遥感场景理解中视觉Transformer的参数高效微调

尹文昕[①③] 于海琛*[①②③] 刁文辉[①③] 孙 显[①②③] 付 琨[①③]

[①](中国科学院空天信息创新研究院 北京 100190)

[②](中国科学院大学电子电气与通信工程学院 北京 100049)

[③](中国科学院空天信息创新研究院网络信息体系技术科技创新重点实验室 北京 100190)

摘 要：随着深度学习和计算机视觉技术的飞速发展，遥感场景分类任务对预训练模型的微调通常需要大量的计算资源。为了减少内存需求和训练成本，该文提出一种名为“多尺度融合适配器微调(MuFA)”的方法，用于遥感模型的微调。MuFA引入了一个多尺度融合模块，将不同下采样倍率的瓶颈模块相融合，并与原始视觉Transformer模型并联。在训练过程中，原始视觉Transformer模型的参数被冻结，只有MuFA模块和分类头会进行微调。实验结果表明，MuFA在UCM和NWPU-RESISC45两个遥感场景分类数据集上取得了优异的性能，超越了其他参数高效微调方法。因此，MuFA不仅保持了模型性能，还降低了资源开销，具有广泛的遥感应用前景。

关键词：遥感图像；场景分类；参数高效；深度学习

中图分类号：TN919.81 **文献标识码**：A **文章编号**：1009-5896(2024)09-3731-08

DOI: 10.11999/JEIT240218

Parameter Efficient Fine-tuning of Vision Transformers for Remote Sensing Scene Understanding

YIN Wenxin[①③] YU Haichen[①②③] DIAO Wenhui[①③] SUN Xian[①②③] FU Kun[①③]

[①](*Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

[②](*School of Electronic*, *Electrical and Communication Engineering*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

[③](*Key Laboratory of Network Information System Technology* (*NIST*), *Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: With the rapid development of deep learning and computer vision technologies, fine-tuning pre-trained models for remote sensing tasks often requires substantial computational resources. To reduce memory requirements and training costs, a method called “Multi-Fusion Adapter (MuFA)” for fine-tuning remote sensing models is proposed in this paper. MuFA introduces a fusion module that combines bottleneck modules with different down sample rates and connects them in parallel with the original vision Transformer model. During training, the parameters of the original vision Transformer model are frozen, and only the MuFA module and classification head are fine-tuned. Experimental results demonstrate that MuFA achieves superior performance on the UCM and NWPU-RESISC45 remote sensing scene classification datasets, surpassing other parameter efficient fine-tuning methods. Therefore, MuFA not only maintains model performance but also reduces resource overhead, making it highly promising for various remote sensing applications.

Key words: Remote sensing; Scene classification; Parameter efficient; Deep learning

1 引言

随着深度学习和计算机视觉技术的飞速发展，深度神经网络已成为遥感场景分类任务中的主要方法[1–3]。遥感场景分类是通过遥感数据识别给定区域的地表覆盖或土地利用类型。遥感场景的背景复杂、类型多变，类内差异较大，但部分类别之间又存在高度相似性，因此，具备强大特征提取能力的深度神经网络非常适用于遥感场景分类。

目前，主流的遥感场景分类方法主要基于两种模型：卷积神经网络(Convolutional Neural Net-

收稿日期：2024-03-29；改回日期：2024-07-17；网络出版：2024-08-02
*通信作者： 于海琛 yuhaichen18@mails.ucas.ac.cn
基金项目：国家重点研发计划(2022ZD0118401)
Foundation Item: The National Key R&D Program of China (2022ZD0118401)

work, CNN)模型和视觉Transformer模型。其中，CNN模型，如Krizhevsky等人[4]提出的深度卷积神经网络(AlexNet)、He等人[5]提出的残差卷积神经网络(ResNet)等主要使用大量可学习的2维卷积算子组合而成。这种2维卷积设计充分利用了2维图像的归纳偏置，因此能够很高效地进行训练。CNN模型以捕捉细节的纹理特征见长，对全局性的语义特征理解能力较弱，因此，在近些年的实践中，CNN模型正逐步被视觉Transformer模型取代。视觉Transformer模型如视觉变换器模型(Vision Transformer, ViT)[6]、滑窗变换器模型(Swin Transformer)[7]等利用自注意力机制，能够对图像的全局信息进行综合性理解，因此具有很强的阅读语义信息的能力。在多种自然场景的应用中，如图像分类、目标检测等任务中，视觉Transformer模型的性能已经全面超过了CNN模型。

视觉Transformer模型不像CNN模型那样利用2维图像先验的归纳偏置，因此需要首先使用大规模数据进行预训练。当前主要使用的是自然场景数据集ImageNet[8]预训练的视觉Transformer模型，为了使其能够应用在遥感场景分类任务中，这些预训练模型还将在特定的遥感数据集上进行微调，并评估分类效果。

