计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009841] [CSTR: 32024.14.csa.009841]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 U-BiFormer 的遥感图像地表分类模型①

安立君, 刘向阳

(河海大学 数学学院, 南京 211100)
通信作者: 刘向阳, E-mail: liuxy@hhu.edu.cn

摘　要: 对遥感图像进行地表分类对于城市规划、土地利用、环境监测和地表温度反演等工作而言十分重要. 针对相似地表类别存在误检的问题以及遥感图像地表类别不均衡的问题, 本文提出了一种 U 型 Transformer 模型 U-BiFormer, 该模型在 BiFormer 的基础上使用 U 型解码器, 使用所有阶段解码器的输出来预测分割图, 提高了模型捕捉图像中的细节和上下文信息的能力, 使模型能更好分割相似类别. 对 U 型解码器特有的混合注意力模块进行改进, 增大当前阶段特征在混合特征中所占的比例, 让解码器更注重对当前阶段特征的细化, 提升模型对相似类别的分割效果. 使用 CE+Focal 混合损失函数替代常规交叉熵损失函数, 应对遥感图像地表类别分布不均的问题. 实验证明在 GID 大型遥感图像数据集上本文方法能更好地分割相似类别, 并且取得了优于当前主流模型的分割结果 (*Acc* (81.99%) 和 *mIoU* (71.04%)).

关键词: 遥感图像; 深度学习; 语义分割; 地表分类; BiFormer

引用格式: 安立君,刘向阳.基于 U-BiFormer 的遥感图像地表分类模型.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9841.html

U-BiFormer-based Model for Land Cover Classification in Remote Sensing Image

AN Li-Jun, LIU Xiang-Yang

(School of Mathematics, Hohai University, Nanjing 211100, China)

Abstract: Land cover classification of remote sensing images is crucial for urban planning, land use, environmental monitoring, and land cover temperature inversion. This study proposes a U-type Transformer network, U-BiFormer to address the issues of misclassification among similar land cover types and the imbalance of land cover classes in remote sensing images. Building upon BiFormer, this model employs a U-shaped decoder and uses the outputs of the decoders in all stages to predict the segmentation map, thereby enhancing the model's ability to capture details and contextual information in images, allowing for better segmentation of similar classes. An improvement is made to the unique hybrid attention module of the U-shaped decoder, increasing the proportion of features from the current stage in the mixed features. This modification enables the decoder to focus more on refining the features at the current stage, enhancing the model's segmentation performance for similar classes. Additionally, the CE+Focal hybrid loss function is employed to replace the conventional cross-entropy loss function to address the issue of class distribution imbalance in remote sensing images. Experiments demonstrate that the proposed method achieves better segmentation results for similar classes on the GID large-scale remote sensing image dataset, outperforming current mainstream models with an accuracy (*Acc*) of 81.99% and a mean intersection over union (*mIoU*) of 71.04%.

Key words: remote sensing image; deep learning; semantic segmentation; land cover classification; BiFormer

① 收稿时间: 2024-10-16; 修改时间: 2024-10-30; 采用时间: 2024-12-04; csa 在线出版时间: 2025-03-24

遥感图像的地表分类是图像处理领域的一个关键研究方向, 对于城市规划、土地利用、环境监测以及地表温度反演等方面具有重要意义[1]. 随着观测比例的增大, 特征提取的难度急剧增加. 传统人工提取特征的方法已经被深度学习的方法所替代.

近些年来基于深度学习的语义分割算法在地表分类领域取得了显著成果. Pan 等[2]专注于使用优化超参数的 CNN 对多光谱激光雷达数据进行土地覆盖分类. Pradhan 等[3]探索了一种基于对象的多尺度卷积神经网络, 用于高空间分辨率遥感图像分类. Fan 等[4]提出了一种半监督多卷积神经网络集成学习方法, 用于使用亚米级高分辨率遥感 (HRRS) 图像进行城市土地覆盖分类, 并结合从未标记数据中自动选择样本的功能. Bui 等[5]研究梯度提升算法 (XGBoost、LightGBM、Catboost) 与 CNN 结合以用于土地覆盖分类, 通过基于对象的图像分析提高了准确性. Horry 等[6]通过提出一种基于视觉变换器的分类方法, 解决了深度学习方法在极化合成孔径雷达 (SAR) 图像土地覆盖分类任务中的局限性. 模型的小感受野和标注样本稀缺是实现最佳性能的挑战. Yao 等[7]扩展了视觉变换器模型用于土地利用和土地覆盖分类, 强调了基于注意机制的深度模型在更广泛领域中的适应性. Weng 等[8]引入了一种局部和全局特征耦合网络 SGformer 用于土地覆盖的语义分割, 在实验中表现出较高的准确性. Lu 等[9]提出了一种变换器-卷积神经网络混合架构, 用于无人机遥感影像的土地覆盖分类, 强调了其特征提取方法在生成高精度土地覆盖地图中的重要性.

