计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(8):252−263 [doi: 10.15888/j.cnki.csa.009901] [CSTR: 32024.14.csa.009901]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合多尺度特征和细节感知策略的遥感图像场景分类模型①

马 惠[1], 霍 然[2]

[1](河南省国土空间调查规划院, 郑州 450016)
[2](中国地质大学 (武汉) 地理与信息工程学院, 武汉 430074)
通信作者: 马 惠, E-mail: 42879575@qq.com

摘 要: 针对遥感图像场景分类中存在的场景尺度多变性、类内多样性和类间相似性, 以及有标签训练样本稀缺的问题, 本文提出了一种结合多尺度特征和细节感知策略的 Vision Transformer (ViT) 模型用于遥感图像场景分类. 该模型通过引入空洞空间金字塔池化模块, 有效捕捉并融合了遥感图像中的多尺度特征, 同时增强了对局部特征信息的利用, 从而进一步提升了特征判别能力. 另外, 采用创新的细节感知掩码策略, 使得模型能够有效利用无标签遥感图像数据, 促进模型学习到更为精细的特征表示, 以实现更高效、更准确的遥感图像场景分类. 在实验部分, 本文首先在大规模无标签遥感图像数据集上进行预训练, 随后将预训练模型迁移至下游场景分类任务中进行微调. 在多个公开遥感图像数据集上的实验结果表明, 所提模型在自监督预训练阶段能够有效提取图像特征, 并在下游场景分类任务中实现较高的准确率, 展现出良好的鲁棒性和有效性.

关键词: 多尺度特征; 细节感知策略; 遥感图像场景分类; 空洞空间金字塔池化; Vision Transformer (ViT)

引用格式: 马惠,霍然.结合多尺度特征和细节感知策略的遥感图像场景分类模型.计算机系统应用,2025,34(8):252–263. http://www.c-s-a.org.cn/1003-3254/9901.html

Remote Sensing Image Scene Classification Model Combining Multi-scale Feature and Detail-aware Strategy

MA Hui[1], HUO Ran[2]

[1](Henan Provincial Institute of Land and Space Investigation and Planning, Zhengzhou 450016, China)
[2](School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China)

Abstract: In response to the challenges of scene scale variability, intra-class diversity, and inter-class similarity in remote sensing image scene classification, a Vision Transformer (ViT) model that integrates multi-scale features with detail perception strategies is proposed for remote sensing image classification. The model effectively captures and fuses multi-scale features from remote sensing images by incorporating a dilated spatial pyramid pooling module, while enhancing the utilization of local feature information, thus improving feature discrimination capabilities. Furthermore, an innovative detail perception masking strategy enables the model to leverage unlabeled remote sensing image data effectively, facilitating the learning of more refined feature representations for more efficient and accurate scene classification. In the experimental section, the model is first pre-trained on a large-scale unlabeled remote sensing image dataset, followed by the fine-tuning of the pre-trained model on downstream scene classification tasks. Experimental results across multiple public remote sensing image datasets demonstrate that the proposed model can effectively extract image features during the self-supervised pre-training phase and achieve high accuracy in downstream scene classification tasks, showcasing robust performance and efficacy.

① 基金项目: 国家自然科学基金 (42471475); 地质探测与评估教育部重点实验室主任基金 (GLAB2024ZR06)
收稿时间: 2024-12-22; 修改时间: 2025-01-15; 采用时间: 2025-02-11; csa 在线出版时间: 2025-05-12
CNKI 网络首发时间: 2025-05-13

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Key words: multi-scale feature; detail perception strategy; remote sensing image scene classification; atrous spatial pyramid pooling; Vision Transformer (ViT)

目前, 遥感图像资源在地质勘探、森林资源管理、海洋监测等多个领域中发挥着至关重要的作用[1–3]. 然而, 传统的人工解译方法存在效率低下和精度不足的问题, 已无法满足当前对地观测任务的需求. 随着光学遥感图像数量的不断增加, 如何快速、高效地对其进行加工和处理, 已成为众多研究者关注的焦点. 因此, 科学有效地解译大量遥感图像显得尤为重要[4–6]. 作为解译遥感图像的一种有效手段, 场景分类问题因其固有的挑战性, 持续吸引着众多研究者的关注[7].

遥感图像场景分类任务旨在对给定图像进行标签预测. 为了有效地进行场景分类, 需要对场景中物体和区域之间的关系进行更高层次的理解和表征[8]. 目前, 遥感图像场景分类方法主要可分为两类: 一类是传统的基于人工特征的方法, 另一类是基于深度神经网络学习的方法. 传统的基于人工特征的方法主要依赖于人工构建的特征, 如颜色、纹理、形状和光谱信息等. 这些特征是图像场景中的关键要素, 包含了进行有效分类所需的重要信息. 典型的传统场景分类方法包括颜色直方图[9]、纹理描述符[10–12]、全局图像特征 (global image statistics, GIST)[13]、尺度不变特征变换 (scale-invariant feature transform, SIFT)[14]和方向梯度直方图 (histogram of oriented gradient, HOG)[15]. 尽管这些传统的基于人工特征的方法各具优势, 但也存在一些局限性. 全局特征如颜色直方图、纹理描述符和 GIST 特征等能够直接作为分类器对场景进行分类, 但无法充分考虑局部结构信息; 而局部特征如 SIFT 和 HOG 则代表了局部结构中的特征和形状信息, 但其计算复杂性较高且耗时, 这限制了它们在处理大规模图像数据时的效率.

目前, 深度学习技术凭借其在大规模视觉识别任务中的卓越表现, 已经逐渐渗透并广泛应用于遥感图像处理领域. 与传统方法相比, 深度学习能够自动挖掘复杂的视觉特征, 这是传统基于手工特征提取方法所难以匹敌的. 通过采用不同的网络策略, 各种基于深度学习的遥感场景分类方法相继涌现. 例如, Fang 等人[16]设计了一种用于场景分类的方法, 该方法通过在卷积神经网络中引入频域分支, 使模型能够获得更具判别性的图像特征; He 等人[17]提出了一种新的端到端学习模型——跳跃连接协方差网络, 该网络在卷积神经网络的基础上融合了跳跃连接和协方差池化, 有效降低了模型参数数量, 并提升了场景分类的准确性; 此外, Bazi 等人[18]将 Vision Transformer 引入到遥感图像场景分类领域, 通过修剪一半的层来压缩网络, 同时保证良好的分类精度. 上述基于有监督深度学习的方法已经取得显著成就. 然而, 这些有监督分类方法通常依赖于大量经过标注的数据样本进行模型训练, 而有标注的遥感图像场景分类样本相对稀缺, 且标注成本较高. 近年来, 半监督学习因其能够结合有标签和无标签数据进行模型训练而在遥感图像场景分类中受到广泛关注[19–21]. 例如, Hong 等人[22]提出了一种名为 X-ModalNet 的新型跨模态深度学习框架, 旨在从小规模高光谱图像中提取具有判别性的信息, 并有效地将其转移至大规模多光谱图像或合成孔径雷达数据的分类任务中. Li 等人[23]通过探索海量无标记图像, 提出了一种基于原型一致性的遥感图像场景分类半监督方法, 该方法通过聚焦于前景区域实现判别性特征的提取, 并引入基于原型的分类器以获取一致的特征表示, 最终通过实验证明了该方法的有效性.

