

频率和空间特征融合的轻量级多尺度遥感 图像场景分类网络

王 威, 孙钰洁, 王 新

(长沙理工大学 计算机与通信工程学院, 长沙 410114)

摘 要: 针对遥感图像分类任务中土地覆盖物尺寸和空间组合多种多样、类间相似性高和类内差异性大的问题, 从特征的有效提取和多尺度特征的充分融合出发, 设计了一种频率和空间特征融合的轻量级多尺度遥感图像场景分类网络(FS-LMFFNet)。首先, 为了结合卷积神经网络(CNN)和 Transformer 的优点, 实现局部和全局特征的充分提取, 提出了一种频率和空间多层感知机模块(FS-MLP), 该模块通过引入频域分析, 补充了传统空间操作在提取全局高频纹理特征方面的不足。其次, 针对遥感场景图像的多尺度特性, 提出了一种轻量级多层特征融合模块(LMFF), 该模块采用轻量级卷积块对前 3 个阶段的多尺度特征进行有效的融合。最后, 基于 FS-MLP 和 LMFF 模块构建的 FS-LMFFNet 在 3 个公开数据集 UC_Merced、RSSCN7 和 AID 上进行实验, 准确率分别达到 99.10%、96.60% 和 95.48%。实验结果表明, 本文提出的 FS-LMFFNet 能更好地提取和融合多尺度特征, 从而取得优于其他先进模型的性能。

关键词: 遥感图像; 深度学习; 卷积神经网络(CNN); 快速傅里叶变换(FFT); 多尺度特征融合
中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1671-5497(2025)10-3361-11

DOI: 10.13229/j.cnki.jdxbgxb.20240054

Lightweight frequency and spatial feature fused multi-scale remote sensing scene classification network

WANG Wei, SUN Yu-jie, WANG Xin

(School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China)

Abstract: To address the issues of diverse land cover sizes and spatial combinations, as well as significant interclass similarity and intraclass variability in remote sensing image classification tasks, a lightweight frequency and spatial feature fused multi-scale remote sensing scene classification network (FS-LMFFNet) is proposed, based on the purpose of effective feature extraction and full integration of multi-scale features. Firstly, to combine the advantages of CNN and Transformer, and achieve an adequate

收稿日期: 2024-01-15.

基金项目: 国家重点行动计划项目(6240XXX0206); 国防科技创新特区项目(2019XXXX0701); 湖南省重点研究开发项目(2020SK2134); 湖南省自然科学基金项目(2019JJ80105, 2022JJ30625); 长沙市科技计划项目(kq2004071).

作者简介: 王威(1974-), 男, 教授, 博士. 研究方向: 智能信息处理. E-mail: wangwei@csust.edu.cn

通信作者: 王新(1976-), 女, 讲师, 硕士. 研究方向: 智能信息处理. E-mail: wangxin@csust.edu.cn

extraction of local and global features, a Frequency and Spatial MLP module (FS-MLP) is proposed, which complements traditional spatial operations in extracting global high-frequency texture features by introducing frequency domain analysis. Secondly, to resolve the multi-scale characteristics of remote sensing scene images, a Lightweight Multi-layer Feature Fusion (LMFF) module is proposed, in which lightweight convolutional blocks are employed to efficiently fuse the multi-scale features in the first three stages. Finally, FS-LMFFNet has been extensively experimented on three publicly available datasets UC_Merced, RSSCN7 and AID datasets and yielded remarkable accuracies of 99.10%, 96.60% and 95.48%, respectively. Experimental results demonstrate the superior multi-scale feature extraction and fusion capability of FS-LMFFNet, which achieves better performance than other state-of-the-art models.

Key words: remote sensing images; deep learning; convolutional neural network (CNN); fast Fourier transform (FFT); multi-scale feature fusion

0 引言

遥感图像场景分类是根据图像内容推断出正确的类别,近年来随着成像技术的发展而迅速发展^[1]。遥感图像场景分类是遥感领域的基础研究方向,可用于城市规划、土地资源管理、农业、林业等。然而,与自然图像中物体占据大部分空间不同,遥感场景图像中通常包含尺寸多样的目标,且存在类间相似性高和类内差异性大的问题,使场景分类任务具有挑战性。因此,如何学习具有强辨别力的鲁棒特征表示对遥感图像场景分类至关重要。

在过去的几十年间出现了大量的方法提高遥感图像场景分类性能。早期的场景分类方法主要采用基于研究人员手工设计的低级特征,如灰度值^[2]、平均值^[3]和马尔可夫随机场(MRF)模型^[4]等进行分类,但其对深层语义的抽象能力不足,限制了这些方法在具有挑战性的场景下的适用性。

卷积神经网络(Convolutional neural networks, CNN)通过多个卷积层的叠加和池化层捕获局部表征。自 2012 年 AlexNet^[5]提出以来, CNN 在计算机视觉领域得到迅速发展。得益于其捕获深度和抽象特征表示的能力,非常适合处理遥感图像中的空间结构化数据。然而,由于卷积核大小的限制, CNN 的感受野有限,缺乏获得长期依赖的能力,不利于捕获遥感图像复杂的多尺度内容^[6]。RepLKNet^[7]和 VAN^[8]利用大卷积核弥补局部感受野的缺陷,但效果仍然有限。

近年来,运用于自然语言处理设计的转换器(Transformer)^[9]架构被引入图像分类领域,解决了局部感受野的问题。视觉转换器(Vision transformer, ViT)^[10]将图像分割成小块(patch),并将

这些 patch 的线性嵌入作为自注意力模块(MSA)的输入,从而获得全局感受野。文献[11]将 ViT 模型应用于遥感场景分类,获得了良好的分类效果。然而, ViT 仍然有一些缺点。首先, MSA 只学习单一尺度的全局信息,限制了对具有复杂的多尺度空间结构的遥感场景特征的探索。其次,由于 MSA 操作, Transformer 的计算复杂度较高。