尽管视觉Transformer已经经过有效训练，但微调这些模型仍然需要大量计算资源，因为它们通常具有千万级别的参数。在这样的背景下，寻找一种仅微调少量参数的高效方法变得至关重要。在计算机视觉领域，这类方法被称为参数高效微调或参数高效迁移学习。经典方法包括偏差微调(BitFit)[9]、适配器微调(Adapter)[10]、低秩分解微调(Low-Rank Adaptation, LoRA)[11]、低秩矩阵适配器微调(LoRand)[12]、并联适配器微调(AdaptFormer)[13]、卷积旁路微调(Convpass)[14]等。然而，这些方法通常仅使用可学习的简单模块，如Adapter使用简单的下采样-上采样瓶颈结构，Convpass则使用下采样-2维卷积-上采样的结构。在复杂的遥感场景分类任务中，这样的简单结构不能有效利用遥感数据复杂的特征信息，因此难以达到最优效果。

为提升在遥感场景分类任务中视觉Transformer模型的迁移效率，在仅微调少量参数的前提下尽可能保持模型的分类能力，本文提出一种基于多尺度融合的适配器微调方法(Multi-Fusion Adapter, MuFA)。MuFA的融合模块将不同下采样倍率的瓶颈模块相融合，并与原有模型进行并联。不同下采样倍率的瓶颈模块会专注于不同维度的图像特征，下采样倍率较低的瓶颈模块会更注重图像的纹理特征信息，而更高倍率的瓶颈模块则侧重于图像的语义信息。因此，将不同下采样倍率的模块进行融合，不同类型的特征信息也能够随之结合，并最终体现为模型分类效果的提升。因此，本文的主要贡献如下：

(1) MuFA是一种参数高效微调方法，能够在仅微调少量模型参数的基础上，实现从自然场景的预训练模型到遥感场景的专用场景分类模型的迁移，节约了迁移学习过程的计算资源。

(2) MuFA通过融合不同下采样倍率的瓶颈模块，实现了纹理特征和语义特征的融合，从而有效地提升了模型的分类效果。

(3)在多个遥感场景分类任务数据集上的实验结果表明，MuFA可以在减少微调参数、节省计算资源的同时，拥有较好的分类能力，优于其他对比方法。

2 相关工作

2.1 视觉神经网络

在视觉深度神经网络领域，许多经典和创新性的模型被提出，为图像分类、目标检测和语义分割等任务带来了显著的进展。首先是卷积神经网络(CNN)模型，其最具代表性的AlexNet是深度学习领域的先驱之一，由Krizhevsky等人[4]于2012年提出。它是第1个在ImageNet[8]图像分类挑战中取得显著成功的卷积神经网络。AlexNet采用了深层卷积层和池化层，以及Dropout等技术，为后续模型的发展铺平了道路。超深度卷积神经网络(Very deep convolutional networks, VGG)是由Simonyan等人[15]于2014年提出的CNN模型。它以其简洁的架构和深度堆叠的卷积层而闻名。VGG的核心思想是使用相同大小的卷积核，使网络更易于理解和实现。ResNet是由He等人[5]于2016年提出的深度残差网络。它通过引入残差块，解决了深层网络训练中的梯度消失问题。ResNet的结构使得网络可以轻松地堆叠数百个卷积层，成为图像分类任务中的标杆模型。

随着Transformer[16]模型在自然语言处理领域得到广泛应用，一系列用于计算机视觉任务的视觉Transformer模型也被相继提出。ViT是一种基于自注意力机制的视觉Transformer模型，由Dosovitskiy等人[6]于2021年提出。它将图像划分为小块(patch)，并使用Transformer的自注意力机制来建模全局上下文信息。ViT在ImageNet分类任务上表现出色，为视觉领域引入了Transformer的思想。Swin Transformer是一种基于滑动窗口注意力机制的Transformer模型，由Liu等人[7]于2021年提出。

它在处理高分辨率图像时具有高效性，并在多个视觉任务上取得了优异的性能。

2.2 参数高效微调

近年来，深度学习的快速发展导致模型参数数量显着增加。预训练模型通常需要迁移学习来适应特定任务。然而，将迁移学习应用于这些任务的大规模模型已被证明是一项艰巨的挑战。为了解决这个问题，自然语言处理(Natural Language Processing, NLP)领域最初提出了一系列参数高效微调方法。Adapter[10]引入了层之间可训练的瓶颈结构以增强模型适应性，BitFit[9]有选择地调整了一些在任务适应中发挥关键作用的关键偏差项，提示微调(Prompt-Tuning)[17]将可学习的标记添加到输入层，以便为模型提供明确的指导，而LoRA[11]将可训练的秩分解矩阵注入到 Transformer 架构的每个注意力层中。在各种NLP 任务中，与完全微调相比，参数高效微调方法始终表现出卓越的性能。 这种现象也成为计算机视觉(Computer Vision, CV)领域进一步研究的灵感来源。