虽然上述半监督学习方法在解决遥感图像场景分类的不确定性问题上展现出显著潜力. 但其仍难以摆脱对有标签样本的依赖. 为了更有效地利用未标注数据, 这类模型通常需要构建复杂的结构, 以同时处理有标签和无标签的数据. 这种复杂性可能导致模型的过拟合, 尤其是在有标签样本数量有限的情况下. 此外, 遥感图像场景中存在多地物目标共存、空间结构复杂、尺度差异、类内差异及类间相似等现象, 这些因素对准确提取场景的语义特征表示仍构成挑战. 针对上述问题, 本文提出了一种结合多尺度特征和细节感知策略 (multi-scale features and detail-aware strategy, MFDS) 的 Vision Transformer (MFDS-ViT) 模型, 用于遥感图像场景分类. 该模型通过引入空洞空间金字塔池化模块, 能够有效捕捉和融合遥感图像中的多尺度特征, 同时增强了对局部特征信息的利用. 此外, 本文改进了细节感知掩码策略, 该策略在保留遥感图像中

© 中国科学院软件研究所 http://www.c-s-a.org.cn

关键对象的同时, 通过部分不完整的掩码避免了不可逆的信息丢失. 这一改进使得模型能够有效利用无标签遥感图像数据, 从而促进模型学习更加精细的特征表示, 最终实现更高效、准确的遥感图像场景分类.

1 方法

本文以 Vision Transformer 网络模型作为主干网络, 并融入多尺度特征和细节感知策略, 提出 MFDS-ViT 模型, 其整体结构如图 1 所示. 另外, 图 2 展示了图 1 中编解码的详细结构.

图 2 中的编码器通过引入空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP)[24]模块, 增强了模型对图像中多尺度特征的捕捉能力, 从而更有效地利用局部特征信息, 进一步提升了模型的特征捕捉与判别能力. 模型的训练过程主要分为两个阶段: 首先是自监督预训练阶段, 此阶段利用未标记的数据集进行训练; 其次是下游任务的微调阶段, 对预训练后的模型进行进一步调整. 模型将输入的遥感场景图像进行分区, 然后应用细节感知掩码策略对所选区域进行掩码处理, 经过线性映射的图像块被送入编码器和解码器进行特征提取与掩码图像重建. 在编码部分, 掩码和未掩码的图像块同时输入编码器进行特征提取, 编码器能够提取掩码图像的潜在表示, 并将其作为重建掩码区域的原始信号. 在解码部分, 轻量级解码器依据编码阶段学习到的特征, 恢复原始图像的像素. 完成预训练后, 进入下游微调阶段, 利用编码器和未掩码的图像块作为场景分类任务的骨干网络, 并对预训练权重的整体参数进行微调.

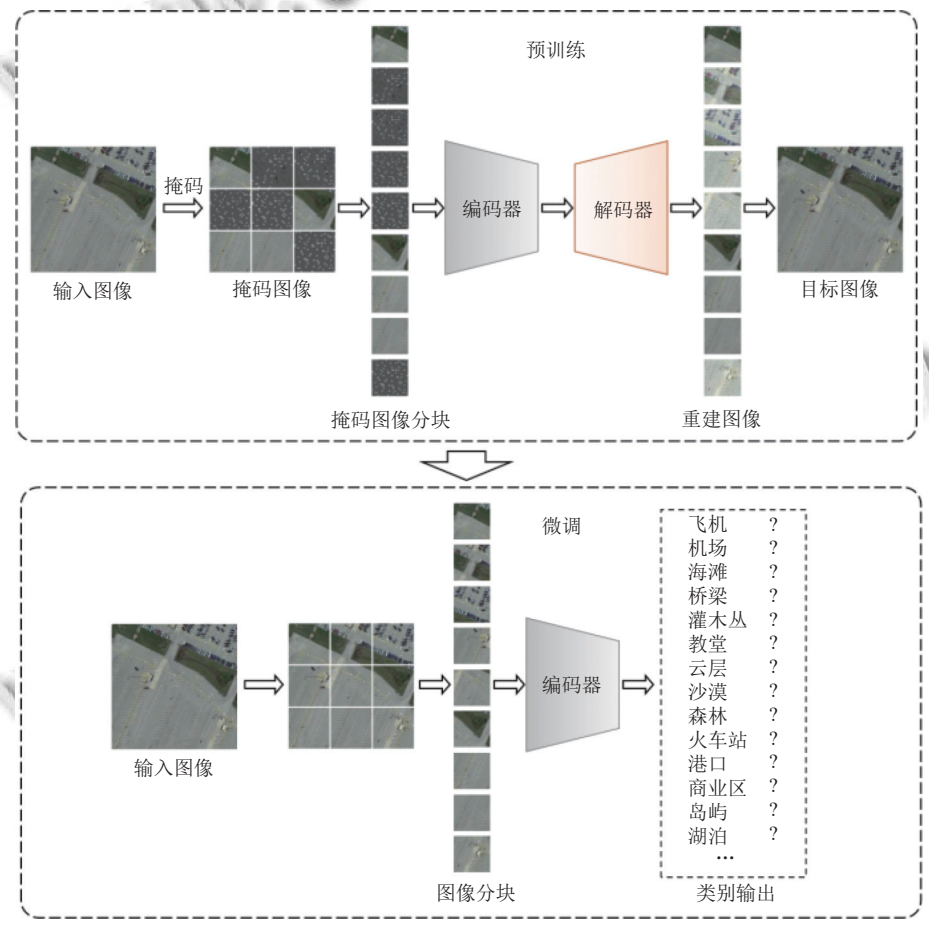

图 1 MFDS-ViT 模型整体框架结构图

1.1 多尺度特征融合模块

在深度学习模型中, 特别是针对图像解译任务, 感受野的扩展对于提升模型对图像的感知能力至关重要. 空洞空间金字塔池化模块正是为此目的而设计, 它通过并行应用不同空间采样率的空洞卷积层以及全局平均池化, 实现了对多尺度特征的有效融合. 如图 2 所示,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

本文提出的 MFDS-ViT 模型中的多尺度特征融合模块由空洞空间金字塔池化模块构成.

图 3 展示了普通卷积与空洞卷积的不同效果. 在图 3 中, 3 幅子图分别呈现了普通卷积和两种不同空洞采样率的空洞卷积, 其中, 红色圆点表示卷积核, 黄色区域代表卷积后的感受野, 最外层黑色线框则表示输入图像. 在图 3(a) 中, 普通卷积 (空洞采样率为 1) 的感受野为 3, 意味着卷积核覆盖了 3×3 的输入区域; 而在图 3(b) 和图 3(c) 中, 通过将空洞采样率设置为 2 和 4, 空洞卷积分别将感受野扩展至 5 和 9. 这表明, 空洞卷积能够在不改变输出特征图尺寸的情况下, 有效增大感受野.