最近,一些基于 MetaFormer^[12]的方法尝试通过将 Transformer 中的多头自注意力块替换为精心设计的令牌混合器(Token mixer),以结合 CNN 和 Transformer 的优点,并展示了优异鼓舞的效果。ADC-CDANet^[13]具有很好的全局特征和局部特征提取能力,文献[14]提出了一种二维特征表示的二维特征表示(Bidimensional feature representation, BFR)块,在提取空间域全局和局部特征的同时,还融合了通道信息。但这些方法只提取网络最后一层的全局图像级特征用于遥感场景分类,忽略了局部对象级特征,限制了场景分类精度的进一步提升。

在本文中引入频域分析,根据傅里叶变换的频域卷积定理,频域中逐点更新会影响傅里叶变换所涉及的全局特征,可以用来弥补卷积等空间操作在下采样过程中导致高频细节信息的丢失,以构建与 MSA 相同的全局感受野的神经结构^[15]。此外,快速傅里叶变换(FFT),比 MSA 的计算效率更高。因此,从空间和频域提取特征的令牌混合器是将 CNN 和 Transformer 的优点相结合的有效方法。

随着深度学习网络层数的不断加深,提取到的特征具有不同的特征尺度和语义层次。低层特征通常包含一些浅层信息,如颜色和位置,中间层

通常是单个对象,而高层则表示整个场景的语义信息。对于遥感图像场景分类任务,因为遥感图像的土地覆盖物尺寸和空间组合复杂,只考虑全局特征可能出现将背景当成关键特征,导致分类错误的情况,因此需要综合考虑不同层之间的上下文相关性。最近的一些方法(如SAFF^[16]、GLFFNet^[17]),将多层特征简单相加聚合它们之间丰富的上下文信息,这会带来对象大小和位置的混乱。

根据上述分析,为了实现高质量的场景分类,需要解决两个问题:特征的有效提取和多尺度特征的充分融合。针对上述问题,本文提出了一种基于频率和空间特征的多尺度遥感场景分类网络(FS-LMFFNet)。本文的主要贡献如下:

(1)为了结合CNN和Transformer的优势,实现局部和全局特征的充分提取,设计了一种频率和空间多层感知机模块(Frequency and spatial MLP, FS-MLP)。该模块运用深度卷积对图像的空间信息建模,采用可学习的全局滤波器补充卷积等空间操作丢失的全局高频纹理特征,通过多层感知机对空间和频率信息进行整合。

(2)为了解决遥感场景图像的多尺度的问题,提出了一种轻量级多层特征融合模块(Light-weight multi-layer feature fusion, LMFF)。该模块利用两个沙漏-位置注意力块对前3个阶段提取的不同特征尺度和语义层次的上下文特征进行有效融合。

(3)基于FS-MLP模块和LMFF模块设计遥感图像场景分类模型FS-LMFFNet。在RSS-CN7、UC-Merced和AID数据集上的实验结果表明,FS-LMFFNet优于其他先进模型。

1 模型结构设计

1.1 频率和空间多层感知机模块(FS-MLP)

遥感图像场景分类的最终效果很大程度上取决于特征提取的能力。当前的很多网络都是在图像的原始空间域对局部特征进行提取,但空间域的卷积、池化等操作会造成高频纹理细节的丢失。为了解决这一问题,本文在遥感图像场景分类中引入频域特征,在频域中学习全局像素之间的相互作用。

遵循MetaFormer的范式,本文设计了频率和空间多层感知机模块,它由频率和空间令牌混合器(Frequency and spatial token mixer, FSTM)和多层感知机(Multilayer perceptron, MLP)组成。MLP由两个全连接层和GELU^[18]激活层组成。FS-MLP的结构如图1(a)所示,公式如下:

$$C'_i = \text{FSTM}[\text{LN}(C_i)] + C_i \quad (1)$$

$$P_i = \text{MLP}[\text{LN}(C'_i)] + C'_i \quad (2)$$

式中: $i=1,2,3$; C_i 为FS-MLP的输入; P_i 为FS-MLP的输出;LN表示层归一化(Layer normalization)。

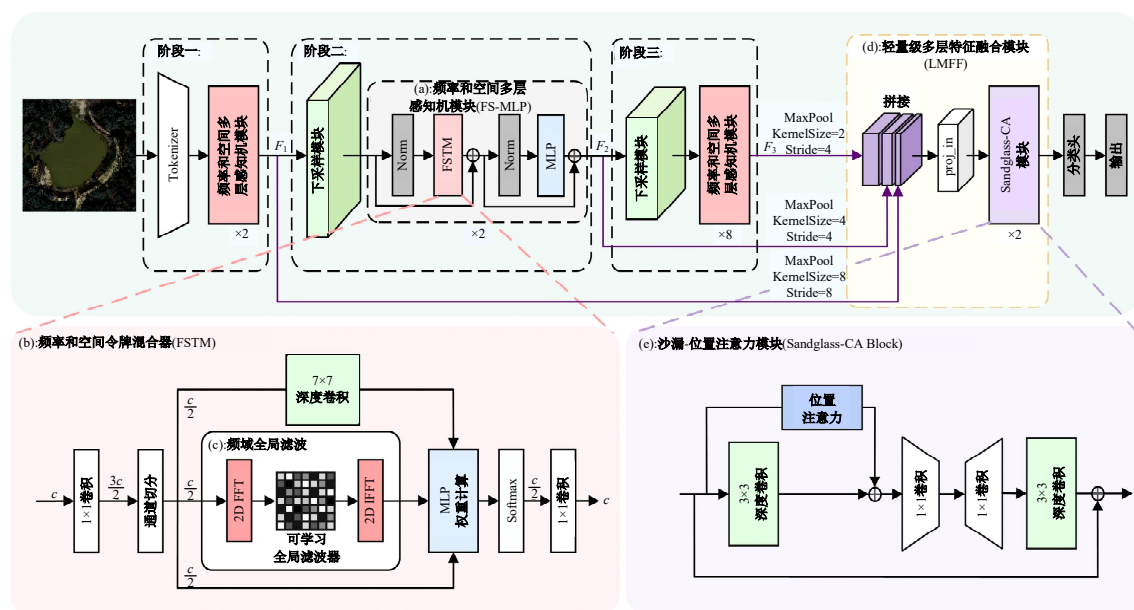


图1 FS-LMFFNet结构

Fig. 1 FS-LMFFNet structure

FSTM的架构如图1(b)所示。设FSTM的输入为 $x_{in} \in \mathbb{R}^{h \times w \times c}$ 。首先,采用 1×1 卷积将输入特征的通道变为 $3c/2$,然后将其分割成3个平行支路 $x_1, x_2, x_3 \in \mathbb{R}^{h \times w \times (c/2)}$,3个支路的操作如下:

(1)空间分支:采用 7×7 深度卷积分支提取空间信息。采用 7×7 的大卷积核可以获得更大的感受野,而深度卷积更关注空间维度的信息。

$$f_s = \text{DWconv}_7(x_1) \quad (3)$$

(2)频率分支:采用频域的可学习全局滤波器学习全局像素之间的相互作用。

$$f_l = \text{GFlayer}(x_2) \quad (4)$$

全局滤波器的流程如图1(c)所示。频域全局可学习滤波器的运算过程如下:

应用二维快速傅里叶变换将输入张量转移到频率域:

$$\text{RealFFT}2d: \mathbb{R}^{h \times w \times (c/2)} \rightarrow \mathbb{R}^{h \times (w/2) \times (c/2)}$$

将实部和虚部连接起来,形成特征张量的额外维度。至此,得到所有空间位置对应于不同频率分量的频率表示:

$$\text{ComplexToReal}: \mathbb{R}^{h \times (w/2) \times (c/2)} \rightarrow \mathbb{R}^{h \times (w/2) \times c}$$

为了提取有用的特征并改进频域表示,对频率表示进一步处理为采用可学的权重矩阵 $K \in \mathbb{C}^{h \times w \times (c/2)}$ 与全局特征相乘:

$$K \odot \mathcal{F}(x_2): \mathbb{R}^{h \times (w/2) \times c} \rightarrow \mathbb{R}^{h \times (w/2) \times c}$$

最后,通过逆傅里叶变换将处理后的频率表示转移回原始空间域:

$$\text{RealToComplex}: \mathbb{R}^{h \times (w/2) \times c} \rightarrow \mathbb{C}^{h \times (w/2) \times c}$$

$$\text{InverseRealFFT}2d: \mathbb{C}^{h \times (w/2) \times c} \rightarrow \mathbb{R}^{h \times w \times (c/2)}$$

全局滤波器可以解释为一个捕捉频域关系的深度全局圆卷积,其滤波器大小随着输入特征的大小自适应地变化。同时,得益于快速傅里叶变换,它还十分高效。

(3)原始分支:第三条分支保留了原始特征。

$$f_i = x_3 \quad (5)$$

不同于简单地将3个分支相加,通过MLP和Softmax激活函数计算3条支路的权重。利用这些步骤,可以互补地将频率特征、空间特征和原始特征进行整合,实现更好的全局和局部场景理解和泛化能力。权重计算公式如下所示:

$$W = \text{Softmax}\left\{\text{MLP}\left[\text{GAP}\left(f_s + f_f + f_i\right)\right]\right\} \quad (6)$$

式中:GAP为全局平均池化; $W \in \mathbb{R}^{3 \times 1 \times 1 \times (c/2)}$ 为3个分支的权重矩阵。重加权后的 f 为:

$$f = f_s \times W[0] + f_l \times W[1] + f_i \times W[2] \quad (7)$$

最后,采用 1×1 卷积将 f 的通道恢复到 $x_{out} \in \mathbb{R}^{h \times w \times c}$ 。

1.2 轻量多层特征融合模块(LMFF)

不同于自然图像分类,遥感图像的地物组成复杂、多尺度性质突出。同时,网络不同深度的特征存在本质的语义差异,对不同层次的上下文进行聚合和对齐可以缓解网络对特征大小的混淆,充分利用不频繁出现的特征。因此,将前3个阶段的输出特征图 $F_1 \in \mathbb{R}^{H/4 \times W/4 \times 64}$ 、 $F_2 \in \mathbb{R}^{H/8 \times W/8 \times 96}$ 、 $F_3 \in \mathbb{R}^{H/16 \times W/16 \times 128}$ 进行池化操作,以获得统一的 $H/32 \times W/32$ 大小的特征图:

$$G_1 = \text{MaxP}_8^8(F_1) \quad (8)$$

$$G_2 = \text{MaxP}_4^4(F_2) \quad (9)$$

$$G_3 = \text{MaxP}_2^2(F_3) \quad (10)$$

式中: MaxP_a^b 为最大池化,步长为 a ,核大小为 b 。选择最大池化而不是平均池化,因为它保留了更多的细节,如边缘和纹理。然后将池化后的特征级联,通过 1×1 卷积,形成特征图 $G \in \mathbb{R}^{H/32 \times W/32 \times 288}$:

$$G = \text{Conv}_1[\text{concat}(G_1, G_2, G_3)] \quad (11)$$

为了最大限度地利用多尺度特征,采用了两个沙漏-位置注意力块(Sandglass-CA block)进行多尺度特征融合,如图1(e)所示。不同于逆残差块^[19]在两个 1×1 卷积之间引入深度卷积,沙漏块^[20]在高维度上进行残差映射和深度卷积空间计算,避免 1×1 卷积导致的空间信息丢失。此外,所有卷积后都有批标准化(Batch normalization, BN)^[21],激活函数ReLU6仅加在第一个深度卷积与最后一个用于升维的 1×1 卷积之后,在用于降维的 1×1 卷积和最后一个深度卷积后不添加激活函数,这是为了避免特征出现零化现象,进而导致信息损失。

坐标注意力^[22](Coordinate attention, CA),如图2所示。它将通道注意力分解为两个一维特征编码过程,分别沿两个空间方向聚合特征。通过这种方式,可以沿一个空间方向捕获远程依赖关系,同时可以沿另一个空间方向保留精确的位置信息。然后将得到的特征图单独编码成一对方向感知和位置敏感的注意力图,可以将不同语义层次中相同或位置的特征进行对齐。