尽管上述工作在遥感图像的地表分类任务中取得了良好的效果, 但都未能有效利用不同阶段特征进行分割, 且在颜色相似的类别中仍存在误检的问题. 针对此问题, 本文提出一种基于 U 型结构 BiFormer 的地表分类方法. 该方法基于 BiFormer[10]语义分割模型, 使用 U 型结构的轻量级解码器, 并对其特有的混合注意力模块进行改进, 同时采用 CE+Focal 作为损失函数. 其中 U 型结构的轻量级解码器和改进后的混合注意力模块, 能使模型有效利用不同阶段特征进行分割, 且解码过程中能够查询不同阶段的特征. 从而改善模型在相似类别区域的误检现象, 提高分割结果.

1 U-BiFormer 地表分类模型

1.1 BiFormer

BiFormer 使用四级金字塔结构. 在第 1 阶段使用重叠补丁嵌入, 在第 2–4 阶段使用补丁合并模块来降低输入空间分辨率, 同时增加通道数量, 每个阶段使用多个连续的 BiFormer 块来提取特征. BiFormer 块的结构如图 1 所示.

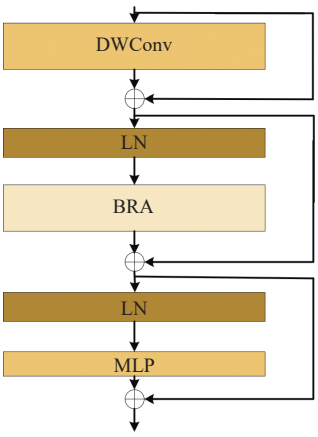

图 1 BiFormer 块结构图

其中每个 BiFormer 块一开始都使用 3×3 深度卷积来隐式地编码相对位置信息, 然后依次应用 BRA (bi-level routing attention) 模块和 MLP 模块. 其中 BRA 模块使用一种双层路由动态稀疏注意力机制过滤不关键的键值对, 然后在路由区域应用精细化的注意力机制. 这样能够使模型重点关注有效信息, 减少无用信息带来的影响, 更好地提取低层次特征与高层次特征, 并减少计算量.

1.2 U-BiFormer 网络结构

针对相似类别中存在误检的问题, 本文提出基于 U-BiFormer 的地表分类模型, 其在 BiFormer 中使用 U 型结构的轻量级解码器[11], 并对 U 型解码器特有的混合注意力模块进行改进, 网络结构如图 2 所示.

本文模型结构与 U-Net[12]的结构类似, 但与其他类似于 U-Net 的变体之间有两点主要区别. 本文模型使用横向连接作为查询的特征, 解码器进行特征细化时隐式地增加了空间分辨率, 不需要在解码器阶段之间进行显式上采样. 使用所有阶段的解码器输出来预测分割图, 而不仅是使用最后阶段的解码器输出. 最后一个阶段的解码器输出的特征图, 空间分辨率是 $H/4\times W/4$, 后面会单独恢复到原始空间分辨率 $H\times W$.

1.3 改进后的混合注意模块

在 Transformer 块中使用的注意力模块计算查询 Q、键 K 和值 V 的点积注意力公式如下:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{1}$$

其中, d_k 表示嵌入维度.

混合注意力模块的核心是选择生成键和值的特征, 在自注意力中用于生成查询、键和值的特征是相同的解码器阶段. 交叉注意力使用了两个不同的特征, 每个特征都来自相同的编码器/解码器阶段. 相比之下, 混合注意力利用了来自多个多尺度阶段的特征. 这允许查询在所有不同的阶段的特征中匹配键和值, 从而促进增强的特征细化, 提升特征质量.

