文章编号:1001-9014(XXXX)XX-0001-14

DOI:10. 11972/j. issn. 1001-9014. XXXX. XX. 001

基于自监督预训练与跨尺度对比学习的多模态遥感图像融合

李朝伟[1,2]， 冯世阳[1,2]， 王 斌[1,2]

（1. 复旦大学 电磁波信息科学教育部重点实验室,上海,200433;
2. 复旦大学 信息学院图像与智能实验室,上海,200433）

摘要:自监督预训练方法具有强大的特征提取和模型迁移能力,然而,目前多模态遥感图像融合中的预训练方法只对所提取多模态特征进行拼接等操作实现简单融合,而未针对多模态信息的融合设计专有模块,导致多模态互补信息融合不充分;其次,这些方法未考虑和利用遥感图像内部的跨尺度一致性先验,导致其对多模态遥感信息的提取和整合有限,因而使得各种下游任务的性能有待提高。针对上述问题,提出一种基于自监督预训练与跨尺度对比学习的多模态遥感图像融合方法,主要包括三部分:1)通过引入交叉注意力融合机制初步融合不同模态提取的特征,再借助于编码器模块进一步提取特征,从而实现各模态互补信息的显式聚合和提取;2)通过引入跨模态融合机制,使每种模态能从所有模态的特征中提取有用的补充信息,分别解码后重构各模态输入;3)基于遥感图像的跨尺度一致性约束,引入跨尺度对比学习,以增强对单模态信息的提取,实现更鲁棒的预训练。在多个公开多模态遥感图像融合数据集上的实验结果表明,与现有方法相比,所提出算法在多种下游任务中均取得了显著的性能提升,在 Globe230k 数据集上达到了 79. 01 的平均交并比(mIoU)、92. 56 的总体准确率(OA)和 88. 05 的平均 F1 分数(mF1),且具有扩展性好、超参数易设置的优点。

关 键 词:多模态遥感图像融合; 自监督预训练; 对比学习; 跨尺度一致性

中图分类号:TP751 **文献标识码**: A

Multimodal Remote Sensing Image Fusion Based on Self-supervised Pre-training and Cross-scale Contrastive Learning

LI Zhao-Wei[1,2], FENG Shi-Yang[1,2], WANG Bin[1,2]

（1. Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China;
2. Image and Intelligence Laboratory, School of Information Science and Technology, Fudan University, Shanghai 200433, China）

Abstract: Self-supervised pre-training methods have strong capabilities in feature extraction and model transfer. However, current pre-training methods in multimodal remote sensing image (RSI) fusion only perform simple fusion operations such as concatenation on the extracted multimodal features without designing dedicated modules for the integration of multimodal information, leading to insufficient fusion of complementary information across modalities. Secondly, these methods do not consider and utilize the cross-scale consistency priors within RSIs, resulting in limited extraction and integration of multimodal remote sensing information, and thus the performance of various downstream tasks needs to be improved. In response to the above issues, a multimodal RSI fusion method based on self-supervised pre-training and cross-scale contrastive learning is proposed, which mainly includes three parts: 1) By introducing a cross-attention fusion mechanism to preliminarily integrate features extracted from different modalities, and then using encoder modules to further extract features, explicit aggregation and extraction of complementary information from each modality are achieved; 2) By introducing a cross-modality fusion mechanism, each modality can extract useful supplementary information from the features of all modalities, and reconstruct each modality's input after separate decoding; 3) Based on the cross-scale consistency constraints of RSIs, cross-scale contrastive learning is introduced to enhance the extraction of single-modality information, achieving more robust pre-training. Experimental results on multiple public multimodal RSI fusion datasets demonstrate that, compared with existing methods, the proposed algorithm has achieved sig-

nificant performance improvements in various downstream tasks. On the Globe230k dataset, our method achieves an average intersection over union (mIoU) of 79.01%, an overall accuracy (OA) of 92.56%, and an average F1 score (mF1) of 88.05%, and it has the advantages of good scalability and easy hyperparameter setting.