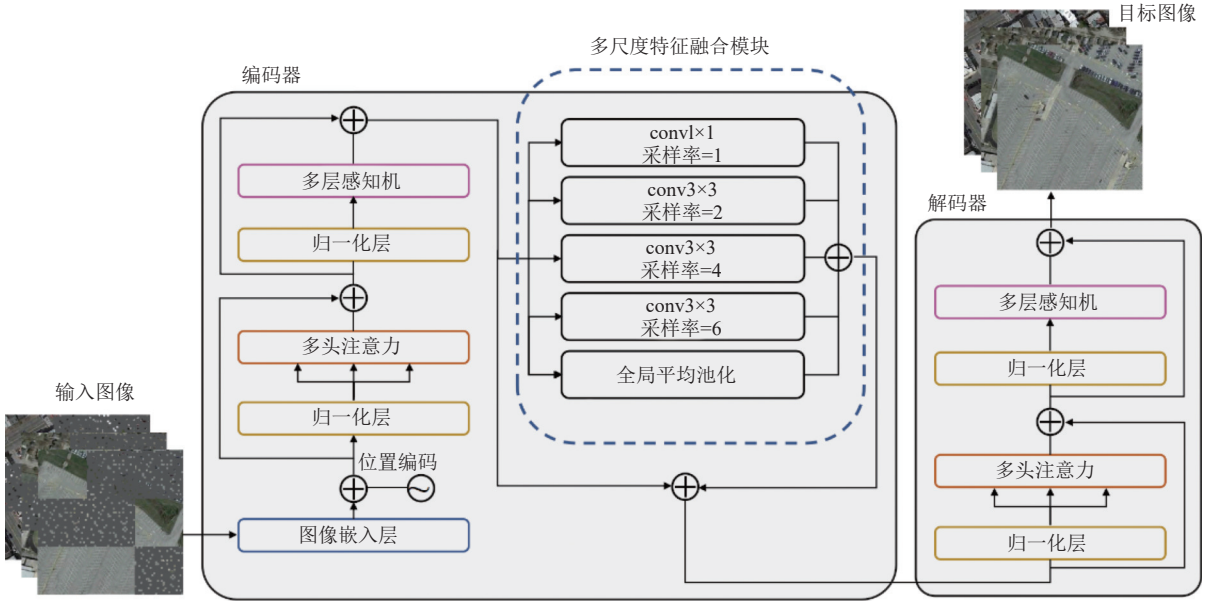

图 2 编解码结构图

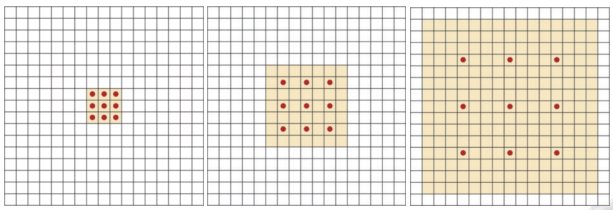

图 3 普通卷积和空洞卷积的对比图

在空洞空间金字塔池化模块的设计中, 空洞卷积的空洞采样率是一个关键参数. 较小的空洞采样率有助于保留更多的局部细节, 而较大的空洞采样率则有助于捕获更广泛的上下文信息. 因此, 如图 4 所示, 本文采用了 4 个具有不同空洞采样率的卷积层, 以及一个全局平均池化层, 以共同对特征图进行采样. 同时, 空洞卷积的空洞采样率由标准的 ASPP (1, 6, 12, 18)[24] 调整为 ASPP (1, 2, 4, 6), 其中括号内的数字代表不同的空洞采样率. 通过这种调整, 模型能够从特征图中提取丰富的多尺度信息. 这些信息随后被送入一个 1×1 卷积层进行融合, 并通过双线性上采样恢复到原始分辨率, 其计算公式如式 (1) 所示.

$$Y = Concat(GAP(X), H_{1,1}(X), H_{2,3}(X), H_{4,3}(X), H_{6,3}(X)) \tag{1}$$

其中, $H_{r,n}(X)$表示对输入特征图X进行空洞卷积操作, 该操作的空洞采样率设为r, 卷积核的尺寸为 n×n. 此外, $GAP(X)$表示通过对X进行全局平均池化处理以获得图像级特征. 经过空洞空间金字塔池化过程, 不仅提升了模型的特征提取能力, 还为后续的场景分类任务提供了更加精确的特征描述.

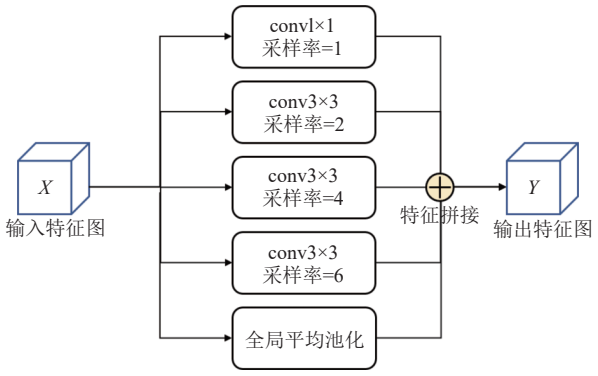

图 4 空洞空间金字塔池化模型图

© 中国科学院软件研究所 http://www.c-s-a.org.cn

当特征经过编码层完成特征提取后, 编码层会产生多个输出. 然而, 在原始的 Vision Transformer 中, 仅选取与分类标记对应的输出作为编码层的最终输出, 用于特征分类. 这种做法导致了许多局部特征信息的丢失, 因为分类标记序列主要用于学习模型中的全局特征表示. 为减少局部特征信息的损失, MFDS-ViT 模型利用了除分类标记序列外的其他标记序列对应的输出特征. 具体而言, 模型将除分类标记序列外的其他标记序列在经过编码层后得到的 768 个 14×14 的特征图输入到多尺度特征融合模块中, 以提取多尺度特征信息, 最终获得 n 个 14×14 的特征图 (n 为类别个数). 随后, 通过 14×14 的卷积核对得到的特征进行降维操作, 以获得分类得分结果. 最终, 将其与分类标记序列对应输出的分类结果进行融合, 完成模型的训练.

1.2 细节感知掩码策略

在大多数掩码图像建模方法中, 常用的掩码策略为随机掩码策略, 如图 5 右侧所示. 传统的随机掩码策略是随机选择一定比例的图像块并进行完全遮罩. 该方法在自然图像的场景分类任务中表现优异, 但在遥感图像的场景分类中却面临诸多问题. 由于遥感图像具有特殊的成像机制, 图像背景更加复杂, 并且存在许多小尺寸物体. 因此, 采用传统的随机掩码策略容易忽视完整的小物体及其细节信息, 针对这一问题, 本文设计了一种细节感知掩码策略, 如图 5 左侧所示.

图 5 本文掩码策略与原随机掩码策略对比

在图 5 中, 右上角红框所标出的区域采用随机掩码策略时, 会导致掩码区域内的小目标信息完全丢失, 这对模型重构小物体目标带来了显著影响, 增加了图像重构的难度. 为此, 本文设计的掩码策略并未对图像块进行完全遮罩, 而是在其中随机保留一些像素信息. 通过这种掩码策略, 可以有效保留小目标的部分像素信息, 从而在模型训练过程中不致于丢失图像中的关键特征. 为确保方法的严谨性, 模型适当增加了掩码图像块的数量, 以保持整体掩码比例不变, 如图 5 中间蓝框区域所示, 原随机掩码策略未对该区域的图像块进行掩码, 而本文提出的掩码策略则增加了该区域的掩码.