参数高效微调在NLP领域的成功启发了计算机视觉(CV)领域的研究人员。受 Prompt-Tuning 的启发，视觉提示微调(Visual Prompt-Tuning, VPT)[18]开发了视觉提示模块，这种视觉提示模块由一组简单的可训练的提示token(即提示模块)组成，并应用在视觉Transformer模型上。考虑到VPT的设计仍然偏向NLP的提示token设计方法，没有考虑视觉任务的独特性，因此，在此基础上，在视觉领域诞生了一系列基于VPT的改进方法。改进视觉提示微调(Improving Visual Prompt-Tuning, IVPT)[19]提出不同层的提示模块应当存在一定关联性，并增加了不同层之间提示token的连接和交互。视觉查询微调(Visual Query-Tuning, VQT)[20]注意到提示模块与模型的中间查询特征可以进行交互，因此选择在提示模块上额外聚合视觉Transformer的Query特征。有效且高效的视觉提示微调(Effective and Efficient Visual Prompt Tuning, E^{2}VPT)[21]则为视觉提示附加了Key约束，在自注意力层和输入层的视觉提示中额外引入了一组可学习的Key值提示。

此外，在另一种经典的高效微调方法Adapter的基础上，一系列全新的视觉Adapter方法也被相继引入。AdaptFormer[13]在视觉任务上使用了并行的适配器结构，并从理论和实验两方面证明并行的适配器结构学习速度更快、效果更好。LoRand[12]使用低秩融合机制创建了紧凑的适配器结构，这种低秩融合的设计进一步降低了微调参数量，改善了微调效果。克罗内克适配器微调(Kronecker Adaptation, KAdaptation)[22]使用Kronecker乘积法进一步分解微调参数，实现了微调参数量的进一步降低。Convpass[14]提出采用卷积型旁路代替经典的Adapter瓶颈结构，并结合实验证实这种微调模式更加适应CV任务。视觉提示搜索(Neural prOmpt seArcH, NOAH)[23]则是一种综合性的方法，融合了VPT, LoRA 和Adapter 3种微调方法，并取得了更优的效果。

然而，这些方法主要应用于自然场景图像分类或对象检测中相对简单的任务。对于遥感场景分类任务而言，其普遍存在背景信息复杂、多尺度信息占比重等问题。对于以VPT为代表的提示微调方法，其没有充分考虑视觉提示模块与遥感图像特性的关联，而对于以Adapter为代表的适配器微调方法，其适配器模块主要是简单的多层感知机(MultiLayer Perceptron, MLP)和卷积模块，难以捕捉遥感图像复杂的多尺度特征和语义特征。因此，在遥感领域，这类研究仍然存在显着的研究空白，值得做进一步的研究。

3 基于多尺度融合的适配器微调方法

多尺度融合适配器微调(Multi-Fusion Adapter, MuFA)的基础是使用视觉Transformer进行遥感场景分类。这种分类方法通常采用的是编码器-解码器结构。其中编码器使用ViT或Swin Transformer模型，解码器则使用简单的多层感知机(MLP)或全连接层，其结构如图1所示。这种编码器-解码器结构是非对称的，编码器一般是深层神经网络，有大规模的参数，其主要作用是将输入图片转化为高维度的特征向量，解码器则通常结构简单，仅为1层或2层全连接层，作用是将高维特征向量转化为图片的类别信息。因此，在模型的微调过程中，编码器部分消耗几乎所有的计算资源，因而参数高效微调聚焦于如何高效地微调编码器部分的参数。

3.1 并联适配器微调

如图2所示，在每个视觉Transformer层中，多尺度融合适配器微调(MuFA)将多尺度融合适配器模块(图2中蓝色模块)与原始的视觉Transformer模块(图2中绿色部分)并联连接。如果用LN代表层归一化，MHA代表多头自注意力模块，MLP代表多层感知机，MuFA代表需要微调的多尺度融合适配器模块，$\boldsymbol{x}$代表前向传播过程中的特征向量，i代表当前特征向量和模块所在的层，那么在单个视觉Transformer层中，前向传播过程可以表示为