Key words: Multimodal remote sensing image fusion, self-supervised pre-training, contrastive learning, cross-scale consistency

PACS:84. 40. Xb

引言

随着遥感成像技术的进步,各种遥感图像的获取变得更加容易,为遥感数据的各类应用提供了可靠的数据支持,并在如自然灾害监测[1]、城市规划[2]、环境监测[3]等任务中得到广泛应用。通常,这些任务依赖于单一模态数据,如高光谱图像(Hyperspectral Images, HSIs)、多光谱图像(Multispectral Images, MSIs)、合成孔径雷达(Synthetic Aperture Radar, SAR)等。然而,在某些复杂场景中,受限于遥感图像不同模态的特有性质,单一模态的图像可能无法提供全面的地表信息[4],限制了其检测和识别目标地物的能力。随着遥感图像成像技术的进步,同时获取同一场景的多模态数据已经成为可能,它们可为目标地物的检测与识别提供不同类型的互补信息。通过有效融合这些模态的互补信息,可得到更全面的地表信息,从而更好地为遥感下游任务提供支持。然而,由于不同模态遥感图像间存在显著差异,如何合理有效地实现多模态遥感图像的融合显得至关重要。

为了解决这一问题,大量的工作探索了多模态遥感图像信息融合结构的设计。早期的工作[5][6][7]主要探索了在模型不同阶段融合多模态信息的方法。这些方法利用卷积神经网络(Convolutional Neural Networks, CNNs)来提取不同模态的特征,并通过设计特定的网络结构来实现特征的融合。根据融合阶段的不同,这些方法可以分为早期、中期、晚期融合以及编码器-解码器融合、交叉融合等。这里,编码器-解码器融合方法[5]将编码器得到的多模态融合特征经解码器重构原始输入,来促进融合特征学习;交叉融合[6]方法则通过通过显式的跨模态信息交互来促进多模态融合。然而,这些基于CNNs的方法对于多模态特征的融合过于简单,不同模态数据的异质性导致多模态信息的融合困难。其次,受限于CNNs中卷积运算的局限性,这类方法往往会忽略遥感图像中存在的全局信息及长程依赖关系,特别是,在遥感图像场景分类等任务中,如果模型只关注学习局部信息,会丢失一些全局的关键特性,从而导致在完成这些任务时的性能降低。因此,如何更好地保留全局长程信息是后续多模态遥感图像融合的研究方向之一。

受到自然图像中视觉Transformer(Vision Transformer, ViT)[8]成功的启发,一系列基于ViT的方法[9][10]被运用到多模态遥感图像融合中来解决上述问题。ViT模型基于注意力机制来捕捉图像不同块间的关系,能有效捕捉图像中不同模态间的长程依赖关系。得益于ViT的特性,它们在处理遥感图像中的复杂地物结构方面具有独特优势,能有效提取和整合不同模态输入的互补信息,提高遥感图像分析和理解的准确性,进而提升下游任务的性能,如目标检测、场景分类和语义分割等。这里,MFT[9]将来自不同模态的特征在编码器中用作分类标记,有助于实现更好的泛化;ExViT[10]则借助可分离卷积模块扩展的位置共享ViT的平行分支,来处理多模态遥感图像块,通过跨模态注意力融合多模态输入,能更好地利用不同模态数据间的互补信息。然而,由于这些模型的表征能力有限,它们无法有效地利用大量多模态数据来进一步提升性能。其次,这些模型在迁移学习和泛化能力方面存在一定局限性。当在一个数据上训练后,面对新领域或数据集时,这些模型可能无法进行迁移或适应,导致性能下降。因此,如何通过预训练来改善这些模型的泛化能力成为一个研究热点。

在后续的工作中,为解决多模态遥感图像融合方法迁移性差的问题,以掩码图像建模[11]为代表的基于自监督学习的预训练方法[11][12][13][14][15]开始被广泛应用在多模态遥感图像融合中。通过在大量数据上利用数据本身的结构信息来生成伪标签、设计预测任务和进行自监督预训练,模型可捕捉到遥感图像的多模态互补特征,在迁移到新数据集时,通过简单的微调就可以得到出色的效果。其中典型的方法是掩码图像建模,通过随机掩码掉原始输入的一部分图像块后送入编码器,再经过解码器重建原始图像,使得模型能充分提取输入图像中的信息。代表性的方法包括SatViT[14],将SAR和光学图

形沿通道维度简单叠加,用于预训练;而3DMAE[15]则探索了遥感图像中的三维掩码策略,以进一步提高预训练的有效性。然而,这类方法仍有一定的局限性:1)大部分预训练方法只是简单地对多模态特征进行拼接等操作实现融合,没有为多模态信息融合设计专有模块,这导致其难以充分融合互补的多模态信息;2)这些方法未考虑遥感图像内部的跨尺度一致性特点,这导致编码器对遥感图像的信息提取不足,难以满足复杂场景中下游任务的需求。