1.3 FS-LMFFNet

针对遥感图像纹理信息丰富和整体结构复杂的特点,基于FS-MLP和LMFF,设计了一种遥感图像场景分类模型FS-LMFFNet。其结构如表1所示。

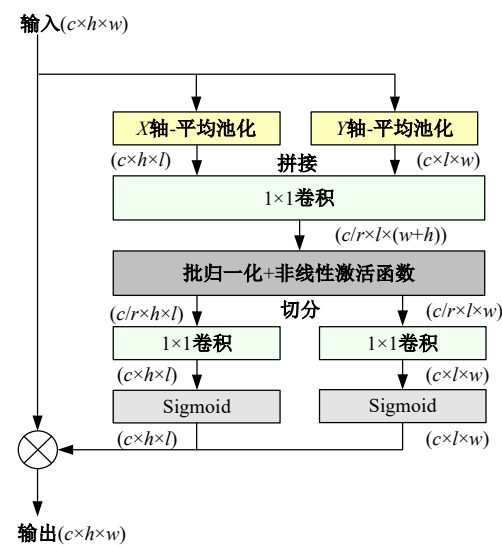


图 2 坐标注意力结构

Fig. 2 Coordinate attention structure

表 1 FS-LMFFNet 网络结构表

Table 1 FS-LMFFNet network structure table

	层	输出大小
阶段一	Tokenizer 层	$(H/4) \times (W/4)$
	FS-MLP 模块 $\times 2$	
阶段二	下采样模块	$(H/8) \times (W/8)$
	FS-MLP 模块 $\times 2$	
阶段三	下采样模块	$(H/16) \times (W/16)$
	FS-MLP 模块 $\times 8$	
阶段四	LMFF 模块	$(H/32) \times (W/32)$
分类器	归一化、全局池化	$C = \text{预测类别}$
	全连接层	

FS-LMFFNet 包括 3 个阶段,每个阶段由下采样模块和 n 个 FS-MLP 模块。第一阶段的下采

样操作用于将图像分割为令牌,称为 Tokenizer,包括一个步幅为 1 的 3×3 卷积、BN、ReLU 激活函数和步幅为 2 的最大池化。其他两个阶段的下采样操作是一个步幅为 2 的 3×3 卷积,因为重叠窗口可以保留更多的空间交集。将 3 个阶段的输出 F_1, F_2, F_3 , 输入 LMFF, 并以 $(H/32) \times (W/32) \times 288$ 的大小输出。最后一步的分类器包括归一化、全局池化和全连接层。

2 实验结果处理与分析

2.1 数据集处理

为了评估 FS-MLFFNet 的有效性,本文在 3 个主流的遥感图像场景分类数据集上进行了实验。

(1) UC_Merced 土地利用数据集^[23] (UC_Merced)

该数据集中有 2 100 张图像,为 21 个场景类别。每个类由 100 张图像组成,大小为 256×256 像素。像素空间分辨率为 0.3 m。该数据集的示例如图 3 所示。

(2) RSSCN7 遥感图像数据集^[24] (RSSCN7)

该数据集包含 7 个类别中的 2 800 张图像。每个类由 400 张的图像组成。每个空间图像的大小为 400×400 像素。

(3) Aerial Image dataset 航空图像数据集^[25] (AID)

该数据集包含 30 类场景的 10 000 张图像。每个类由 220 到 420 张图像组成。每个图像的大

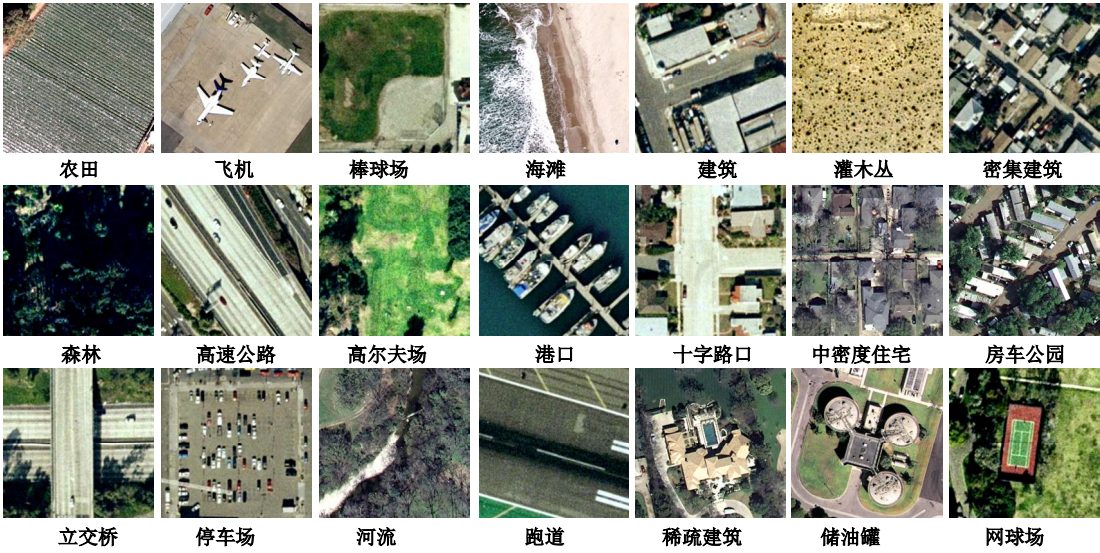


图 3 UC_Merced 数据集示例

Fig.3 Examples of the UC_Merced dataset

小为 600×600 像素。AID 的像素分辨率从大约 8 m 变化到大约 1/2 m, 具有很强的多尺度特性。

为了将实验数据的大小统一, 在训练之前, 将 RSSCN7、UC-Merced 和 AID 数据集中的图像裁剪为 224×224 像素的大小。然后, 将每个数据集中的图像随机分成 80% 用于训练和 20% 用于测试。数据扩充仅在训练集上执行, 包括随机翻转(水平、垂直)和旋转 45° 。