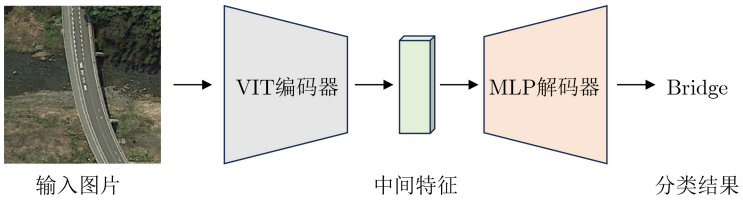


图 1 基于视觉Transformer进行图片分类

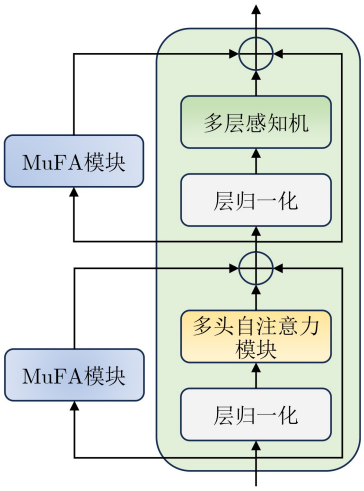


图 2 多尺度融合适配器微调的并联连接

$$\boldsymbol{x}_i' = \boldsymbol{x}_i + \mathrm{MHA}_i(\mathrm{LN}_i^1(\boldsymbol{x}_i)) + \mathrm{MuFA}_i^1(\boldsymbol{x}_i) \tag{1}$$

$$\boldsymbol{x}_{i+1}' = \boldsymbol{x}_i' + \mathrm{MLP}_i(\mathrm{LN}_i^2(\boldsymbol{x}_i')) + \mathrm{MuFA}_i^2(\boldsymbol{x}_i') \tag{2}$$

模型的这种并联设计有两个原因，其一是并行设计使用独立的分支而非将模块嵌入，这样可以维护模型原始的特征，并进一步增强模型的语义特征。其二是原始的串联设计相当于增加了更多的层数，可能会导致模型优化困难。在模型微调过程中，原始的视觉Transformer模块，包括层归一化模块、多头自注意力模块和多层感知机模块将不再进行参数更新，仅MuFA模块将进行微调训练。经过专门设计的MuFA模块是一个轻量级的模块，这样的设计使得模型在训练过程中只需要微调很少量的参数便能达到较优的性能。

3.2 多尺度融合适配器模块

当某个预训练模型应用于特定下游任务时，并非该预训练模型提供的所有特征向量都是有用的。相反，某些特征向量可能会使模型产生负迁移。基于这种理解，本文通过在模型上附加多尺度融合适配器模块以引入偏差项。多尺度融合适配器模块的结构如图3所示。输入特征图首先会经过若干个下采样层，得到的若干个中间特征图在经过激活后将进行对应的上采样计算，并经过加权融合后，模型得到最终的输出特征图。这种方法的特点是在某些通道中进行更高程度的下采样，以确保关键的特征信息得到保留。

对于多尺度融合适配器模块，如果使用 GELU 作为激活函数，Up和Down分别表示上采样和下采样计算，那么多尺度融合适配器模块的前向推理过程可以表示为

$$\mathrm{MuFA}(\boldsymbol{x}) = \sum_{i=1}^{n} \mathrm{Up}_i(\mathrm{GELU}(\mathrm{Down}_i(\boldsymbol{x}))) \tag{3}$$

在具体实现中，上采样和下采样通过单个全连接层实现，全连接层带有额外的bias。如果将通道为c的特征向量下采样到c'，那么全连接层的权重将是一个$c\times c'$的权重矩阵，全连接层的bias则是一个长度为c'的向量。多尺度融合的适配器模块的个数被设置为3个(即上式中的$n=3$)，对应的其下采样倍数依次为8倍、16倍和32倍。较高的下采样倍数降低了额外引入的可训练参数量，并加快了模型的训练速度。