针对上述问题,本文提出一种基于自监督预训练与跨尺度对比学习的多模态遥感图像融合方法,其主要由三部分构成:1)在多模态融合编码器模块中,我们引入了交叉注意力融合(Cross-Attention Fusion, CAF)机制来初步融合不同模态的特征,融合后的特征通过Transformer编码器模块进一步聚合和提取特征;2)在模态特定解码器模块中,通过引入跨模态融合(Cross-Modality Fusion, CMF)机制,使得解码过程中的各模态特征能从所有模态特征中提取有用的补充信息,最终重建各模态的原始图像;3)基于遥感图像跨尺度一致性的先验(即在不同尺度下捕获到的同一场景的图像是相似的),采用对比学习策略,确保同一场景内不同尺度的图像具有相似潜在表示,以增强对单模态信息的提取,实现更鲁棒的预训练。此外,本文方法还可方便地扩展到更多输入模态数目的情况。在大量多模态遥感图像融合数据集上进行的预训练和迁移实验,也验证了该方法在各类下游任务中的有效性和可扩展性。

本文的主要贡献可简要总结如下:

1) 设计了多模态融合编码器模块,通过在编码器模块中引入了CAF机制,实现了多模态互补信息的显式聚合和提取,促进了多模态信息的有效融合;

2) 设计了模态特定解码器模块,通过引入了CMF机制,允许每种模态从所有模态特征中获得补充信息,促进了模态解码和原始图像的有效重建;

3) 基于多模态遥感图像内部的跨尺度一致性先验,采用对比学习方法,充分利用图像内部的跨尺度特征,促进了单模态信息的充分提取,提升了预训练的鲁棒性。

1　相关工作

1.1　自监督预训练

自监督预训练(Self-Supervised Pre-training, SSP)是一种机器学习范式,特别是在深度学习领域中,用于训练模型以学习数据的表示或特征。与传统的监督训练不同,自监督预训练不需要标注数据,它通过利用数据本身的结构信息来生成伪标签,设计预测任务,使模型能够在未标注的数据上进行预训练,学习到通用的特征表示。然后,预训练到的模型可被微调,以适应特定的下游任务。总的来说,通过自监督预训练学习到的特征能提高模型在特定任务上的性能,尤其是在数据标注成本高昂或难以获得的情况下。

掩码自编码器[11](Masked Autoencoders, MAE)是一种典型的自监督预训练方法,它通过重构输入数据中被掩码的部分来学习数据的表示。在掩码自编码器中,输入数据首先被随机掩码一部分,然后,模型基于未被掩码的部分来预测被掩码的部分。掩码自编码器的训练目标可表示为最小化重构误差,即模型重构输出和原始数据间的差异。具体来说,对于一个给定输入x,其中随机一部分被掩码后得到x_M。模型的目标是最小化重建损失,该损失是输入x和解码器输出$\mathrm{D}(\mathrm{E}(x_M))$之间的差异,其中,E和D分别表示编码器和解码器,其损失可表示为:

$$\mathrm{L}_{\text{reconstruction}} = \mathrm{E}_x\left[\left\| x - \mathrm{D}(\mathrm{E}(x_M)) \right\|_2^2\right] \tag{1}$$

其中$\|\cdot\|_2^2$表示L2范数的平方,E_x表示对输入的期望,该损失鼓励模型学习到能准确重构原始数据的表示。

在实践中,掩码自编码器通常采用变分自编码器(Variational Autoencoder, VAE)或生成对抗网络(Generative Adversarial Networks, GAN)的结构。在VAE的情况下,模型会学习到一个潜在空间,其包含了输入数据的压缩表示,模型通过从这个潜在空间采样来重构数据。对于GAN,生成器网络负责重构数据,而判别器网络则确保重构数据与原始数据无法区分。掩码自编码器的一个关键优势是它能捕捉到数据的复杂结构,如图像中的轮廓等细节信息。此外,由于不需要外部标签,因此可在大量未标记数据上进行训练,这在多模态遥感图像融合这种数据标注成本高昂的情况下尤其有效。

1.2　对比学习

对比学习[16](Contrastive Learning)通过学习数据的正负样本对来揭示数据的内在结构和特征。该方法的核心思想是将相似的样本拉近,将不相似的样本推远,从而学习到一个能够区分不同样本的