首先, 模型将输入图像分割为多个不重叠的图像块. 为充分发挥细节感知掩码策略的优势, 我们将图像块的尺寸设置为 32×32. 然后对这些图像块进行掩码操作. 本文所设计的掩码策略的数学计算过程见式 (2).

$$\begin{aligned} m_{\text{tag}} &= RTG\left(\alpha, len\left(C_{\text{patches}}\right)\right) \\ &= RTG(\alpha, len(\{patch_i\})),\ i = 1,2,3,\cdots,n \end{aligned} \tag{2}$$

如式 (2) 所示, 模型以掩码图像块的比例α和图像块总数$len\left(C_{\text{patches}}\right)$作为输入, 通过随机标记生成器 ($RTG$) 函数确定每个图像块的掩码标签. 因此, 输出的m_{tag}是一个一维向量, 包含了所有图像块的掩码标签.

在获得图像块的掩码标签$m_{\text{tag}}[i]$后, 模型能够判断该图像块是否被部分掩码. 如式 (3) 所示, 模型以掩码内部比β和图像块的尺寸 $dim\ (patch_i)$ 作为输入, 通过随机标记生成器 RTG 函数生成每个图像块的掩码张量$IncompleteMask_i$.

$$IncompleteMask_i = \begin{cases} RTG(\beta, dim(patch_i)), & m_{\text{tag}}[i] = 1 \\ RTG(0, dim(patch_i)), & m_{\text{tag}}[i] = 0 \end{cases} \tag{3}$$

最后, 模型直接将 $patch_i$ 与 $(I - IncompleteMask_i)$ 相乘, 如式 (4) 所示, 其中I表示单位矩阵, ×表示逐元素相乘.

$$patch^i_{\text{mask}} = patch_i \times (I - IncompleteMask_i) \tag{4}$$

2 实验对比与分析

2.1 数据集介绍及实验设置

2.1.1 数据集介绍

在本节实验中, 共使用了 7 个不同的公开遥感图像场景分类数据集, 分别为 AID、NWPU-RESISC45、

© 中国科学院软件研究所 http://www.c-s-a.org.cn

WHU-RS19、SIRI-WHU、RSC11、RSI-CB256[25]、VGoogle[26]和 UC-Merced, 每个数据集的详细介绍见表 1.

表 1 实验数据集属性表

数据集	场景类	每类图像数量	总数量	图像尺寸	空间分辨率 (m)
AID	30	220–420	10000	600×600	0.5–8
NWPU-RESISC45	45	700	31500	256×256	0.2–30
WHU-RS19	19	50–61	1005	600×600	0.5
SIRI-WHU	12	200	2400	200×200	2
RSC11	11	≈100	1232	512×512	0.2
RSI-CB256	35	≈690	24000	256×256	0.3–3
VGoogle	38	1052–1847	59404	256×256	0.075–9.555
UC-Merced	21	100	2100	256×256	0.3

表 1 中, WHU-RS19、SIRI-WHU、RSC11、RSI-CB256 和 VGoogle 数据集被用于自监督学习模型的预训练. 这 5 个数据集均为带标签的公开遥感场景分类数据集, 包含多种不同场景类别及不同空间分辨率的遥感图像, 总计 88041 张. 在预训练阶段, 模型对所有图像进行了统一尺寸调整, 将其缩放至 192×192 像素, 并未使用标签信息. 为便于后续叙述, 本文将该数据集命名为公共遥感图像数据集 (public remote sensing image dataset, PRSID). 此外, 为验证自监督学习方法的性能, 本文选择 AID、NWPU-RESISC45 和 UC-Merced 数据集作为场景分类任务中的训练和测试数据集.

2.1.2 实验设置

为验证本文提出的自监督学习模型 MFDS-ViT 的性能, 首先在无标签的 PRSID 数据集上进行预训练实验, 然后分别在 AID、NWPU-RESISC45 和 UC-Merced 数据集上对模型进行微调, 以完成场景分类实验, 并与其他自监督分类模型进行对比.

在遥感图像场景分类的自监督预训练阶段, 本实验将学习率设置为 1E–2, 批次大小设为 64, 训练迭代次数为 200, 图像的随机掩码比率设定为 0.6, 即在编码阶段, 每幅图像的 60% 区域被随机掩盖, 随后在解码阶段进行重建. 在模型微调阶段, 将学习率设置为 1E–3, 批次大小保持为 64, 训练迭代次数设置为 150. 对于 AID、NWPU-RESISC45 和 UC-Merced 数据集的训练比率, 分别设置为 10% 和 20%, 其余部分用于测试. 本文通过总体分类精度和混淆矩阵评估模型性能. 为确保结果的可靠性, 场景分类过程中均重复进行了 10 次实验, 最终计算 10 次实验的总体分类精度平均值作为模型的分类精度, 并使用在训练过程中的最佳模型计算混淆矩阵.

2.2 图像重建可视化分析

本文构建了一个大规模的无标签公共遥感图像数据集, 并基于该数据集完成了自监督学习模型的训练. 在预训练过程中, 模型通过学习未被掩码的图像区域的特征, 以恢复被掩码的图像块像素. 经过预训练的模型最终可作为特征提取器, 以支持下游场景分类任务的微调.

为了验证模型在图像重建方面的性能, 本文采用细节感知掩码策略对模型的图像重建效果进行了可视化分析. 图像像素重建的质量直接反映了自监督学习框架中编码器特征提取的能力, 同时也指示了该预训练模型作为下游场景分类任务特征提取器的适用性. 下面随机选取了 3 个场景进行示例展示, 如图 6 所示.

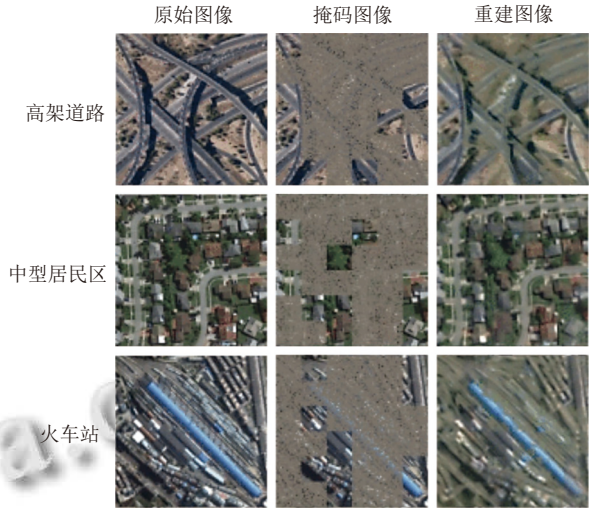

图 6 预训练模型的图像重建可视化示例

图 6 中自上而下依次展示了高架道路、中型居民区和火车站 3 个场景的原始图像、掩码图像和重建图像. 重建结果表明, 模型能够有效恢复被掩码的像素. 通过对比原始图像与重建图像, 我们观察到两者在颜色和纹理上都保持了高度一致性. 尽管重建图像与原始图像存在细微差异, 这恰恰表明本文提出的模型是基于未掩码区域进行图像重建, 而非简单记忆与复制原始图像. 综合分析表明, 本文提出的自监督学习模型能够有效地提取具有区分性的图像特征, 且预训练模型在预测被掩码像素方面表现出了良好的鲁棒性, 为其在下游场景分类任务中的应用奠定了坚实基础.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.3 精度对比分析

为了进一步验证本文所提出方法的有效性与优越性, 将在 AID、NWPU-RESISC45 和 UC-Merced 数据集上与一些先进的自监督方法进行对比. 对比的方法包括: MobileViT[27]、DaViT[28]、EVA[29]、BEiT[30]、MAE[31]和 MixMAE[32]. 下面将对 AID、UC-Merced 和 NWPU-RESISC45 数据集上的检测结果进行详细的对比分析.