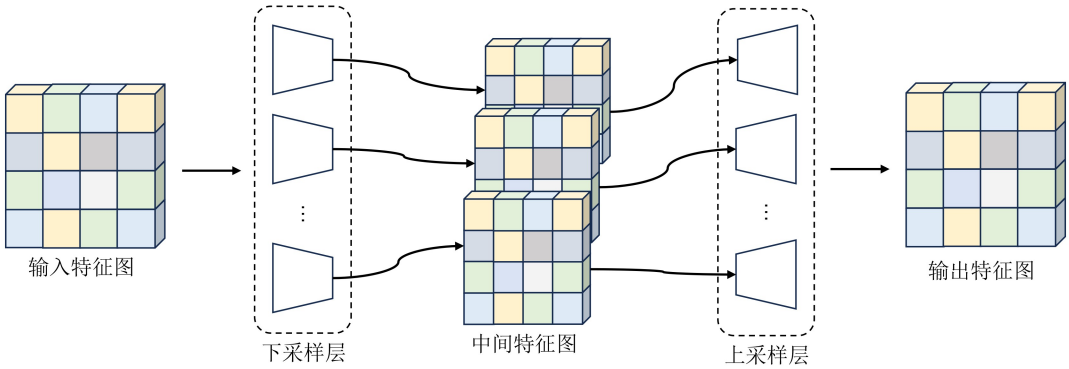


图 3 多尺度融合适配器模块

3.3 训练过程

由于参数冻结策略的存在，使用MuFA进行模型微调时，整个模型的训练方法是有所不同的。如果有数据集$D(\boldsymbol{x}_i, y_i)$，$D(\boldsymbol{x}_i, y_i)$其中$\boldsymbol{x}_i$为输入图像，y_i为真值标签，那么对于模型f_θ，其中θ为模型参数，完全微调的参数更新过程为

$$L(D,\theta)=\sum_{i}^{n}\text{Loss}(f_\theta(\boldsymbol{x}_i),y_i) \tag{4}$$

$$\theta \leftarrow \arg\min L(D,\theta) \tag{5}$$

而如果仅微调模型的部分参数，则模型参数可以拆解为两部分，即θ_{f}和θ_{t}。其中θ_{t}表示需要进行微调的模型参数，即本文所提方法的MuFA模块部分和分类头，θ_{f}为无需微调的参数，即本文所提方法的视觉Transformer编码器部分。因此在这种情形下，模型的参数更新过程为

$$L(D,\theta_{\text{f}},\theta_{\text{t}})=\sum_{i}^{n}\text{Loss}(f_\theta(\boldsymbol{x}_i),y_i) \tag{6}$$

$$\theta_{\text{t}} \leftarrow \arg\min L(D,\theta_{\text{f}},\theta_{\text{t}}) \tag{7}$$

综合上述公式，在模型训练过程中，通过引入MuFA模块，可以减少最后一步参数更新过程的计算量，从而达到提升训练效率、节省训练时间的效果。此外，当使用多个GPU进行并行训练时，优于需要进行梯度更新的参数更少，因此GPU通信的时间损失将会大大减小，从而实现更高效的多GPU并行训练。

4 实验结果与分析

4.1 实验设置

在本文的研究中，两个预训练模型被选择为遥感场景分类的基准模型： Vit-Base[6]和Swin-Tiny[7]。这两个模型都在ImageNet-1K数据集上进行了预训练，因此具有比较丰富的特征表示能力。Vit-Base是一种经典的视觉Transformer结构，而Swin-Tiny是一个轻量级的视觉Transformer模型。

本文复现了多种经典的参数高效微调方法，以评估所提方法在遥感场景分类任务上的性能。这些对比方法包括完全微调、完全冻结(仅微调分类头)、仅微调Layer Norm层[24]、Adapter[10], LoRA[11], BitFit[9], AdaptFormer[13]和Convpass[14]。这些对比方法大多发表在顶级机器学习期刊和会议上，与之对比能够体现所提方法在遥感场景分类任务上的优势。此外，为综合评估所提方法，两个常用的遥感场景分类数据集被用于实验。加州大学默塞德分校发布的土地利用图像遥感数据集(University of California at Merced land-use dataset, UCM)[25]是一个包含21个类别的数据集，涵盖了不同类型的遥感图像。它具有一定的多样性，适合用于模型性能评估。西北工业大学发布的45类遥感场景分类数据集(NorthWestern Polytechnical University REmote Sensing Image Scene Classification, NWPU-RESISC45)[26]则包含45个遥感场景类别，是一个更具挑战性的数据集，涵盖了更广泛的场景，包括城市、农田、森林等。对于这两个数据集，实验沿用了标准的训练/测试集划分方式，其中UCM按照8:2划分训练和测试集，NWPU-RESISC45则按照2:8划分训练和测试集。训练时，所有图片被调整为224×224像素大小。本文所有实验使用两张NVIDIA RTX 4 090显卡完成，批次大小(batch size)设为128，共微调100个轮次(epoch)。训练过程使用AdamW优化器，基础学习率为0.001(完全微调为0.000 1)，带有5个epoch的预热(warm up)，并采用余弦退火学习率衰减设置。

4.2 对比实验结果

以Vit-Base为基准模型的对比实验结果如表1所示。其中，微调参数表示除分类头外，模型中被训练的参数占总参数的比例。完全微调意味着所有参数都被训练，而完全冻结表示除分类头外的所有参数都未被训练。分类的评价指标为分类正确率(Accuracy, Acc.)在Vit-Base为基准模型的对比试验中，本文提出的 MuFA 在场景分类性能方面表现出色。在 UCM和RESISC45两个数据集上，其性能仅次于完全微调，优于所有参数高效微调方法。值得注意的是，得益于对整个原有模型的冻结，

表 1　以Vit-Base为基准模型的对比试验(%)