特征空间。对比学习的关键优势是其灵活性和可扩展性，通过学习数据的内在结构来提高模型的泛化能力，只需适当定义正负样本对，即可应用于不同的数据类型和任务，在图像、文本和语音等多种类型的数据上都取得了显著效果。此外，该方法也可在大量未标记数据上进行训练。

在对比学习中，每个样本通常被映射成一个高维特征向量。对于一个给定的输入x，通过一个编码器f将其映射到特征空间，得到特征向量$z=f(x)$。它的目的是使特征向量能捕捉到样本x的语义信息，其关键挑战在于如何定义正负样本对。一种常见的方法是使用数据增强技术来生成正负样本对，如通过旋转、裁剪或颜色变换等操作来改变原始图像以得到正样本；负样本则是随机选择的，正负样本之间没有相似性。对比学习中典型的损失函数是基于互信息概念的InfoNCE损失。对于一个样本x_i，其正样本对x_j和负样本对x_k，InfoNCE损失定义为：

$$L_{\text{InfoNCE}} = -\log\frac{\exp(\text{sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{N}\exp(\text{sim}(z_i,z_k)/\tau)} \tag{2}$$

其中，$\text{sim}(z_i,z_j)$是样本的特征向量z_i和z_j间的相似度，通常可使用余弦相似度来计算。τ是温度参数，用于控制相似度的缩放，N是负样本的数量。通过最小化InfoNCE损失，模型能够学习到能区分正负样本对的特征表示。

2 模型构建

本节将详细介绍所提出的预训练方法，其整体结构如图1所示，主要由三部分构成：多模态融合编码器、模态特定解码器和跨尺度对比学习分支。其中，多模态融合编码器用于提取和聚合多模态互补特征；模态特定解码器利用所得到的各模态特征进行解码，重建原始图像；跨尺度对比学习模块则基于跨尺度一致性，进行对比学习，促进特征提取；最后，采用对比损失和重建损失函数进行模型优化。

2.1 多模态融合编码器

多模态融合编码器基于ViT[8]架构，用于提取并聚合多模态输入的互补信息。对于不同模态的输入，将其划分为16×16的图像块，随机掩码一部分后，使用不同的线性投影层将图像块投影到具有Transformer编码器维度的标记中；然后，这些不同模态的标记被拼接成标记序列，作为Transformer编码器的输入。考虑到自注意力机制关于序列长度的二次方计算复杂度，我们仅将每种模态掩码后的剩余标记进行拼接，然后输入到编码器中，以显著减