2.3.1 AID 数据集实验结果分析

表 2 详细列出了在 AID 数据集的 10% 和 20% 训练比率条件下, MobileViT、DaViT、EVA、BEiT、MAE、MixMAE 和 MFDS-ViT 方法的分类准确率. 结果显示, MFDS-ViT 在这两个训练比率下均实现了较高的分类准确率, 分别为 93.58% 和 96.15%, 较高于其他自监督场景分类方法, 同时也表明 MFDS-ViT 在处理有限训练数据时展现出良好的鲁棒性和有效性.

表 2 不同自监督方法在 AID 数据集上采用不同训练比例的分类结果 (%)

方法	10%	20%
MobileViT	83.11	88.78
DaViT	92.08	94.05
EVA	86.06	89.26
BEiT	89.79	91.21
MAE	85.16	91.69
MixMAE	91.53	94.23
MFDS-ViT	**93.58**	**96.15**

MAE 和 BEiT 方法均采用了原始的随机掩码策略, 而所提出的 MFDS-ViT 则引入了一种细节感知掩码策略. 对比结果显示, 在 10% 的训练比率下, MFDS-ViT 的分类准确率比 MAE 方法高出 8.42 个百分点, 较 BEiT 也提升了 3.79 个百分点; 在 20% 的训练比率下, MFDS-ViT 相较于 MAE 提高了 4.46 个百分点, 与 BEiT 相比同样领先 4.94 个百分点. 整体来看, MFDS-ViT 在各个训练比率下均取得了较高的分类准确率, 进一步验证了本文所提出掩码策略的有效性. 值得注意的是, MixMAE 作为 2023 年提出的新方法, 在计算机视觉领域获得了广泛认可, 在 10% 的训练比率下, 其分类精度达到了 91.53%, 而在 20% 的训练比率下达到了 94.23%. 然而, MFDS-ViT 在 10% 的训练比率下相比 MixMAE 高出 2.05 个百分点, 在 20% 的训练比率下也高出 1.92 个百分点. 这不仅证明了 MFDS-ViT 的优越性, 同时也表明其在遥感图像分类任务中的潜力和可靠性.

MFDS-ViT 在 AID 数据集上场景分类的混淆矩阵如图 7 所示. 结果表明, MFDS-ViT 能够有效识别大部分场景类别, 整体分类准确率超过 90%, 特别是在沙漠、森林、山脉、停车场和高架道路等场景类别中, 其分类准确率达到了 100%. 其余大部分场景类别的识别准确率也均超过了 90%, 仅有 3 个场景类别的分类精度低于 90%, 分别为公园 (87%)、度假胜地 (80%) 和学校 (86%). 另外, 从图 7 中可以看出, 5% 的学校被错误分类为商业区, 3% 的学校被误判为公园, 5% 的公园被识别为度假胜地, 而 10% 的度假胜地则被错误归类为公园. 这种错误分类现象的发生, 主要是由于这些场景中普遍存在树木、草地及稀疏建筑物, 导致其在光谱特征和空间布局上具有一定的相似性, 从而增加了识别的难度.

2.3.2 NWPU-RESISC45 数据集实验结果分析

不同自监督方法在 NWPU-RESISC45 数据集上的场景分类结果如表 3 所示. MFDS-ViT 在两个训练比率下均取得了较高的分类准确率, 分别为 92.61% 和 94.71%. 这表明其在遥感图像场景分类任务中具有较为显著的优势. 另外, MobileViT 在 10% 的训练比率下的准确率为 80.89%, 在 20% 的训练比率下提高至 87.15%. DaViT 和 EVA 方法的表现相对接近, 分别在两个训练比率下达到 88.32% 和 91.04% 以及 88.01% 和 90.27%. 采用原始随机掩码策略的 BEiT 方法在 10% 的训练比率下的准确率为 85.56%, 在 20% 的训练比率下则达到了 90.60%. 同样采用原始随机掩码策略的 MAE 方法在两个训练比率下的表现分别为 83.69% 和 90.90%. 而采用混合掩码策略的 MixMAE 方法在两种训练比率下的分类准确率分别为 90.29% 和 92.69%. 具体而言, 在 10% 的训练比率下, MFDS-ViT 比 MAE 方法高出 8.92 个百分点, 比 MixMAE 方法高 2.32 个百分点. 在 20% 的训练比率下, MFDS-ViT 比 MAE 方法高 3.81 个百分点, 比 MixMAE 方法也高出 2.02 个百分点. 这表明 MFDS-ViT 在处理遥感图像数据时, 不仅在分类准确率上有所提升, 而且在不同训练比率下均能保持较高的稳定性和性能, 即便在标签样本较少的情况下, 仍能确保良好的检测效果.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