2.2 实验配置

实验采用计算机图像分类任务中常用的评价标准, 即采用参数量(Parameters, Param)、浮点运算次数(Floating point operations, FLOPs)作为模型复杂度的评价指标, 采用准确率(Accuracy)作为模型性能的评价标准。设模型预测正确的正样本数量为 T_P , 预测错误的正样本数为 F_N , 预测正确的负样本数为 T_N , 预测错误的负样本数为 F_P , 准确率的计算公式如下:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (12)$$

本文所有实验都基于配备 NVIDIA A10 GPU 的服务器, 使用 Pytorch 1.3 深度学习框架完成。使用 AdamW 优化器, 学习率设置为 0.000 5, 个别模型如 SwinTransformer、GFNet 的学习率需要另作调整时将其置为 0.000 1。训练集和测试集的批尺寸(Batch size)设置为 16。采取的损失函数为标签平滑损失函数 LabelSmoothSoftmax-CEV2。为了寻找合适的训练迭代次数(epoch), 统计了 FS-LMFFNet 在 3 个数据集上的训练损失值, 结果如图 4 所示。当迭代次数为 300~350 时, 所有模型都收敛。将迭代次数设置为 400, 是为了多一些训练次数使训练更稳定。此外, 为了获得稳定的结果, 对所有网络进行了 5 次重复实验, 计算平均场景分类准确率(ACC)和标准差(STD)作为最终的总体准确率。

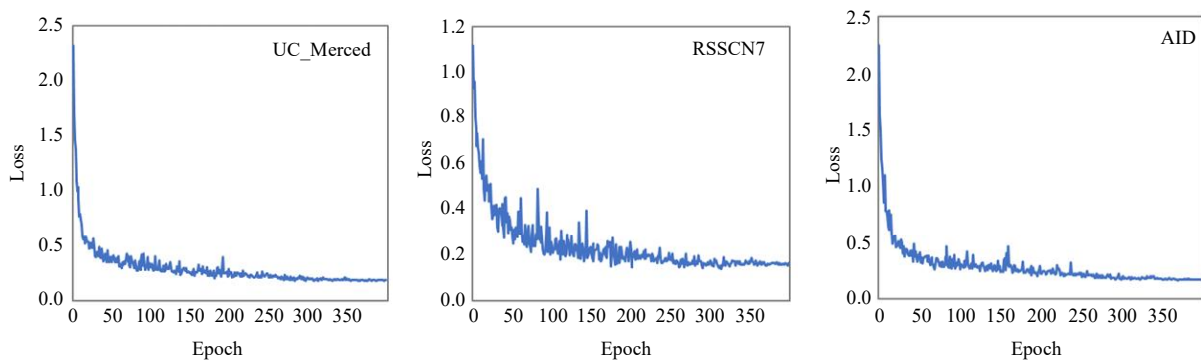


图 4 FS-LMFFNet 在不同数据集上的损失曲线

Fig. 4 Loss curves of FS-LMFFNet on different dataset

2.3 消融实验

为评估每个模块的有效性, 本文在 UC_Merced 数据集上对 FS-LMFFNet 进行消融实验。

2.3.1 FS-MLP 的消融实验

FS-MLP 的关键组件为 FSTM, 它通过 MLP 学习 3 条分支的权重。因此, 本文对 3 条分支和 MLP 权重计算进行了消融实验, 结果如表 2 所示。

当去掉 3 条分支的任何一条后, 网络的性能都有所下降。特别是当将频域分支去掉时准确率显著降低 0.74%。这说明, 从频域分析中提取的全局高频信息对场景分类任务是至关重要的, 虽然频域分支带来一定的参数量, 但这是可以接受的。将空间分支删除之后, 网络的准确率降低 0.48%, 这表明, 大卷积核深度卷积是一种非常有效的建模空间信息的方法。

此外, 将深度卷积的卷积核大小设置为 3 后, 相比于没有空间信息的情况, 此时网络性能提升 0.1%, 而与卷积核大小为 7 的性能还是有 0.38% 的差距, 这证明了增加卷积核的大小可以获得更大的感受野, 实现捕获更多尺度的空间信息的目的。将原始分支删除后, 网络性能降低 0.24%, 说明原始信息应该被保留。此外, 将计算 3 条支路权重的 MLP 删除, 改为将 3 条支路简单相加, 此时准确率为 98.50%。

2.3.2 LMFF 的消融实验

在 LMFF 中通过不同池化核大小的池化层将前 3 个阶段的输出转换为同一尺寸后, 输入沙漏-坐标注意力模块进行多尺度特征聚合。因此, 本文评估了不同的下采样策略和多尺度特征聚合方法对遥感图像场景分类的影响。

不同的下采样策略消融实验结果如表 3 所

表 2 FSTM 的消融实验

Table 2 Ablation experiments with FSTM

频率分支	空间分支	原始分支	权重计算	参数量/M	计算量/G	准确率/%
×	Kernel size=7	✓	✓	1.99	1.09	98.36±0.26
✓	×	✓	✓	2.31	1.07	98.62±0.20
✓	Kernel size=3	✓	✓	2.41	1.11	98.72±0.11
✓	Kernel size=7	×	✓	2.34	1.09	98.86±0.26
✓	Kernel size=7	✓	×	2.36	1.12	98.50±0.35
✓	Kernel size=7	✓	✓	2.43	1.12	99.10±0.22

注:粗体表示最优值。下表同。

示。首先,将池化核的大小设为恒定的 2,此时 F_1 、 F_2 的步距大于池化核的大小,下采样时无法获取特征图的全部信息,所以网络的性能下降了 0.4%。其次,将 LMFF 的输入变为 F_3 ,此时采用步幅为 2 的 3×3 卷积下采样块进行下采样,网络性能降低了 0.94%。这证明了对多尺度特征进行聚合的重要性。