方法	微调参数	数据集	
		UCM	RESISC45
		Acc.	
完全微调	100.00	99.05	91.66
完全冻结	0	90.95	76.55
微调Layer Norm	0.07	94.58	86.40
Adapter	2.08	97.66	89.12
LoRA	4.04	98.12	89.44
BitFit	0.24	94.76	87.22
AdaptFormer	2.08	98.12	90.98
Convpass	4.12	97.94	90.42
LoRand	1.84	97.80	90.17
VPT	0.78	95.47	88.15
IVPT	0.78	97.41	90.55
E^{2}VPT	0.39	97.90	90.24
MuFA(本文方法)	3.64	**98.57**	**91.29**

MuFA仅微调了模型中的3.64%参数，这大大降低了模型迁移过程的时间成本和计算成本。

以Swin-Tiny为基准模型的对比试验结果如表2所示。Swin-Tiny作为一种轻量级的视觉Transformer模型，利用了滑动窗口注意力机制，更适用于遥感场景的任务。在Swin-Tiny为基准模型的对比实验中，MuFA同样表现出极佳的分类性能，在两个数据集上的分类正确率都优于所有参数高效微调方法，同时减少了所需的微调参数量。

表 2 以Swin-Tiny为基准模型的对比试验(%)

方法	微调参数	数据集	
		UCM	RESISC45
		Acc.	
完全微调	100.00	99.76	94.70
完全冻结	0	96.67	84.21
微调Layer Norm	0.07	97.66	91.70
Adapter	2.08	98.80	92.26
LoRA	4.04	98.58	92.55
BitFit	0.24	98.12	91.76
AdaptFormer	2.08	99.05	93.88
Convpass	4.12	99.05	93.61
LoRand	1.84	98.53	92.75
VPT	0.78	98.17	92.10
IVPT	0.78	98.73	92.69
E^2VPT	0.39	99.32	93.10
MuFA(本文方法)	3.64	**99.52**	**94.10**

图4列出了ViT-Base模型上UCM数据集的对比试验和Swin-Tiny模型上RESISC45数据集的对比实验的可视化结果。在与诸多方法的比较中，本文所提方法具有最好的微调效果。横坐标为微调参数量所占的百分比(Tuning Parameters, Param.)，纵坐标为在对应数据集上的分类正确率(Accuracy, Acc.)。

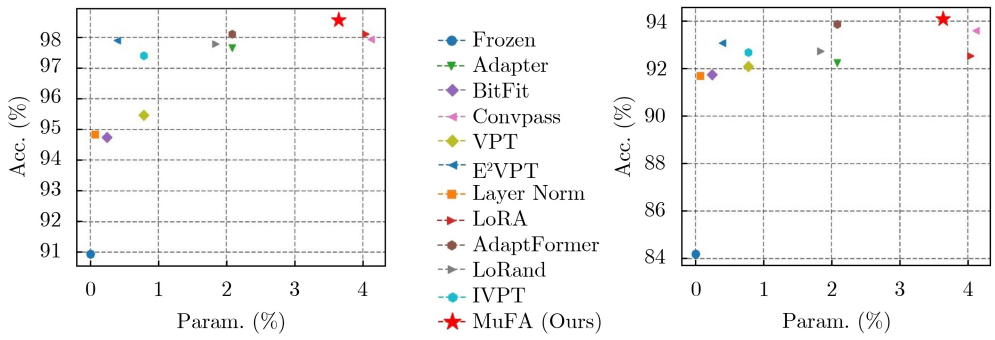
MuFA在不同视觉 Transformer模型上展现出稳定的性能。它不仅降低了微调参数，节省了训练成本，而且其性能与完全微调相差无几。这表明MuFA 是一种有效的模型微调方法，既保持了性能，又减少了资源开销。

4.3 消融研究

以Swin-Tiny为基准模型的消融实验结果如表3所示。在前面的对比试验中，已经观察到MuFA相对于计算机视觉领域的参数微调方法具有显著的优势。然而，为了更全面地证明MuFA的有效性，本文还进行了消融实验的研究。MuFA的设计包含两个关键方面，即融合模块的并联连接和多尺度结构。接下来，本文将详细探讨这两个设计对MuFA性能的影响。

表 3 以Swin-Tiny为基准模型的消融研究(%)

序号	消融研究					数据集	
	连接方式		融合结构			UCM	RESISC45
	串联	并联	单尺度	双尺度	多尺度	Acc.	
(1)	√		√			98.80	92.26
(2)	√			√		98.80	92.76
(3)	√				√	99.12	93.44
(4)		√	√			99.05	93.88
(5)		√		√		99.30	93.96
(6)		√			√	**99.52**	**94.10**