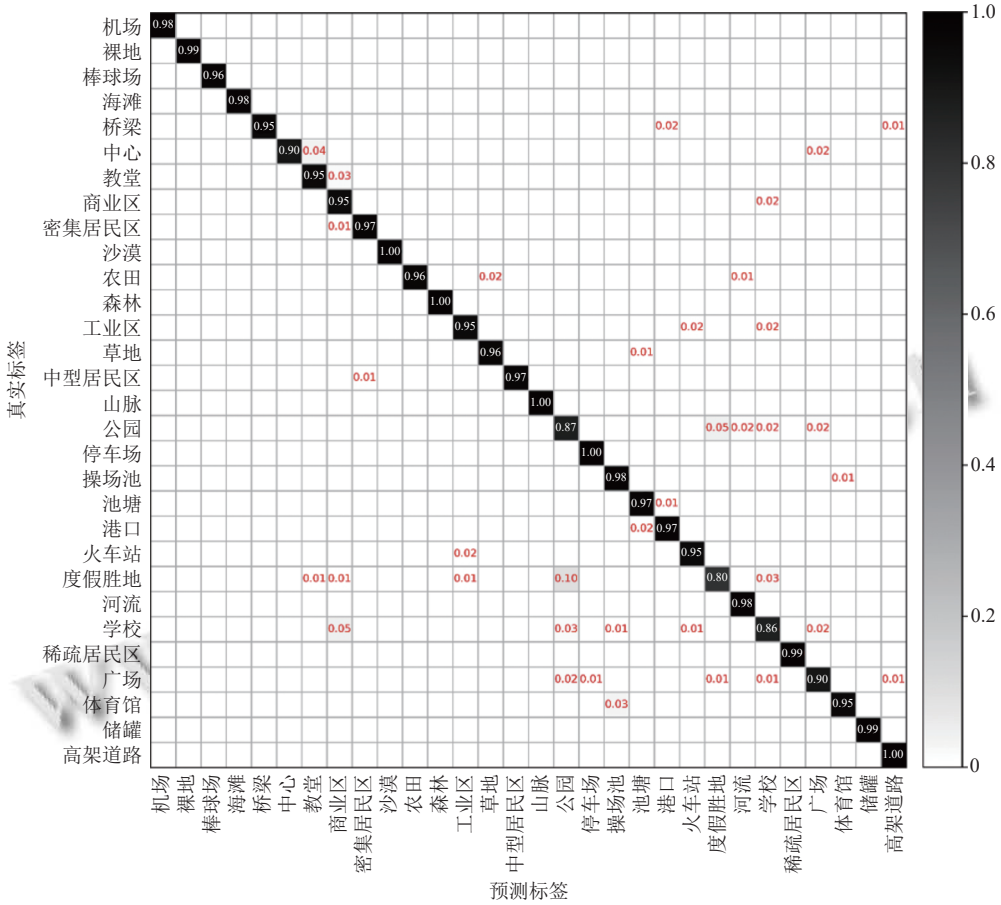

图 7 所提出的 MFDS-ViT 在 20% 的 AID 数据集上的分类混淆矩阵

表 3 不同自监督方法在 NWPU-RESISC45 数据集上采用不同训练比例的分类结果 (%)

方法	10%	20%
MobileViT	80.89	87.15
DaViT	88.32	91.04
EVA	88.01	90.27
BEiT	85.56	90.60
MAE	83.69	90.90
MixMAE	90.29	92.69
MFDS-ViT	**92.61**	**94.71**

图 8 展示了 MFDS-ViT 在 20% 的 NWPU-RESISC45 训练数据集上场景分类的混淆矩阵. 结果表明, MFDS-ViT 能够有效识别大部分场景类别, 其中有 40 类场景的分类准确率超过 90%. 但是, 个别场景类别的识别准确率相对较低, 例如教堂 (83%)、高速公路 (85%)、中型居民区 (88%)、宫殿 (82%) 和网球场 (88%). 具体而言, 9% 的教堂被识别为宫殿, 9% 的宫殿被错误归类为教堂, 这主要是由于教堂与宫殿在建筑结构上有许多相似之处. 4% 的山脉被错误判定为沙漠, 原因在于植被稀疏的山脉与沙漠在颜色和纹理上的相似性. 3% 的火车站被判定为铁路, 4% 的铁路被判定为火车站, 这主要是因为火车站周围通常存在密集的铁路设施, 导致两者易于混淆. 此外, 7% 的中型居民区被错误判定为密集型居民区, 2% 的密集型居民区被归类为中型居民区. 这一现象的发生, 主要是由于中型居民区与密集型居民区之间缺乏明确的判定界限, 往往依赖人为主观判断, 从而导致这两类场景容易被误分类.

2.3.3 UC-Merced 数据集实验结果分析

表 4 汇总了在 10% 和 20% 的训练比率下, 多个自监督学习方法在场景分类任务中的准确率. 所涉及的方法包括 MobileViT、DaViT、EVA、BEiT、MAE、MixMAE 以及我们提出的 MFDS-ViT.

首先, MFDS-ViT 在 10% 和 20% 训练比率下分别取得了 93.81% 和 97.48% 的分类准确率, 显示出其在有限有标签训练数据中有效提取有用特征的能力. 这一结果进一步验证了 MFDS-ViT 在自监督学习框架下的潜力与适用性. 其次, 与其他自监督学习方法的分类性能进行对比时, DaViT 在 20% 训练样本比率下达到

© 中国科学院软件研究所 http://www.c-s-a.org.cn

了 96.71% 的准确率, 表现与 MFDS-ViT 相近; EVA 和 BEiT 在两个训练样本比率下均展现出良好的性能, 准确率分别为 93.80% 和 94.96%. 在 10% 训练样本比率下, MAE 的准确率为 87.32%, 而在 20% 训练样本比率下显著提升至 93.05%, 这一显著提升表明 MAE 在数据量增加时具有更好的泛化能力, 但也反映出其在处理有限遥感图像数据时的鲁棒性不足. 此外, 采用混合掩码策略的 MixMAE 在 10% 训练样本比率下的准确率为 92.14%, 在 20% 训练样本比率下提升至 96.42%. 尽管 MixMAE 在 20% 训练样本比率下的表现接近 MFDS-ViT, 但 MFDS-ViT 在两个训练样本比率下均保持了更高的准确率. 总体而言, 通过对比分析, MFDS-ViT 在 UC-Merced 数据集上的表现优于其他自监督学习方法, 进一步验证了其有效性.

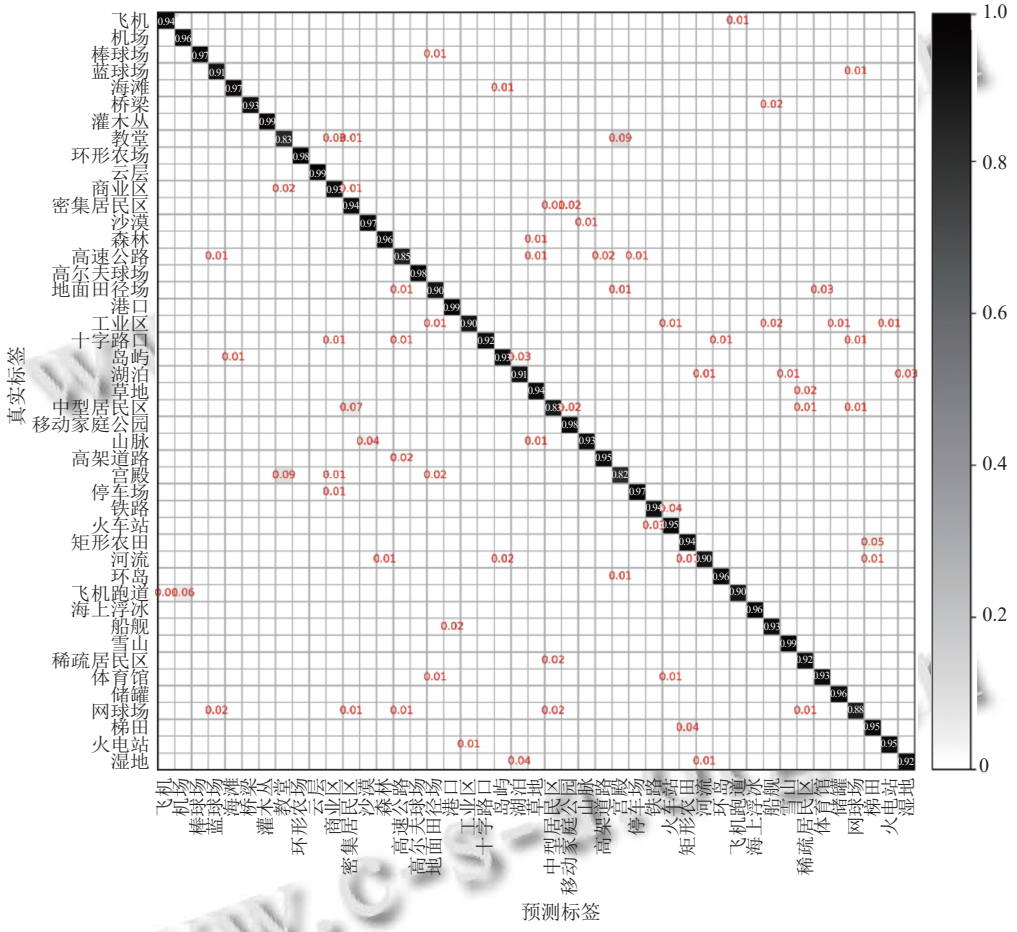

图 8 MFDS-ViT 在 20% 的 NWPU-RESISC45 训练数据集上的分类混淆矩阵

表 4 不同自监督方法在 UC-Merced 数据集上采用不同训练比例的分类结果 (%)