表 3 LMFF 下采样操作的消融实验

Table 3 Ablation experiments for LMFF's downsample operations

下采样方法	参数量/M	计算量/G	准确率/%
(一,一,下采样块)	2.67	1.14	98.16±0.30
($\text{MaxP}_2^8, \text{MaxP}_2^4, \text{MaxP}_2^2$)	2.44	1.12	98.70±0.26
($\text{MaxP}_8^8, \text{MaxP}_4^4, \text{MaxP}_2^2$)	2.43	1.12	99.10±0.22

此外,本文还评估了不同的融合策略对场景分类的影响,结果如表 4 所示。首先,使用本文提出的 FS-MLP 进行多尺度特征聚合,相比于沙漏-坐标注意力模块,准确率下降了 0.68%,这是因为在经过 3 个阶段的 FS-MLP 之后网络对特征的提取已经饱和,所以在网络的最后一个阶段的任务主要是多尺度特征的聚合,此时再使用 FS-MLP 是冗余的,还会带来更多的计算开销。其次,评估了位置注意力对分类效果的影响,插入位置注意力后网络的性能提升了 0.12%。

表 4 LMFF 特征融合方法的消融实验

Table 4 Ablation experiments for LMFF's feature fusion method

模型	参数量/M	计算量/G	准确率/%
FS-MLP	4.32	1.12	97.42±0.42
Sandglass Block	2.42	1.12	98.98±0.20
Sandglass_CA Block	2.43	1.12	99.10±0.22

2.4 对比实验

为了验证 FS-LMFFNet 的有效性,在本节中设置 3 类对照实验。第一类包含基于 CNN 的自然图像分类网络;第二类包括基于 MetaFormer 架

构的自然图像分类模型;第三类是遥感图像场景分类方法。为保证实验结果公平,所有的方法都在没有经过预训练的条件下进行实验。

由表 5 可知,与上述方法相比,FS-LMFFNet 的参数量是最小的,计算量仅高于 3 个轻量级的分类网络。特别是与其他遥感图像场景分类方法相比,FS-LMFFNet 的参数量和计算量都非常低。FS-LMFFNet 在 UC_Merced、RSSCN7 和 AID 数据集上的准确率分别为 99.10%、96.60%、95.48%,除准确率高外,FS-LMFFNet 的标准差也较低,这表明本文的网络性能更加稳定。

基于 CNN 的方法比一些基于 MetaFormer 的方法准确率更高,是因为 CNN 的局部特征提取能力更适合目标尺寸多样的遥感图像数据,而 Swin Transformer^[26]等方法更适用于数据集样本数量较大的任务。2023 年提出的 Strip-MLP^[27]有效地增强了令牌交互和局部令牌聚合能力,Strip-MLP 较高的准确率也验证了多尺度特征提取的重要性。此外,遥感图像场景分类方法的准确率比自然图像分类方法的准确率普遍要高,得益于它们在网络结构设计时考虑了遥感图像的多尺度特性。但 SAFF^[16]的骨干网络为较原始的 VGG16^[28],导致其性能较低。

UC_Merced 数据集通过聚合密集特征生成全局图像表示,FS-LMFFNet 通过对频率和空间特征的提取和分析,获得比 EMTCAL 的准确率高 0.32% 的结果。虽然 RSSCN7 数据集只有 7 类,但是其训练样本较少,且类间相似性高和类内差异性大,这增加了分类的难度。得益于 FS-LMFFNet 中的频率和空间多层感知机能够从频域和空间域互补地对特征进行提取,其在 RSSCN7 数据集上的分类准确率为 96.60%,与分类效果较好的 EMTCAL 和 Strip-MLP 相比,准确

表 5 不同方法在 3 个数据集上的准确率和计算复杂度对比

Table 5 Overall ACC and computation complexity of different methods on three datasets

模型	参数量/M	计算量/G	准确率/%		
			UC_Merced	RSSCN7	AID
ResNet18 ^[29]	11.15	1.82	98.28±0.25	94.96±0.42	94.50±0.19
FcaNet18 ^[30]	11.23	1.82	97.96±0.47	94.96±0.52	94.34±0.34
MobileNeXt ^[20]	2.86	0.29	97.14±0.46	94.64±0.73	95.08±0.16
MobileNeXt_CA ^[22]	3.26	0.29	97.36±0.37	95.10±0.39	95.26±0.11
SwinTransformer_Tiny ^[26]	26.96	4.2	95.48±0.45	93.40±0.52	90.20±0.20
VAN_b0 ^[8]	3.91	0.88	97.62±0.18	94.28±0.41	93.88±0.23
GFNet_PyramidTi ^[31]	12.18	1.90	95.18±0.68	91.54±0.44	90.90±0.57
WaveMLP_T ^[32]	16.40	2.48	96.86±0.72	92.48±0.41	92.54±0.40
ConvMLP_S ^[33]	8.60	2.30	95.24±0.86	94.62±0.28	93.38±0.22
Strip-MLP-T* ^[27]	18.24	2.54	98.72±0.16	95.32±0.51	95.12±0.23
SAFF ^[16]	14.76	15.38	95.58±0.39	93.78±0.63	94.18±0.17
RaNet ^[34]	21.47	3.85	98.38±0.39	95.24±0.15	95.38±0.22
EMTCAL ^[35]	27.30	4.23	98.78±0.27	95.32±0.37	94.96±0.19
FS-LMFFNet	2.43	1.12	99.10±0.22	96.60±0.24	95.48±0.13

率高 1.28%。AID 数据集的图像像素分辨率变化较大,多尺度特性明显。尽管分类难度较大,FS-LMFFNet 对前 3 个阶段提取的不同特征尺度和语义层次的上下文特征进行了有效融合。最终得到比 RaNet 的准确率高 0.1% 的实验结果,而 RaNet 的参数量是 FS-LMFFNet 的 6 倍。综上所述,FS-LMFFNet 较其他方法具有更高的效