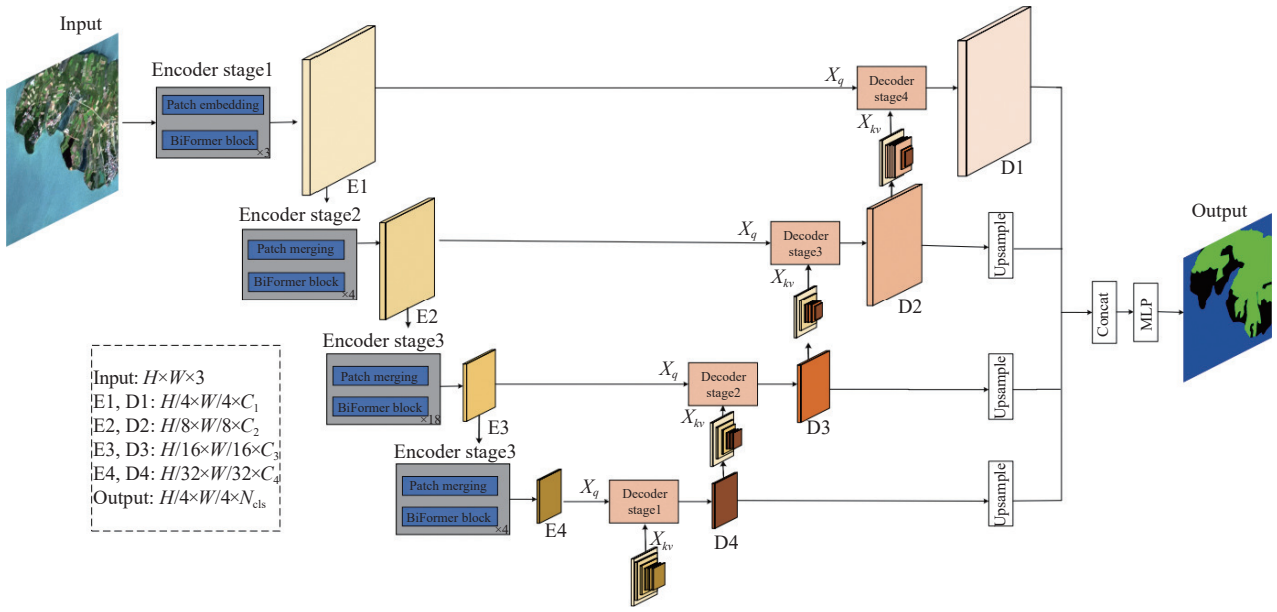

图 2 U-BiFormer 的网络结构图

虽然混合后的特征图有着多尺度阶段的混合特征, 但经过解码后一定程度降低了对当前尺度阶段特征的关注, 也会因此影响对当前尺度阶段特征的细化. 针对此问题, 本文对混合注意力模块进行改进, 将原混合特征比例 1:1:1:1 中的当前尺度阶段特征比例调整为 K, 目的是让混合注意力模块在引入多尺度特征的同时更加关注当前阶段的特征, 从而进一步增强特征细化, 该方法的有效性在消融实验部分进行了验证. 结果表明 K=3 时, 模型效果最优.

1.4 解码器结构

本文中的解码器没有使用传统的 Transformer 解码器中的自注意力与交叉注意力. 而是使用改进后的混合注意力进行替代, 其结构如图 3 所示.

图 3 中 $MixAtt_K$ 表示本文改进后的混合注意力模块、LN 表示归一化层、FFN 表示前馈网络, 解码器输出的计算公式如下:

$$A_i=LN(MixAtt_K(LN(X_{kv}^i,X_q^i))+LN(X_q^i)) \tag{2}$$

$$D_{N-i+1}=FFN(A_i)+A_i \tag{3}$$

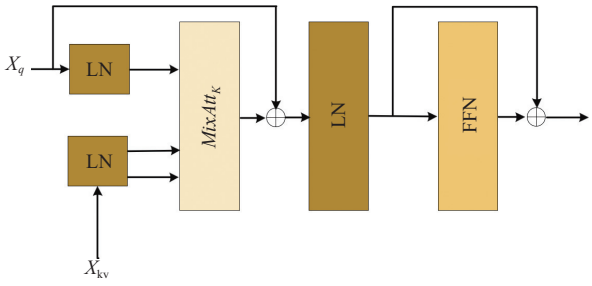

图 3 解码器结构图

1.5 损失函数

遥感图像地表类别分布往往是不均匀的, 比如本文所使用的数据集, 林地类别占比不足 5%. 为解决遥感图像地表类别分布不均的问题, 本文使用 CE+Focal[13] 混合损失函数, 其定义如下:

$$Loss=Loss_{\mathrm{CE}}+Loss_{\mathrm{Focal}} \tag{4}$$

其中, 交叉熵损失函数与焦点损失函数在多类别任务中的单个像素点计算公式分别为式 (5)–式 (7):

$$Loss_{\mathrm{CE}}=-\sum_{i=1}^{C}y_i\log(p_i) \tag{5}$$

$$Loss_{\text{Focal}} = \sum_{i=1}^{C} FL(y_i, p_i) \tag{6}$$

$$FL(y_i, p_i) = \begin{cases} -\alpha(1-p_i)^{\gamma}\log(p_i), & \text{if } y_i = 1 \\ -(1-\alpha)p_i^{\gamma}\log(1-p_i), & \text{otherwise} \end{cases} \tag{7}$$

其中, C 是地表类别个数, y_i表示预测样本的真实标签, p_i表示模型预测像素属于第 i 个类别的概率, $FL(y_i, p_i)$表示第 i 个类别的焦点损失, α是一个平衡因子用于调整正类别和负类别之间的权重, γ是一个调节参数用于控制难易样本的权重.

2 实验分析

2.1 实验数据集

本文数据集是基于遥感图像大规模分类集 (GID-5)[14–16]构建的. 大规模分类集 (GID-5) 包含建筑、农田、林地、草地和水域 5 个土地覆盖类别, 共计 150 幅像素级标注的高分辨遥感图像. 实验中将遥感图像裁切成 512×512 分辨率的图像后随机选取 20 000 张图像按照 8:1:1 进行数据集划分.

2.2 实验设置

本文实验在 NVIDIA A10 (24 GB 显存) GPU 上使用 PyTorch 深度学习框架来实现的,代码语言及版本为 Python 3.7. 实验中使用 AdamW 优化器, 初始学习率设置为 0.000 5, 采用 Poly 衰减策略, 权重衰减 (weight decay) 设置为 0.001, 批量大小 (batch size) 设置为 4, 样本迭代次数设置为 80 000 次.

2.3 评价指标

遥感图像的地表分类任务是一个像素级多分类任务, 本文选择交并比 (*IoU*)、像素准确率 (*Acc*) 平均交并比 (*mIoU*)、平均像素准确率 (*mAcc*) 和平均误检率 (*mFPR*) 作为评价指标来衡量网络的性能[17]. *IoU* 表示单一类别预测值和真实值两个集合的交集与并集之比; *Acc* 表示所有像素准确率. *mIoU* 表示所有类别预测值和真实值两个集合的交集与并集之比的平均值; *mAcc* 表示所有类别的像素准确率之和的平均值; *mFPR* 表示所有类别的像素误检率之和的平均值. 具体的计算公式为式 (8)–式 (12):

$$IoU = \frac{TP}{TP+FP+FN} \tag{8}$$

$$Acc = \frac{TP}{TP+FP} \tag{9}$$

$$mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU \tag{10}$$

$$mAcc = \frac{1}{C}\sum_{i=1}^{C} \frac{TP}{TP+FP} \tag{11}$$

$$mFPR = \frac{1}{C}\sum_{i=1}^{C} \frac{FP}{FP+TN} \tag{12}$$

其中, *TP* 表示正确预测为该类别的像素; *FP* 表示将其他类别错误预测为该类别的像素; *FN* 表示将该类别错误预测为其他类别的像素; *TN* 表示正确预测为其他类别的像素.

2.4 改进后的混合注意力 $MixAtt_K$ 中 K 值的选取

本文中对混合注意力 MixAtt 进行改进, 增大当前特征在混合时的比例, 使其在混合多阶段多尺度特征的同时更加注重于当前阶段的特征. 为了确定最佳混合比例 *K*, 本文进行实验, 结果如表 1 所示.