方法	10%	20%
MobileViT	80.48	90.31
DaViT	92.92	96.71
EVA	89.76	93.80
BEiT	91.07	94.96
MAE	87.32	93.05
MixMAE	92.14	96.42
MFDS-ViT	**93.81**	**97.48**

图 9 展示了在训练比率为 20% 时, MFDS-ViT 在 UC-Merced 数据集上进行场景分类的混淆矩阵. 从图 9 可以看出, 近一半的场景类别的分类准确率达到了 100%, 具体包括农业区、灌木丛、森林、高速公路、高尔夫球场、港口、河流和飞机跑道. 然而, 仅有一个场景类别的整体准确率低于 90%, 即中型居民区, 其分类准确率为 80%. 进一步分析显示, 有 2% 的十字路口被误分类为高架道路, 这两类场景在遥感图像中均表现为道路交叉, 且在光谱和空间布局上具有相似性, 从而导致误判. 此外, 6% 的中型居民区被误判为密集型居民区, 4% 的中型居民区被误判为稀疏型居民区, 4% 的密集型居民区被误分类为中型居民区, 4% 的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

建筑群被误分类为密集型居民区, 另有 4% 的建筑群被误判为稀疏型居民区. 上述误分类现象与第 2.3.2 节在 NWPU-RESISC45 数据集中提到的原因类似, 建筑群、稀疏型居民区、中型居民区和密集型居民区在类别划分时缺乏明确的界定标准, 导致模型在识别过程中可能产生错误判断, 从而影响最终的分类效果.

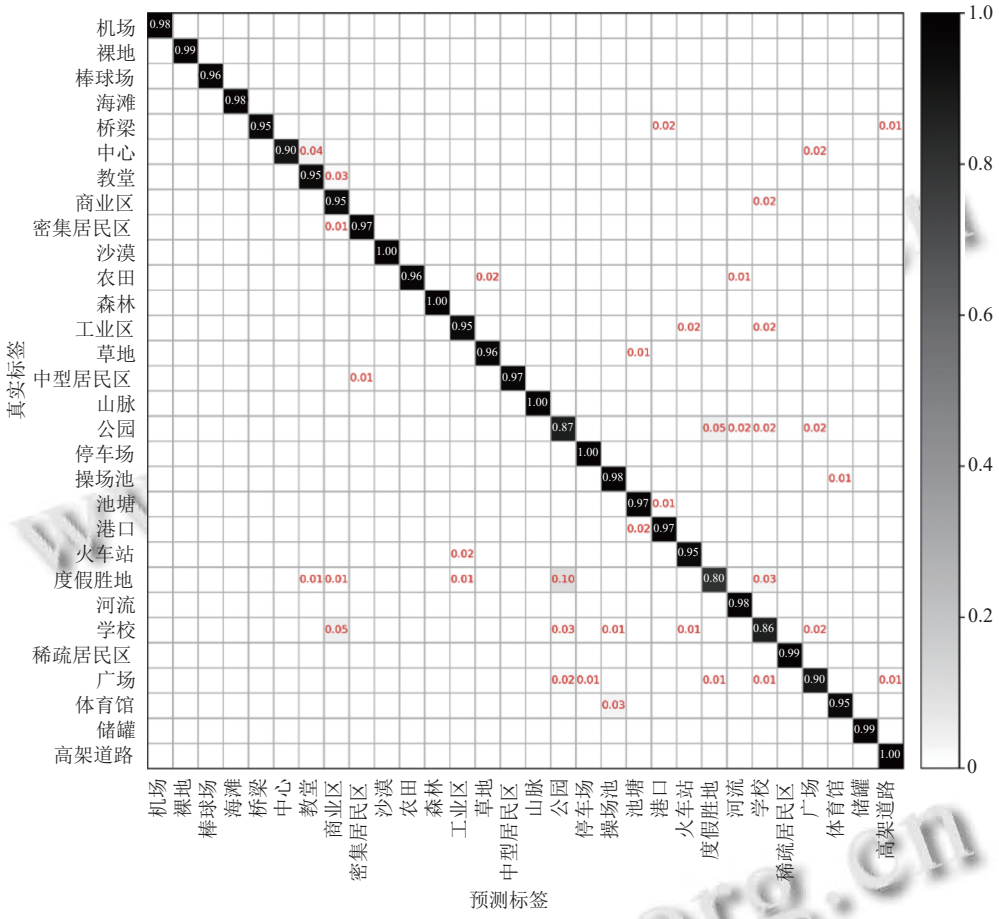

图 9 MFDS-ViT 在 20% 的 UC-Merced 训练数据集上的分类混淆矩阵

2.4 消融实验

为验证和评估所提出的 MFDS-ViT 中各部分对分类精度的影响, 本节设计了 4 个消融实验方案, 分别为 Test1、Test2、Test3 和 Test4. 其中, Test1 采用原始随机掩码策略, 并使用 Vision Transformer 作为特征编码器; Test2 同样采用原始随机掩码策略, 但将所提出的 MFDS-ViT 中的编码器作为特征编码器, 解码器部分保持不变; Test3 则采用细节感知掩码策略, 同时使用 Vision Transformer 作为特征编码器, 解码器部分保持不变; Test4 采用细节感知掩码策略, 将 MFDS-ViT 中的编码器作为特征编码器, 解码器部分保持不变. 然后, 在 AID、NWPU-RESISC45 和 UC-Merced 数据集上以不同的训练比例进行实验. 表 5 展示了本文模型中各部分改进对场景分类结果的影响.

表 5 消融实验结果 (%)

方案	AID		NWPU-RESISC45		UC-Merced	
	10%	20%	10%	20%	10%	20%
Test1	85.16	91.69	83.69	90.90	87.32	93.05
Test2	88.34	93.03	87.94	92.44	89.06	94.80
Test3	90.63	94.34	89.12	92.98	90.72	95.93
Test4 (本文方法)	**93.58**	**96.15**	**92.61**	**94.71**	**93.81**	**97.48**

实验结果表明, 在 Test2 中, 采用 MFDS-ViT 中的编码器作为特征编码器后, 模型在不同数据集和训练比例下的分类精度提高了 1.34%–4.25%. 这一结果表明, 增强编码器的特征提取能力能够有效改善场景分类效果. 然而, 仅通过增强编码器的特征提取能力并不足以实现较好的场景分类效果, 分类精度仍有进一步提升的空间. 这主要是因为模型中原始的随机掩码策略主要针对自然图像场景设计, 而遥感图像场景中存

© 中国科学院软件研究所 http://www.c-s-a.org.cn

在大量小且密集的物体, 且背景复杂度较高. 原随机掩码策略可能导致许多不可逆的信息丢失, 从而对最终的训练效果产生负面影响.