率和更好的分类准确率,更适应于数据规模较小、多尺度特性明显的场景。

2.5 可视化

为了进一步研究不同场景类别的分类效果,绘制了 AID 数据集中准确率最高的一次实验结果的混淆矩阵,如图 5 所示。在 AID 数据集的 30 个类别中,有 14 个类别的分类准确率较高(≥

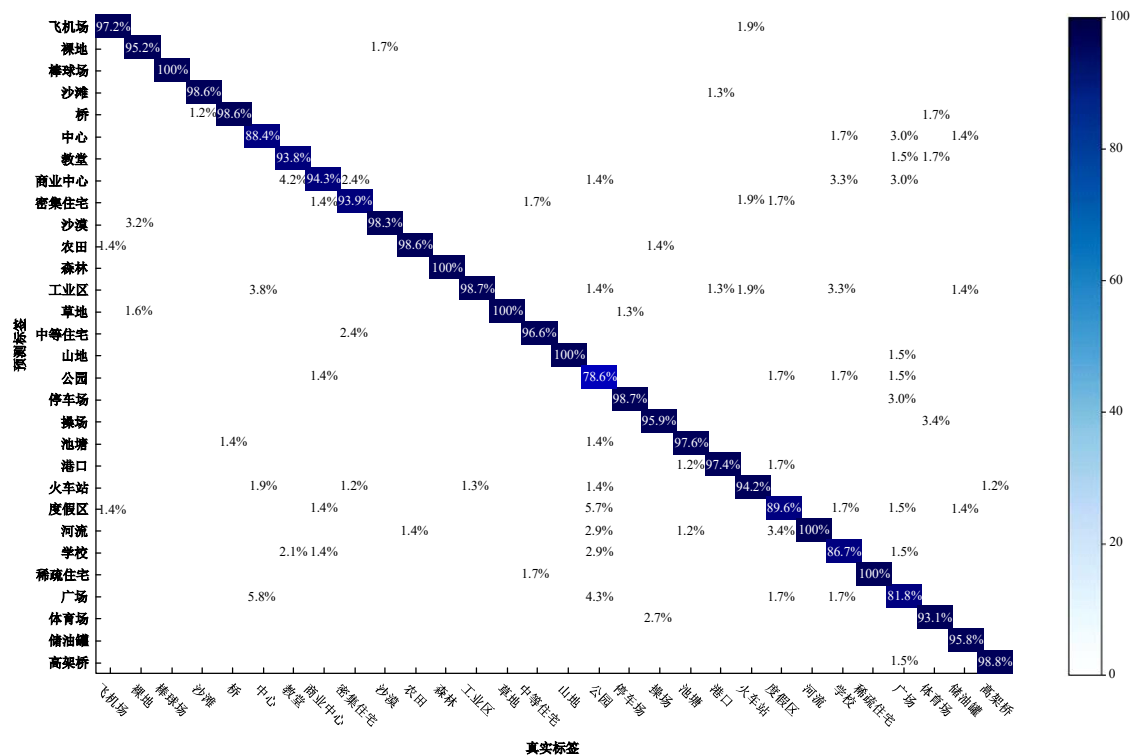


图 5 FS-LMFFNet 在 AID 数据集上的混淆矩阵

Fig. 5 Confusion matrix of the FS-MLP method on AID dataset

98%),对于“篮球场”“森林”“草地”“山地”和“河流”等具有明显判别特征的类别,FS-LMFFNet的分类准确率达到100%。对于一些类间差异较小的类,如“密集住宅”(94%)、“中等住宅”(97%)、“稀疏住宅”(100%),土地覆盖物相似但密度不同,得益于从频域对目标出现的频率分析,FS-LMFFNet可以对建筑物出现频率进行分析,从而做出较为准确的判断。主要的混淆发生在“公园”和“度假村”,是因为两个类别土地覆盖物(如游乐设施、绿化带)种类相同、空间分布相似,极大地提高了分类难度。尽管如此,FS-LMFFNet获得较有竞争力的结果。

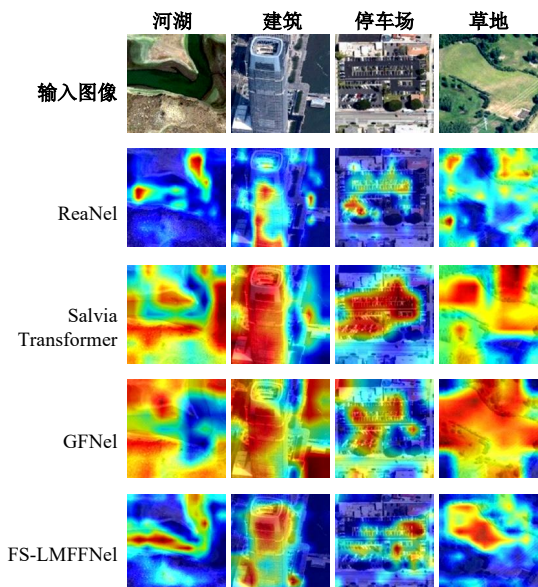


图6 FS-LMFFNet在RSSCN7数据集上的Grad-CAM可视化结果

Fig.6 Grad-CAM visualization results on RSSCN7 dataset

此外,本文通过GradCAM^[36]绘制热力图进行可视化网络做出判断所关注的区域,如图6所示,从上到下分别是原始图像和不同网络的热力图。可以看出,基于CNN的ResNet倾向于关注特定对象,而基于Transformer的SwinTransformer和使用全局滤波器的GFNet更关注全局特性。相比之下,FS-LMFFNet可以更准确地聚焦于显著的物体,同时保持全局感知的能力。特别是在第一列中,FS-LMFFNet的热力图对河流做出准确的定位,而其他方法有的只关注到河流的局部,有的没有关注到具体目标。这进一步解释了为什么FS-LMFFNet在3个数据集中都能达到最好的结果。

