提高对地面小对象分类的准确性。模型(d)进一步添加FEM,进一步提升对细节特征的增强效果,从第2幅图像上方可以看出由于树木阴影的遮挡,物体被错误归类为树木,增加FEM后得到正确分类。因此,根据上述结果与分析,表明SRAU-Net有较好的遥感地物分割效果。

5 结 论

针对现有遥感影像分类方法主流深度学习网络模型存在的小尺度目标错分漏分、相似地物易错分、地物边界较模糊等问题,提出了一种具有Swin Transformer和残差模块双分支编码器的遥感图像语义分割网络模型SRAU-Net。该模型的关键是将残差模块辅助编码器的空间特征融合到Swin Transformer的主编码器路径中,用局部空间特征补充全局特征。针对图块内部的空间信息损失问题,设计了具有图块注意的特征融合模块,自适应地融合不同尺度的特征。此外,在解码器部分还构造了一个特征增强模块将通过跳跃连接获得的编码器特征与解码器上采样特征有效融合。在ISPRS Vaihingen数据集上对SRAU-Net进行了验证,并与同类模型进行比较,研究结果表明,SRAU-Net的总体分类精度较原始模型提升了2.57百分点,并且对小目标地物和地物边缘的分类效果获得明显提升,优势如下。

1) 建立了一种具有Swin Transformer和残差模块

双分支编码器的SRAU-Net模型。通过同时获得粗粒度分辨率和细粒度分辨率特征信息,SRAU-Net能够有效地收集全局上下文和详细空间信息。此外,在解码器的上采样特征图中加入来自编码器的信息后,SRAU-Net可以恢复更清晰的边缘图,获得令人满意的分割结果。

2) 构建了特征融合模块(FFM),以关注空间和通道维度上的像素级特征相关性,以缓解下采样过程中小尺度目标特征遗漏缺失的问题,收集更多关于小尺度对象的特性,并减少详细信息的丢失。结果表明,加入FFM后模型与基础模型相比精度有明显提升,MIoU与OA分别提升了2.09百分点和2.32百分点。

3) 在解码器中提出了一个特征增强模块(FEM),减少地面物体遮挡造成的语义模糊,可以有效地区分相似性较高的地面物体。结果表明,在基础模型中依次加入FFM、FEM后,得到的SRAU-Net模型可有效处理ISPRS Vaihingen数据集,MIoU、OA和F1三项指标分别较未添加FEM的模型提高了0.41百分点、0.25百分点和0.28百分点。

基于Swin Transformer和卷积神经网络提出的遥感图像语义分割网络模型改进效果良好,可为遥感大数据时代土地资源监测等方面提供支撑。但SRAU-Net仍存在着一定不足:采用全监督的方式进行网络训练,训练过程中需要大量的标签数据,获取高质量的标签数据通常需消耗大量的人力物力。遥感图像包括多模态数据,如DSM、DEM和NDVI等,这些数据包含大量可用于遥感智能解译的先验知识,本文未利用

多模态数据辅助模型分割影像。在未来的研究中可利用半监督或弱监督的方法减少对样本数据的依赖,同时充分利用多模态遥感数据提高遥感影像分割性能。此外,将进一步考虑减少参数量,构建一个更轻量级的模型,加快模型训练速度。