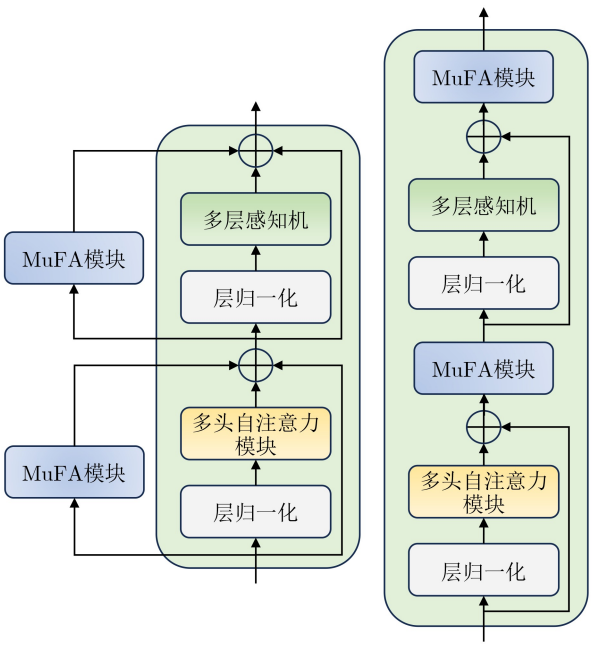
(a) ViT-Base模型上UCM数据集的对比试验结果　　(b) Swin-Tiny模型上RESISC45数据集的对比实验结果

图 4 Vit-Base模型和Swin-Tiny模型上的对比试验

如图5所示，参考现有的Adapter微调方法，MuFA的融合模块可以采取两种不同的连接方式：串联连接和并联连接，已有的方法中，如Convpass, AdaptFormer采用并联，LoRand采用串联。本文方法选择并联连接。具体而言：当MuFA采取串联连接，并仅使用单尺度时，MuFA将退化为经典的Adapter结构，这种结构下原模型的特征向量会直接经过Adapter的瓶颈模块或MuFA的融合模块，而仅采用并联并使用单尺度的MuFA则与AdaptFormer结构相同，特征向量会分别经过原模型和AdaptFormer瓶颈模块或MuFA融合模块，再进行组合。在消融研究中，模块与原模型并联的性能通常优于与原模型串联，这在对比实验中表现为AdaptFormer的性能优于Adapter，在消融研究中则表现为(4), (5)和(6)的分类性能分别优于(1), (2)和(3)。

MuFA采用了多尺度融合的结构，该模块使用多个不同采样倍数的适配器模块实现，这使得融合模块能够更充分地利用不同尺度的信息。本文进一步探讨了不同尺度数目对MuFA性能的影响。具体而言，增加MuFA融合模块数目有效提升了模型最终的分类性能，在消融研究中，(1), (2)和(3)以及(4), (5)和(6)依次表现出性能的提升。

综上所述，MuFA设计的有效性在不同连接方式和尺度融合数量的情况下都得到了验证，既保持了性能，又降低了资源开销。这为进一步推广MuFA在视觉Transformer遥感场景应用提供了有力的支持。



(a) MuFA融合模块并联连接　(b) MuFA融合模块串联连接

图 5 MuFA模块连接方式

5　结束语

本文提出了一种基于多尺度融合的适配器微调方法(MuFA)，旨在提升视觉Transformer模型在遥感场景分类任务中的迁移效率。MuFA通过融合不同下采样倍率的瓶颈模块，有效地结合了图像的纹理特征和语义特征，从而在仅微调少量参数的前提下，保持了模型的分类能力。实验结果表明，MuFA在减少微调参数、节省计算资源的同时，具有较好的分类性能，优于其他对比方法。

未来的研究方向可以进一步探索更复杂的遥感场景分类任务，并考虑更多不同类型的遥感任务和数据集。此外，对于更大规模的视觉Transformer模型，特别是专门的遥感无监督预训练模型，如空天·灵眸大模型(RingMo)[27]、多模态遥感大模型(SkySense)[28]等，如何进一步提高迁移效率也是一个值得研究的问题。

参 考 文 献

[1] 王佩瑾, 闫志远, 容雪娥, 等. 数据受限条件下的多模态处理技术综述[J]. 中国图象图形学报, 2022, 27(10): 2803–2834. doi: 10.11834/jig.220049.
WANG Peijin, YAN Zhiyuan, RONG Xue'e, *et al*. Review of multimodal data processing techniques with limited data[J]. *Journal of Image and Graphics*, 2022, 27(10): 2803–2834. doi: 10.11834/jig.220049.

[2] BI Hanbo, FENG Yingchao, YAN Zhiyuan, *et al*. Not just learning from others but relying on yourself: A new perspective on few-shot segmentation in remote sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5623621. doi: 10.1109/TGRS.2023.3326292.