3 结束语

针对遥感图像土地覆盖物尺寸和空间组合多种多样、类间相似性高和类内差异性大的特点,本文从特征的有效提取和多尺度特征的充分融合出发,提出了频率和空间特征融合的多尺度遥感场景分类网络(FS-LMFFNet)。本文通过引入频域分析,结合CNN和Transformer的优点,构建了频率和空间多层感知机模块。同时,为了解决遥感场景图像的多尺度问题,提出了一种轻量级多层特征融合模块。在3个基准测试上的实验结果表明,本文所提出的FS-LMFFNet能为遥感图像场景分类带来显著的性能改进,而且具有更低的计算复杂度。然而,当遥感图像土地覆盖物种类相同、分布相似时,很难提高分类精度。下一步工作将进一步探索其他的频域特征提取方法,如小波变换等,进一步提高模型的表征能力,从而提高遥感图像场景分类的精度。

参考文献:

- [1] 徐从安,吕亚飞,张筱晗,等. 基于双重注意力机制的遥感图像场景分类特征表示方法[J]. 电子与信息学报, 2021, 43(3): 683-691.
Xu Cong-an, Lyu Ya-fei, Zhang Xiao-han, et al. A discriminative feature representation method based on dual attention mechanism for remote sensing image scene classification[J]. Journal of Electronics & Information Technology, 2021, 43(3): 683-691.
- [2] Morell-Monzó S, Sebastián-Frasquet M T, Estornell J. Land use classification of VHR images for mapping small-sized abandoned citrus plots by using spectral and textural information[J]. Remote Sensing, 2021, 13(4): No. 681.
- [3] Liang S, Cheng J, Zhang J. Maximum likelihood classification of soil remote sensing image based on deep learning[J]. Earth Sciences Research Journal, 2020, 24(3): 357-365.
- [4] Fatemighomi H S, Golalizadeh M, Amani M. Object-based hyperspectral image classification using a new latent block model based on hidden Markov random fields[J]. Pattern Anal Applic, 2022, 25: 467-481.
- [5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

- [6] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 764-773.
- [7] Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31×31 : revisiting large kernel design in CNNs[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, Louisiana, USA, 2022: 11953-11965.
- [8] Guo M H, Lu C Z, Liu Z N, et al. Visual attention network[J]. Computational Visual Media, 2022, 9(4): 733-752.
- [9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of 31st Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000-6010.
- [10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. [2022-10-18]. <https://arxiv.org/abs/2010.11929>.
- [11] Bazi Y, Bashmal L, Rahhal M M A, et al. Vision transformers for remote sensing image classification[J]. Remote Sensing, 2021, 13(3): No. 516.
- [12] Yu W H, Luo M, Zhou P, et al. Meta former is actually what you need for vision[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 10809-10819.
- [13] 王威, 李希杰, 王新. ADC-CPANet: 一种局部-全局特征融合的遥感图像分类方法[J]. 遥感学报, 2024, 28(10): 2661-2672.
Wang Wei, Li Xi-jie, Wang Xin. ADC-CPANet: a remote sensing image classification method based on local-global feature fusion[J]. National Remote Sensing Bulletin, 2024, 28(10): 2661-2672.
- [14] Wang W, Hu T, Wang X, et al. BFRNet: bidimensional feature representation network for remote sensing images classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-13.
- [15] Huang Z, Zhang Z, Lan C, et al. Adaptive frequency filters as efficient global token mixers[EB/OL]. [2023-03-22]. <https://arxiv.org/abs/2307.14008>.
- [16] Cao R, Fang L, Lu T, et al. Self-attention-based deep feature fusion for remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18(1): 43-47.
- [17] 王威, 邓纪伟, 王新, 等. 面向遥感图像场景分类的 GLFFNet 模型[J]. 测绘学报, 2023, 52(10): 1693-1702.
Wang Wei, Deng Ji-wei, Wang Xin, et al. GLFFNet model for remote sensing image scene classification[J]. Acta Geodaetica ET Cartographica Sinica, 2023, 52(10): 1693-1702.
- [18] Hendrycks D, Gimpel K. Gaussian error linear units (GELUs) [EB/OL]. [2024-01-10]. <https://arxiv.org/pdf/1606.08415>.
- [19] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 4510-4520.
- [20] Zhou D, Hou Q, Chen Y, et al. Rethinking bottleneck structure for efficient mobile network design[J]. In Computer Vision-ECCV 2020, Lecture Notes in Computer Science, 2020, 12348: 680-697.
- [21] Sergey I, Christian S. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015: 448-456.
- [22] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Montreal, Canada, 2021: 13713-13722.
- [23] Yang Y, Shawn N. Bag-of-visual-words and spatial extensions for land-use classification[C]//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose California, USA, 2010: 270-279.
- [24] Zou Q, Ni L H, Zhang T, et al. Deep learning based feature selection for remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(11): 2321-2325.
- [25] Xia G S, Hu J, Hu F, et al. AID: a benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.
- [26] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021: 10012-10022.
- [27] Cao G, Luo S, Huang W, et al. Strip-MLP: effi-

- cient token interaction for vision MLP[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France , 2023: 1494-1504.
- [28] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2023-03-18]. <https://arxiv.org/abs/1409.1556>.
- [29] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778.
- [30] Qin Z, Zhang P, Wu F, et al. FcaNet: frequency channel attention networks[C] //Proceedings of the IEEE International Conference on Computer Vision, Xi'an, China, 2020: 763-772.
- [31] Rao Y, Zhao W, Zhu Z, et al. Global filter networks for image classification[J]. Advances in Neural Information Processing Systems, 2021, 2: 980-993.
- [32] Tang Y, Han K, Guo J, et al. An image patch is a wave: phase-aware vision MLP[EB/OL]. [2023-03-18]. <https://arxiv.org/abs/2111.12294>.
- [33] Li J, Hassani A, Walton S, et al. ConvMLP: Hierarchical Convolutional MLPs for Vision[C] //IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Vancouver , Canada, 2023: 6307-6316.
- [34] Wang X, Duan L, Ning C, et al. Relation-attention networks for remote sensing scene classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 422-439.
- [35] Tang X, Li M, Ma J, et al. EMTCAL: efficient multiscale transformer and cross-level attention learning for remote sensing scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-15.
- [36] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C] //Proceedings of 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 618-626.