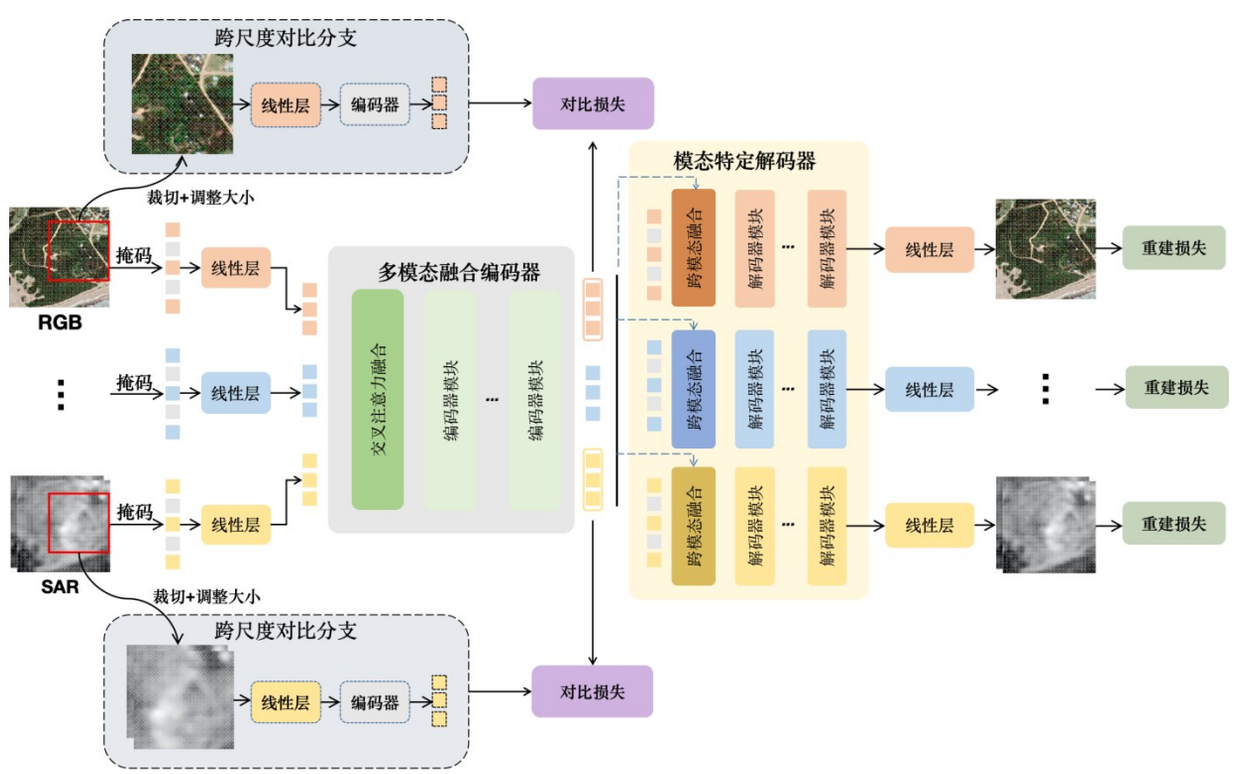

图1 本文方法的整体框架

Fig. 1 The overall framework of our method

少计算资源的消耗,同时极大提高训练效率。类似于ViT中的方法,引入了一个额外的嵌入可学习的全局标记,类似于ViT中的分类标记。另外,考虑到所有模态的二维结构,在线性投影之后加入了二维正弦余弦位置嵌入。

尽管这种拼接不同模态的标记的方法允许在编码过程中进行交互,但由于不同模态输入间存在的显著差异,这种简单的融合方法可能无法有效实现多模态信息聚合。因此,在此引入了CAF机制,以显式地促进不同模态间的信息融合。具体来说,考虑到来自不同模态的一组原始标记 M_i,其中 $i \in \{1,2,\ldots,n\}$,n 代表模态的数量。首先通过交叉注意力进行模态间交互:

$$\text{Cross}-\text{Attention}(x,y)=\text{Softmax}\left(\frac{Q_x K_y^T}{\sqrt{d_q}}\right)V_y \tag{3}$$

$$M_i'=\text{Cross}-\text{Attention}(M_i,M_{j\neq i}) \tag{4}$$

其中 Q_x、K_y 和 V_y 分别表示查询向量、键向量和值向量,d_q 表示查询向量的维度,$M_{j\neq i}$ 表示其它模态的特征。即每种模态的标记充当查询,而其它模态的标记充当键和值,这允许每种模态的标记聚合来自其它模态的补充信息。CAF可以表示为:

$$\text{CAF}(M_i)=M_i'+\text{MLP}(M_i') \tag{5}$$

其中MLP是一个多层感知机。然后,使用 $N-1$ 个Transformer编码器块来进一步提取多模态特征并聚合信息。最后,获得了第 i 个模态的潜在表示 Z_i。

考虑到来自多个模态特征间存在的显著差异,直接使用普通的编码器结构会导致每种模态的特征主要关注其自身模态内的信息,从而导致跨模态交互不足。本文的方法则可促进早期的跨模态交互融合,确保了每种模态的标记表示整合了来自其它模态的补充信息。

2.2　模型特定解码器

在上述多模态融合编码器之后,获得了未掩码标记的潜在表示。为了基于这些可见标记的潜在表示重建掩码标记,对每种模态使用单独的解码器,具体结构如图2所示。鉴于每种模态都需要自己的解码器,解码器的计算成本与模态数量成线性关系,为了保持预训练的效率,这里采用了对整体计算成本影响很小的轻量级解码器,即一个跨模态融合模块后跟两个Transformer解码器块,其维度较低,为256维。

每个解码器前面有一个线性投影层,以将编码器输出转换到解码器维度。每个模态解码器的输入包括相应模态的完整可见标记集合与一组掩码标记,这些输入一起被解码,其中,掩码标记作为解码器重建的占位符。为防止由于每个模态解码器仅依赖于其自身模态的标记而导致次优的解码结果,在每个解码器中引入了CMF机制,以整合来自其它模态编码标记的信息。

在CMF中,首先,将正弦余弦位置嵌入和可学习的模态嵌入加入到所有标记中,以帮助模型理解不同标记的位置和模态;然后,引入一个交叉注意力层,使用每种模态的标记作为查询,所有模态的标记作为键和值,促进多模态互补信息的进一步交互,后面是一个小型多层感知器。CMF可表示如下:

$$\text{CMF}(Z_i)=\text{MLP}(\text{Cross}-\text{Attention}(Z_i,Z_i+\text{pos}-\text{emb}+\text{mod}-\text{emb})) \tag{6}$$