表 1 不同 *K* 值的实验结果 (%)

K	*Acc*	*mAcc*	*mIoU*	*mFPR*
1	81.62	82.71	70.71	16.88
2	81.78	82.96	70.84	16.71
3	81.99	82.99	71.04	16.64
4	81.60	82.90	70.58	16.89
5	81.54	82.74	70.54	16.95

选取混合比例 *K*=3 时, 能够使模型在关注多尺度特征和更注重当前阶段特征之间达到平衡, 模型的 *Acc*、*mAcc*、*mIoU* 分别提升了 0.37%、0.28%、0.33%, 并且 *mFPR* 降低了 0.24%, 证明了此改进的有效性.

2.5 对比实验

为了验证改进后模型的有效性, 本文在同样的数据集上对比了近年提出的经典分割模型, 如 U-Net[12]、SegFormer[18]、SegNeXt[19]和 Swin Transformer[20], 实验结果如表 2 所示.

表 2 不同模型地表分类实验结果 (%)

方法	*Acc*	*mAcc*	*mIoU*	*mFPR*
U-Net	79.45	79.41	65.83	20.35
SegFormer	79.81	80.63	68.14	19.79
SegNeXt	77.65	65.83	55.15	20.92
Swin Transformer	75.46	72.61	59.88	22.25
BiFormer	76.46	78.69	64.56	21.02
U-BiFormer	81.99	82.99	71.04	16.64

从表 2 中可以看到, 本文改进后的模型表现最佳, *Acc*、*mAcc*、*mIoU* 均优于两种经典卷积模型 U-Net 和 SegNeXt, 同时也均优于两种经典 Transformer 模型

SegFormer 和 Swin Transformer. 本文模型对比表现最好的 SegFormer 模型, *Acc*、*mAcc*、*mIoU* 这 3 个指标上分别提升了 2.18%、2.36%、2.90%, 并且比未进行改进的 BiFormer 模型分别提升了 5.53%、4.30%、6.48%, 同时也有着最低的误检率. 这表明本文改进后的模型在地表分类任务中是有效的.

为了更明显地对比其他方法与本文方法对遥感图像地表分类任务的分割效果, 选取部分实验结果进行对比分析, 结果如图 4 所示. 从第 1 个示例中可以看出本文方法对比除 SegFormer 之外的方法明显减少了对草地的误检, 而且建筑用地区域边缘更加清晰. SegFormer 将部分草地误检为建筑, 而本文方法不存在这种误检. 从第 2 个示例中可以看出本文方法比其他方法能更精准地分割草地与农田, 且比 U-Net, SegNeXt 和 Swin Transformer 更好的保持水域的连通性. 从第 3 个示例可以看出 BiFormer、U-Net 和 Swin Transformer 将大片农田误检为草地, Swin Transformer 更是在草地区域误检出建筑, 而本文方法比其他方法分割边缘更加准确且误检区域更小. 因此证明了本文在 BiFormer 的基础上使用 U 型解码器并且对混合注意力模块进行改进后的模型能对像素值分布相近的类别和不同类别的边缘区域进行更精确地分割, 减少误检.