通过对比 Test3 和 Test1 的实验结果, 可以明显看出, 采用细节感知掩码策略后, 模型的分类精度得到了显著提升. 在 10% 和 20% 的 AID 数据集上, 分类精度分别提高了 5.47% 和 2.65%; 在 10% 和 20% 的 NWPU-RESISC45 数据集上, 分类精度分别提高了 5.43% 和 2.08%; 在 10% 和 20% 的 UC-Merced 数据集上, 分类精度分别提高了 3.4% 和 2.88%. 以上实验结果充分表明, 本文提出的细节感知掩码策略在遥感图像场景分类任务中具有良好的有效性.

进一步对比 Test4 和 Test1 的实验结果显示, 在 10% 的 AID 数据集上, 分类精度较改进前提高了 8.42%; 在 20% 的 AID 数据集上, 分类精度提高了 4.46%; 在 10% 的 NWPU-RESISC45 数据集上, 分类精度较改进前提高了 8.92%; 在 20% 的 NWPU-RESISC45 数据集上, 分类精度提高了 3.81%; 在 10% 的 UC-Merced 数据集上, 分类精度较改进前提高了 6.48%; 在 20% 的 UC-Merced 数据集上, 分类精度提高了 4.43%. 上述实验结果表明, 所提出的 MFDS-ViT 模型通过混合使用多尺度特征模块和细节感知掩码策略, 能够高效处理遥感图像场景分类任务, 并显著提升分类精度. 尤其是在 10% 的训练比率下, 分类精度的提升幅度较大, 进一步证明了 MFDS-ViT 在处理少量有标签样本时的良好泛化性.

3 结论

本文提出一种结合多尺度特征和细节感知策略的模型 (MFDS-ViT) 用于遥感图像场景分类. MFDS-ViT 中引入了细节感知掩码策略, 该策略在保留遥感图像中关键对象的同时, 通过部分不完整的掩码有效避免了不可逆的信息丢失, 因而更适合于遥感图像场景分类任务. 另外, MFDS-ViT 中还引入了空洞空间金字塔池化模块, 能够有效捕捉和融合遥感图像中的多尺度特征, 同时增强了模型对局部特征信息的利用. 实验结果表明, MFDS-ViT 在自监督预训练阶段能够有效提取图像特征, 并在下游场景分类任务中实现较高的分类准确率. 与其他自监督学习方法相比, MFDS-ViT 在处理有限训练数据集时展现出良好的有效性和鲁棒性.

参考文献

1 李世龙. 遥感图像解译在铁路沿线地质灾害调查中的应用研究. 资源信息与工程, 2018, 33(2): 163–164. [doi: 10.3969/j.issn.2095-5391.2018.02.078]

2 杨帆. 遥感影像判读在洮河国家级自然保护区森林资源二类调查中的应用. 南方农业, 2023, 17(24): 111–113.

3 李晓威, 范儒彬, 马荣华, 等. 空天地海一体化海洋监测体系研究. 科技与创新, 2023(10): 145–148, 153.

4 宝音图. 基于深度学习的光学遥感图像场景分类与语义分割 [硕士学位论文]. 郑州: 战略支援部队信息工程大学, 2021.

5 Cheng G, Xie XX, Han JW, *et al*. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3735–3756. [doi: 10.1109/JSTARS.2020.3005403]

6 丁鹏. 基于深度卷积神经网络的光学遥感目标检测技术研究 [博士学位论文]. 长春: 中国科学院大学 (中国科学院长春光学精密机械与物理研究所), 2019.

7 Zhao XM, Wang HJ, Wu J, *et al*. Remote sensing image segmentation using geodesic-kernel functions and multi-feature spaces. Pattern Recognition, 2020, 104: 107333. [doi: 10.1016/j.patcog.2020.107333]

8 Cheng G, Han JW, Lu XQ. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 2017, 105(10): 1865–1883. [doi: 10.1109/JPROC.2017.2675998]

9 Swain MJ, Ballard DH. Color indexing. International Journal of Computer Vision, 1991, 7(1): 11–32. [doi: 10.1007/BF00130487]

10 Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC-3(6): 610–621.

11 Jain AK, Ratha NK, Lakshmanan S. Object detection using gabor filters. Pattern Recognition, 1997, 30(2): 295–309. [doi: 10.1016/S0031-3203(96)00068-4]

12 Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971–987. [doi: 10.1109/TPAMI.2002.1017623]

13 Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 2001, 42(3): 145–175. [doi: 10.1023/A:1011139631724]

14 Lowe DG. Distinctive image features from scale-invariant

© 中国科学院软件研究所 http://www.c-s-a.org.cn

keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

15 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893.

16 Fang J, Yuan Y, Lu XQ, *et al*. Robust space-frequency joint representation for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(10): 7492–7502. [doi: 10.1109/TGRS.2019.2913816]

17 He NJ, Fang LY, Li ST, *et al*. Skip-connected covariance network for remote sensing scene classification. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(5): 1461–1474.

18 Bazi Y, Bashmal L, Al Rahhal MM, *et al*. Vision Transformers for remote sensing image classification. Remote Sensing, 2021, 13(3): 516. [doi: 10.3390/rs13030516]

19 夏英, 李骏垚, 郭东恩. 基于 GAN 的半监督遥感图像场景分类. 光子学报, 2022, 51(3): 0310003.

20 郭东恩, 吴泽琛. 基于生成对抗网络的半监督遥感图像场景分类. 南阳理工学院学报, 2022, 14(6): 53–59.

21 周国华, 蒋晖, 顾晓清, 等. 基于半监督子空间迁移的稀疏表示遥感图像场景分类方法. 浙江大学学报 (理学版), 2021, 48(6): 684–693.

22 Hong DF, Yokoya N, Xia GS, *et al*. X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 167: 12–23. [doi: 10.1016/j.isprsjprs.2020.06.014]

23 Li Y, Li Z, Wang Z, *et al*. Semi-supervised remote sensing image scene classification with prototype-based consistency. Chinese Journal of Aeronautics, 2024, 37(2): 459–470.

24 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.

25 Li HF, Dou X, Tao C, *et al*. RSI-CB: A large scale remote sensing image classification benchmark via crowdsource data. arXiv:1705.10450, 2020.

26 Hou DY, Miao ZL, Xing HQ, *et al*. V-RSIR: An open access web-based image annotation tool for remote sensing image retrieval. IEEE Access, 2019, 7: 83852–83862. [doi: 10.1109/ACCESS.2019.2924933]

27 Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision Transformer. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

28 Ding MY, Xiao B, Codella N, *et al*. DaViT: Dual attention vision Transformers. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 74–92.

29 Fang YX, Wang W, Xie BH, *et al*. EVA: Exploring the limits of masked visual representation learning at scale. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 19358–19369.

30 Bao HB, Dong L, Piao S, *et al*. BEiT: BERT pre-training of image Transformers. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

31 He KM, Chen XL, Xie SN, *et al*. Masked autoencoders are scalable vision learners. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15979–15988.

32 Liu JH, Huang X, Zheng JL, *et al*. MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision Transformers. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 6252–6261.

(校对责编: 张重毅)

© 中国科学院软件研究所 http://www.c-s-a.org.cn