其中,Z_i 表示第 i 个模态的嵌入表示,Cross-Attention即公式(3)中的定义;pos-emb表示位置编码,mod-emb表示模态编码,分别用于区分序列中不同位置和模态的标记。在CMF之后,两个Transformer解码器块用于进一步解码,其输出经过变形,重建出原始图像,用于计算重建损失。

2.3　跨尺度对比学习

原始的MAE[11]通过在图像中掩盖大部分像素来简单地训练一个图像重建网络。然而,与自然图

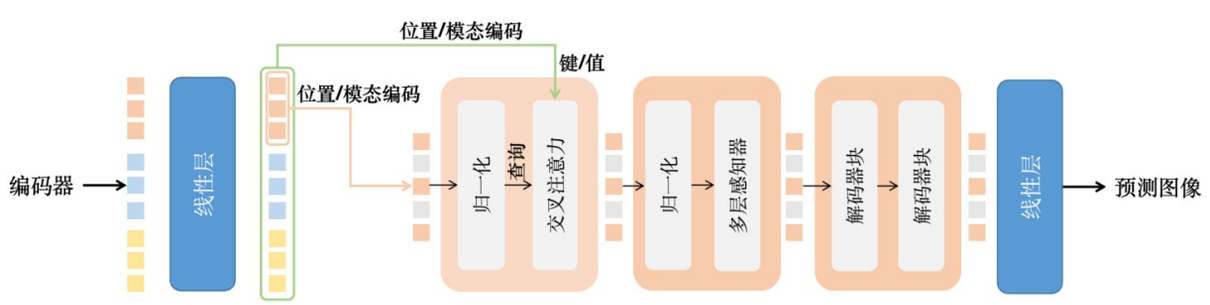

图2　解码器的整体结构

Fig. 2　The overall architecture of the decoder

像不同，遥感图像表现出跨尺度一致性的特性[17]，即在不同尺度下捕获的同一场景的图像是相似的，具有高度相关性，因此，即使对人类来说可能存在显著的视觉差异，它们在通过编码器后也应该有一致的表示；相反，不同场景捕获的图像则具有显著不同的表示。基于这一特性，我们引入跨尺度对比学习到多模态遥感图像融合的预训练中，以更好地学习图像特征。

跨尺度对比学习的目标是最大化同一场景不同尺度表示间的共享信息，同时最小化来自不同位置图像的共享信息。具体来说，对于一个训练样本，除了将原始输入图像输入编码器，还会随机裁切原始图像以生成子图像，再调整其大小到原始图像的大小，并将它与原始图像一起输入编码器，然后，基于原始图像和新图像的嵌入进行对比学习。与传统的数据增强方法不同，该方法中调整后的图像不是作为新的独立样本进行训练，而是在训练过程中动态生成的，并与原始图像一起输入编码器以更好地提取特征。

对于一个图像样本，通过编码器后，获得了对应于两种不同尺寸图像的嵌入。在计算对比损失时，用z_i^j表示第i个模态的第j个样本通过编码器后的表示，用$\hat{z}_i^j$表示生成的新图像的编码器表示，将跨尺度一致性损失定义如下：

$$L_{contra}=\frac{1}{2MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(L_{InfoNCE}\left(z_i^j,\hat{z}_i^k\right)+L_{InfoNCE}\left(z_i^k,\hat{z}_i^j\right)\right) \tag{7}$$

其中，M是模态的数量，N是批次大小，$L_{InfoNCE}$即公式（2）中定义的对比损失。该损失函数鼓励模型学习在不同尺度之间一致的特征表示，从而提高模型对遥感图像的理解和分类能力。

2.4 掩码策略与优化目标

为确保掩码自编码器的有效运行，需要对大量的标记进行掩码处理，不同的掩码策略对最终的预训练和迁移实验结果有着至关重要的影响。特别是，在多模态场景中，不同模态的掩码策略在模型预测不同模态和空间位置的掩码标记中起着至关重要的作用。为在所有实验中提高效率和简化训练，选择使用固定数量的可见标记。具体来说，对每种模态的补丁进行75%的掩码处理，保持不同模态间的掩码位置一致性，以确保每种模态在相同位置提供可见标记，降低注意力层捕获跨模态信息的难度。例如，当输入为三个模态224×224像素大小的遥感图像时，每个模态划分为16×16的图像块，则每个模态包含196个标记。训练时每种模态只保留49个标记，其余的进行掩码处理，最终三种模态共147个未掩码标记。