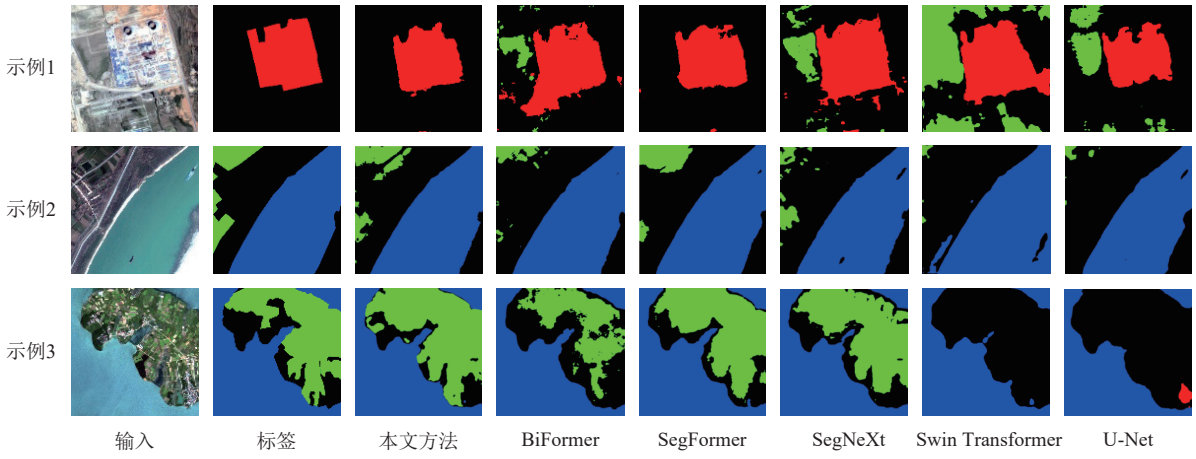

图 4 不同模型地表分类结果

2.6 消融实验

为了验证本文对 BiFormor 网络改进的有效性, 本文分别在 U 型解码器、混合注意力 *MixAtt* 及改进后的混合注意力 $MixAtt_K$ 上进行消融对比实验, 得到的实验结果如表 3 所示.

表 3 消融实验结果 (%)

方法	*Acc*	*mIoU*	建筑	农田	草地	林地	水域	*mFPR*
BiFormer	76.46	64.56	60.77	60.56	57.53	58.41	85.53	21.02
+U型结构	81.26	70.01	64.23	68.31	64.65	64.92	88.37	17.65
+U型结构+MixAtt	81.62	70.71	64.60	68.75	65.22	66.39	88.61	16.88
+U型结构+$MixAtt_K$	81.99	71.04	64.69	69.22	65.74	66.88	88.69	16.64

选择 BiFormer 网络作为基线模型, 其 *Acc*、*mIoU*、*mFPR* 分别为 76.46%、64.56%、21.02%. 使用 U 型解码器后, 其 *Acc* 与 *mIoU* 分别提升至 81.26%、70.01%, *mFPR* 降低至 17.65%. 在使用 U 型解码器的基础上使用混合注意力 MixAtt 后, 其 *Acc* 与 *mIoU* 分别提升至 81.62%、70.71%, *mFPR* 降低至 16.88%. 在使用 U 型解码器的基础上使用改进的混合注意力 $MixAtt_K$ 后, 其 *Acc* 与 *mIoU* 分别提升至 81.99%、71.04%, *mFPR* 降低至 16.64%. 从评价指标 *Acc*、*mIoU*、*mFPR* 可以看出, 本文提出的改进是有效的.

遥感图像中农田、林地与草地这 3 个类别像素分布比较相似分割难度大. 从基线模型对 5 个地表覆盖类别的分类评价指标来看, 建筑、农田、林地和草地的 *IoU* 明显低于水域的 *IoU*. 使用 U 型解码器后, 5 个地表分覆盖类别的 *IoU* 都有明显提高, 其中农田、林地与草地这 3 个类别提升最为明显. 在使用 U 型解码器的基础上使用混合注意力 MixAtt 后, 5 个地表分覆盖类别的 *IoU* 都有一定程度提高, 其中农田、林地与草地这 3 个类别提升较大. 在使用 U 型解码器的基础上使用改进的混合注意力 $MixAtt_K$ 后, 建筑与水域的 *IoU* 只有小于 0.10% 的提高, 农田、林地与草地这 3 个类别的 *IoU* 均提升 0.50% 以上. 综上, 证明了本文为应对相似类别 (农田、林地、草地) 存在误检的问

题, 对 BiFormer 模型的改进是有效的.

为了验证本文所使用损失函数的有效性, 分别使用不同的损失函数在数据集上进行消融实验, 得到的实验结果如表 4 所示.

表 4 不同损失函数的消融实验结果

损失函数		评价指标 (%)		
CE	Focal	*Acc*	*mAcc*	*mIoU*
√	×	80.77	82.07	69.72
×	√	81.62	82.70	70.60
√	√	81.99	82.99	71.04

从表 4 中可以看出使用 CE+Focal 作为损失函数实验效果最好, 评价指标 *Acc*、*mAcc* 和 *mIoU* 均高于单独使用 CE 或 Focal 作为损失函数的模型. 其中对比单独使用 CE 作为损失函数的模型, 评价指标 *Acc*、*mAcc* 和 *mIoU* 分别提高了 1.22%、0.92% 和 1.32%. 证明了本文使用 CE+Focal 作为损失函数的有效性.