参 考 文 献

- [1] di Gregorio A. Land cover classification system: classification concepts and user manual: LCCS[EB/OL]. [2023-05-06]. <https://www.fao.org/3/y7220e/y7220e00.htm>.
- [2] 刘纪远, 张增祥, 张树文, 等. 中国土地利用变化遥感研究的回顾与展望: 基于陈述彭学术思想的引领[J]. 地球信息科学学报, 2020, 22(4): 680-687.
Liu J Y, Zhang Z X, Zhang S W, et al. Innovation and development of remote sensing-based land use change studies based on Shupeng Chen's academic thoughts[J]. Journal of Geo-Information Science, 2020, 22(4): 680-687.
- [3] 舒弥, 杜世宏. 国土调查遥感 40 年进展与挑战[J]. 地球信息科学学报, 2022, 24(4): 597-616.
Shu M, Du S H. Forty years' progress and challenges of remote sensing in national land survey[J]. Journal of Geo-Information Science, 2022, 24(4): 597-616.
- [4] 高练, 周勇. 武汉市土地利用/土地覆盖变化的生态环境效应分析[J]. 农业工程学报, 2008, 24(S1): 73-77.
Gao L, Zhou Y. Environmental effect of land use and land cover change in Wuhan City[J]. Transactions of the Chinese Society of Agricultural Engineering, 2008, 24(S1): 73-77.
- [5] 陈霆, 徐伟铭, 吴升, 等. 国土空间规划视角下的城镇开发边界划定和空间管控体系构建[J]. 地球信息科学学报, 2022, 24(2): 263-279.
Chen T, Xu W M, Wu S, et al. Delimitation of urban development boundary and construction of space control system from the perspective of territorial spatial planning [J]. Journal of Geo-Information Science, 2022, 24(2): 263-279.
- [6] Bi H X, Xu F, Wei Z Q, et al. An active deep learning approach for minimally-supervised PolSAR image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(11): 9378-9395.
- [7] 张继贤, 顾海燕, 杨懿, 等. 自然资源要素智能解译研究进展与方向[J]. 测绘学报, 2022, 51(7): 1606-1617.
Zhang J X, Gu H Y, Yang Y, et al. Research progress and trend of intelligent interpretation for natural resources features[J]. Acta Geodaetica et Cartographica Sinica, 2022, 51(7): 1606-1617.
- [8] 黄冬青, 徐伟铭, 许文迪, 等. 基于 DeeplabV³⁺ 网络的高分遥感影像分类[J]. 激光与光电子学进展, 2023, 60(16): 1628001.
Huang D Q, Xu W M, Xu W D, et al. High-resolution remote sensing image classification based on DeeplabV³⁺ network[J]. Laser & Optoelectronics Progress, 2023, 60(16): 1628001.
- [9] Zhu Q Q, Lei Y, Sun X L, et al. Knowledge-guided land pattern depiction for urban land use mapping: a case study of Chinese cities[J]. Remote Sensing of Environment, 2022, 272: 112916.
- [10] Fan Z Y, Zhan T, Gao Z C, et al. Land cover classification of resources survey remote sensing images based on segmentation model[J]. IEEE Access, 2022, 10: 56267-56281.
- [11] 杨潇宇, 汪西莉. 结合多尺度注意力和边缘监督的遥感图像建筑物分割模型[J]. 激光与光电子学进展, 2022, 59(22): 2228004.
Yang X Y, Wang X L. Building segmentation model of remote sensing image combining multiscale attention and edge supervision[J]. Laser & Optoelectronics Progress, 2022, 59(22): 2228004.
- [12] Maggiori E, Tarabalka Y, Charpiat G, et al. Convolutional neural networks for large-scale remote-sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(2): 645-657.
- [13] Wu X, Hong D F, Chanussot J. Convolutional neural networks for multimodal remote sensing data classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5517010.
- [14] 李宏达, 高小红, 汤敏. 基于 CNN 的不同空间分辨率影像土地覆被分类研究[J]. 遥感技术与应用, 2020, 35(4): 749-758.
Li H D, Gao X H, Tang M. Land cover classification for different spatial resolution images from CNN[J]. Remote Sensing Technology and Application, 2020, 35(4): 749-758.
- [15] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(4): 640-651.
- [16] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.
- [17] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08) [2023-05-04]. <https://arxiv.org/abs/2102.04306>.
- [18] 梁正印, 汪西莉. 基于波段-位置自适应选择的多光谱遥感图像语义分割[J]. 激光与光电子学进展, 2023, 60(14): 1410016.
Liang Z Y, Wang X L. Semantic segmentation of multispectral remote sensing images based on band-location adaptive selection[J]. Laser & Optoelectronics Progress, 2023, 60(14): 1410016.
- [19] 菅永胜, 朱大明, 付志涛, 等. 多层次特征优化融合的遥感图像分割网络[J]. 激光与光电子学进展, 2023, 60(4): 0428002.
Jian Y S, Zhu D M, Fu Z T, et al. Remote sensing image segmentation network based on multi-level feature refinement and fusion[J]. Laser & Optoelectronics Progress, 2023, 60(4): 0428002.
- [20] Li R, Zheng S Y, Duan C X, et al. Multistage attention

- ResU-net for semantic segmentation of fine-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 8009205.
- [21] 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.
- [22] Mou L C, Hua Y S, Zhu X X. Relation matters: relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(11): 7557-7569.
- [23] Zheng S X, Lu J C, Zhao H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 6877-6886.
- [24] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-05-06]. <https://arxiv.org/abs/2010.11929>.
- [25] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.
- [26] Lin A L, Chen B Z, Xu J Y, et al. DS-TransUNet: dual Swin Transformer U-net for medical image segmentation [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 4005615.
- [27] Cao H, Wang Y Y, Chen J, et al. Swin-unet: unet-like pure transformer for medical image segmentation[M]//Karlinsky L, Michaeli T, Nishino K. Computer vision-ECCV 2022 workshops. Cham: Springer, 2023, 13803: 205-218.
- [28] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.
- [29] Shafiq M, Gu Z Q. Deep residual learning for image recognition: a survey[J]. Applied Sciences, 2022, 12(18): 8972.
- [30] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.
- [31] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 3141-3149.
- [32] 吴新辉, 毛政元, 翁谦, 等. 利用基于残差多注意力和 ACON 激活函数的神经网络提取建筑物[J]. 地球信息科学学报, 2022, 24(4): 792-801.
- Wu X H, Mao Z Y, Weng Q, et al. A neural network based on residual multi-attention and ACON activation function for extract buildings[J]. Journal of Geo-Information Science, 2022, 24(4): 792-801.
- [33] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 833-851.