训练中的损失函数由两部分组成：跨尺度一致性对比损失和图像重建损失。重建损失是所有模态的重建损失之和，即所有样本所有模态重建损失的均值，定义如下：

$$L_{recon}=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left\| X_i^j-D(E(X_i^j))\right\|_2^2 \tag{8}$$

其中，X_i^j表示第i个模态的第j个样本，$D(E(X_i^j))$表示重建结果，最终损失函数是两种损失的加权和。

$$L=\alpha L_{contra}+L_{recon} \tag{9}$$

其中，L_{contra}和L_{recon}分别是公式（7）和公式（8）中定义的对比损失和重建损失，α是对比损失的权重，用于权衡不同损失的重要性。在后续实验中，对该参数进行了消融实验分析。

另外，表1中给出了本文方法的训练流程。首先，给定多种模态的遥感图像X_i，确定训练过程的超参数：训练的迭代次数N、对比损失函数的权重α。然后，执行如下步骤：1）各模态的遥感图像经过映射P_i，得到来自不同模态的一组原始标记M_i；2）使用多模态融合编码器进行编码，同时进行多模态信息的聚合，得到每个模态的潜在表示Z_i；3）原始图像裁切调整大小得到新图片$\hat{X}_i$，按照前两步获得其潜在表示$\hat{Z}_i$；4）模态特定解码器基于每种模态的潜在表示进行解码，进一步重构出原始图像R_i；5）基于重构图像和原始图像计算重建损失L_{recon}，基于原始图像和新图像的潜在表示计算对比损失L_{contra}，然后加权求出整体损失；6）反向传播更新网络参数。最终，我们可得到训练的网络。

3 实验结果与分析

3.1 实验数据

本文采用了BigEarthNet-MM、SEN12MS①、C2Seg-BW②以及Globe230k③四个公开的多模态遥感图像融合数据集来测试和评估所提出方法的性能，它们包含了多种模态和任务，涵盖丰富的场景。

①https://mediatum.ub.tum.de/1474000

②https://www.ieee-whispers.com/cross-city-challenge/

③https://zenodo.org/records/8429200

表1　本文算法训练过程的伪代码

Table 1　The pseudocode for the algorithm in training process

本文算法训练过程的伪代码
输入:多模态遥感图像 X_i
训练的总迭代次数N,对比损失权重α
循环迭代N次:
步骤1:$M_i = P_i(X_i)$
步骤2:$Z_i = E(M_i)$
步骤3:$\hat{Z}_i = E(P_i(\hat{X}_i))$
步骤4:$R_i = D(Z_i)$
步骤5:$L = \alpha L_{contra} + L_{recon}$
步骤6:反向传播更新参数
输出:训练好的网络

BigEarthNet-MM 数据集[18]包含了 590,326 对 Sentinel-1 和 Sentinel-2 图像,涵盖 19 个类别,这些图像覆盖了 10 个欧洲国家的地区。数据集中的图像具有不同的分辨率,包括 10 米、20 米和 60 米的波段像素。训练集和验证集与先前的研究一致,包括 354,196 个训练样本和 118,065 个验证样本。

SEN12MS 数据集[19]包含 180,662 个三元组,每个三元组由 SAR Sentinel-1 图像、多光谱 Sentinel-2 图像和 MODIS 土地覆盖图组成,这些图像覆盖了全球所有有人居住的大陆。数据集的地面采样距离为 10 米。

C2Seg-BW 数据集[20]包括 7990 个高光谱、多光谱和 SAR 三元组数据,是中国的北京和武汉城市的多模态遥感数据,涵盖 13 个前景类别。图像地面分辨率为 10 米。训练集中有 7140 个样本,测试集中有 850 个样本。

Globe230k 数据集[21]包含 232,819 个 RGB、NDVI、SAR(VV、VH)和 DEM 四元组数据,分别有 3、1、2 和 1 个波段。这些数据分布在 10 个前景类别中,这些标注图像来自世界各地的不同地点,覆盖面积超过 60,000 平方公里。数据集遵循 7:1:2 的训练集、验证集、测试集的比例。