3 总结与展望

本文提出了一种基于 U-BiFormer 的语义分割模型来对遥感图像进行地表分类. 该方法在 BiFormer 的基础上使用类似 U 型结构的解码器, 对解码器原本的混合注意力模块进行改进, 使其在混合不同阶段不同尺度特征的同时更加注意当前阶段的特征, 有效利用了不同阶段的特征进行预测. 使用 CE+Focal 作为损失函数, 提升了模型在地表类别不均匀数据集上的分割性能. 通过消融实验和对比实验, 证实了改进后的模型能够提升各个地表类别的精度, 并显著减少了图像中相似类别区域的误检. 后续会考虑在本文基础上进一步对不同解码器进行研究, 让模型能更好地解决地表分类任务.

参考文献

1 朱凡, 罗小波. 改进的 DeeplabV3Plus 高分辨率遥感影像土地覆盖分类. 计算机工程与应用, 2024, 60(13): 266–275. [doi: 10.3778/j.issn.1002-8331.2309-0228]

2 Pan SY, Guan HY, Chen YT, *et al*. Land-cover classification of multispectral LiDAR data using CNN with optimized hyper-parameters. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 241–254. [doi: 10.1016/j.isprsjprs.2020.05.022]

3 Pradhan B, Al-Najjar HAH, Sameen MI, *et al*. Unseen land cover classification from high-resolution orthophotos using integration of zero-shot learning and convolutional neural networks. Remote Sensing, 2020, 12(10): 1676. [doi: 10.3390/rs12101676]

4 Fan RY, Feng RY, Wang LZ, *et al*. Semi-MCNN: A semisupervised multi-CNN ensemble learning method for urban land cover classification using submeter HRRS images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 4973–4987. [doi: 10.1109/JSTARS.2020.3019410]

5 Bui QT, Chou TY, Hoang TV, *et al*. Gradient boosting machine and object-based CNN for land cover classification. Remote Sensing, 2021, 13(14): 2709. [doi: 10.3390/rs13142709]

6 Horry MJ, Chakraborty S, Pradhan B, *et al*. 2-speed network ensemble for efficient classification of incremental land-use/land-cover satellite image chips. arXiv:2203.08267, 2022.

7 Yao J, Zhang B, Li CY, *et al*. Extended vision Transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5514415.

8 Weng LG, Pang K, Xia M, *et al*. SGFormer: A local and global features coupling network for semantic segmentation of land cover. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 6812–6824. [doi: 10.1109/JSTARS.2023.3295729]

9 Lu TY, Wan LH, Qi SQ, *et al*. Land cover classification of UAV remote sensing based on Transformer-CNN hybrid architecture. Sensors, 2023, 23(11): 5288. [doi: 10.3390/s23115288]

10 Zhu L, Wang XJ, Ke ZH, *et al*. BiFormer: Vision Transformer with bi-level routing attention. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 10323–10333.

11 Yeom SK, von Klitzing J. U-MixFormer: UNet-like Transformer with mix-attention for efficient semantic segmentation. arXiv:2312.06272, 2023.

12 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

13 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2999–3007.

14 Wang W, Tang C, Wang X, *et al*. A ViT-based multiscale

feature fusion approach for remote sensing image segmentation. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4510305.

15 Wu XS, Wang LL, Wu CY, *et al*. Semantic segmentation of remote sensing images using multiway fusion network. Signal Processing, 2024, 215: 109272. [doi: 10.1016/j.sigpro.2023.109272]

16 Wang Y, Sun ZC, Zhao W. Encoder-and decoder-based networks using multiscale feature fusion and nonlocal block for remote sensing image semantic segmentation. IEEE Geoscience and Remote Sensing Letters, 2021, 18(7): 1159–1163. [doi: 10.1109/LGRS.2020.2998680]

17 贾克斌, 何岩, 魏之皓. 融合多尺度特征的高分辨率森林遥感图像分割. 北京工业大学学报, 2024, 50(9): 1089–1099. [doi: 10.11936/bjutxb2023010021]

18 Xie EZ, Wang WH, Yu ZD, *et al*. SegFormer: Simple and efficient design for semantic segmentation with Transformers. Advances in Neural Information Processing Systems, 2021, 34. 12077–12090.

19 Guo MH, Lu CZ, Hou QB, *et al*. SegNeXt: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems, 2022, 35. 1140–1156.

20 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 10012–10022.

(校对责编: 王欣欣)