[3] WANG Peijin, SUN Xian, DIAO Wenhui, *et al*. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(5): 3377–3390. doi: 10.1109/TGRS.2019.2954328.

[4] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012: 1097–1105.

[5] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[C]. 9th International Conference on Learning Representations, Austria, 2021.

[7] LIU Ze, LIN Yutong, CAO Yue, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows[C]. The IEEE/CVF International Conference on Computer

Vision, Montreal, Canada, 2021: 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[8] DENG Jia, DONG Wei, SOCHER R, *et al*. ImageNet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248–255. doi: 10.1109/CVPR.2009.5206848.

[9] BEN ZAKEN E, GOLDBERG Y, and RAVFOGEL S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models[C]. The 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Dublin, Ireland, 2022: 1–9. doi: 10.18653/v1/2022.acl-short.1.

[10] HOULSBY N, GIURGIU A, JASTRZEBSKI S, *et al*. Parameter-efficient transfer learning for NLP[C]. 36th International Conference on Machine Learning, Long Beach, USA, 2019: 2790–2799.

[11] HU E J, SHEN Yelong, WALLIS P, *et al*. LoRA: Low-rank adaptation of large language models[C]. 10th International Conference on Learning Representations, 2022.

[12] YIN Dongshuo, YANG Yiran, WANG Zhechao, *et al*. 1% vs 100%: Parameter-efficient low rank adapter for dense predictions[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 20116–20126. doi: 10.1109/CVPR52729.2023.01926.

[13] CHEN Shoufa, GE Chongjian, TONG Zhan, *et al*. AdaptFormer: Adapting vision transformers for scalable visual recognition[C]. The 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1212.

[14] JIE Shibo and DENG Zhihong. Convolutional bypasses are better vision transformer adapters[EB/OL]. https://arxiv.org/abs/2207.07039, 2022.

[15] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. 3rd International Conference on Learning Representations, San Diego, USA, 2015.

[16] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[17] LESTER B, AL-RFOU R, and CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]. The 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021: 3045–3059. doi: 10.18653/v1/2021.emnlp-main.243.

[18] JIA Menglin, TANG Luming, CHEN B C, *et al*. Visual prompt tuning[C]. 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 709–727. doi: 10.1007/978-3-031-19827-4_41.

[19] YOO S, KIM E, JUNG D, *et al*. Improving visual prompt tuning for self-supervised vision transformers[C]. 40th International Conference on Machine Learning, Honolulu, USA, 2023: 40075–40092.

[20] TU Chenghao, MAI Zheda, and CHAO Weilun. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 7725–7735. doi: 10.1109/CVPR52729.2023.00746.

[21] HAN Cheng, WANG Qifan, CUI Yiming, *et al*. E^{2}VPT: An effective and efficient approach for visual prompt tuning[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 17445–17456. doi: 10.1109/ICCV51070.2023.01604.

[22] HE Xuehai, LI Chunyuan, ZHANG Pengchuan, *et al*. Parameter-efficient model adaptation for vision transformers[C]. Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, USA, 2023: 817–825. doi: 10.1609/aaai.v37i1.25160.

[23] ZHANG Yuanhan, ZHOU Kaiyang, and LIU Ziwei. Neural prompt search[EB/OL]. https://arxiv.org/abs/2206.04673, 2022.

[24] QI Wang, RUAN Yuping, ZUO Yuan, *et al*. Parameter-efficient tuning on layer normalization for pre-trained language models[EB/OL]. https://arxiv.org/abs/2211.08682, 2022.

[25] YANG Yi and NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. The 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, USA, 2010: 270–279. doi: 10.1145/1869790.1869829.

[26] CHENG Gong, HAN Junwei, and LU Xiaoqiang. Remote sensing image scene classification: Benchmark and state of the art[J]. *Proceedings of the IEEE*, 2017, 105(10): 1865–1883. doi: 10.1109/JPROC.2017.2675998.

[27] SUN Xian, WANG Peijin, LU Wanxuan, *et al*. RingMo: A remote sensing foundation model with masked image modeling[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, 61: 5612822. doi: 10.1109/TGRS.2022.3194732.

[28] GUO Xin, LAO Jiangwei, DANG Bo, *et al*. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery[EB/OL]. https://arxiv.org/abs/2312.10115, 2023.

尹文昕：女，助理研究员，研究方向为遥感图像智能解译.
于海琛：男，硕士生，研究方向为遥感智能模型轻量化.
刁文辉：男，副研究员，研究方向为遥感图像智能解译.
孙 显：男，研究员，研究方向为计算机视觉与遥感图像理解.
付 琨：男，研究员，研究方向为遥感大数据智能解译.

责任编辑：余 蓉