在 BigEarthNet-MM 数据集上进行双模态的预训练实验;在迁移实验中,使用 SEN12MS 数据集进行实验,这两个双模态数据集的下游任务都是遥感图像场景多分类任务。更进一步,还将预训练模型迁移到 C2Seg-BW 数据集进行语义分割下游任务的验证。对于更多模态的情况,使用 Globe230k 多模态分割数据集进行实验,在更多模态数量上验证本文方法的有效性和易扩展性。

3.2　评价指标和实验设置

评价指标:在实验中,为了公平比较,报告了用于比较的所有模型在广泛使用的评价指标上的结果。不同下游任务使用不同的评价指标,具体来说,对于多标签分类任务,使用的评估指标是平均准确率(mAP)、召回率(Recall)以及 F1-分数(F1-score)。对于分割任务,使用的评估指标是总体准确率(OA)、平均 F1 分数(mF1)和平均交并比(mIoU)。这些统计指标为所提出方法和其它方法提供了公平的性能比较。为了更直观地展示不同方法的性能比较,本文还提供了可视化结果。

训练细节:使用 Pytorch 框架实现了所提出的模型。所有实验都在 8 个 NVIDIA A6000 GPU 上进行训练,迭代 100,000 步,批量大小设置为 16。对比损失的损失权重 α 设置为 0.5。在训练阶段,使用 Adam 优化器来优化网络模型的参数。为了增强模型性能,通过随机翻转和顺时针旋转来进行数据增强。所有数据集输入都划分为 16×16 图像块。

对比基线模型:为了验证本文方法的有效性和通用性,在多个数据集上进行了实验,由于此前缺少在这些数据集上均进行实验的工作,我们分别选取了每个数据集上的一系列最先进方法进行对比。BigEarthNet-MM 和 SEN12MS 数据集上,对比的基线模型是一些基于 Transformer 架构的专门针对 SAR-RGB 图像融合任务设计的模型,包括 SatVIT[14]、Fus-MAE[22]以及 DINO-MM[23]、。对于 C2Seg-BW 数据集,对比的基线模型包括 FastFCN[24]和 Segformer[25],它们是语义分割任务中的经典模型;此外,还对比了 HighDAN[20]、AdaptSeg[26]、DSAN[27]和 DualHR[28]等方法,这些方法着重解决城市环境中的域偏移问题[29]。

对于 Globe230k 数据集,由于该数据集公开时间较短,此前的研究很有限,因此,我们实现了一个简单但有效的基线模型。具体来说,该基线模型的输入图像通过两个卷积层处理进行特征提取,以生成相同通道数量的嵌入。在多模态图像输入的情况下,特征沿通道维度连接,然后通过两层 MLP 进行映射。模型使用在 ImageNet1k 上预训练的 ViT-Base 模型作为信息提取的主干网络。语义分割解码包括一个 UperNet 语义分割头以及一个使用 FCN 的辅助头,以提高分割准确性并确保训练期间的稳定收敛。实现的基线模型可应用于任何输入模态的场景。因此,在 Globe230k 数据集对比方法的选

取上，对于只有RGB输入的单模态场景，一些代表性的语义分割算法被选取作为基线模型，包括DeepLabv3+[30]、Segformer[25]、FCN[31]、OCNet[32]、PSPNet[33]、PointRend[34]、SwinTransformer[35]和ViT[8]。在多模态输入情况下，上述所提出的基线模型用于与本文的方法进行比较。

3.3 与其它方法的对比

3.3.1 BigEarthNet-MM实验结果

在BigEarthNet-MM数据集进行预训练之后，在预训练的编码器上添加一个线性分类头进行了微调，并在表2中报告了测试集上的mAP指标。分类头在单模态和多模态数据上分别进行训练，表中S1代表仅使用SAR数据进行训练，S2仅使用多光谱数据进行训练，S1+S2表示同时使用两种模态数据进行训练。此外，为了在数据稀缺条件下评估预训练模型的有效性，按照Fus-MAE的设置，在1%的训练数据上进行了微调实验，训练中冻结了预训练模型权重，仅训练线性分类器，结果展示在表2的mAP-1% labels列中。

与其它方法相比，本文方法在100% labels条件下得到的mAP值为90.4%，在1%labels条件下达到的mAP值为71.4%，证明了本文预训练方法的有效性，即使在数据有限的情况下，本文的方法也有较好的实验结果。此外，在Fus-MAE等基线模型上观察到，与仅使用多光谱图像训练相比，同时使用SAR和多光谱图像训练会导致性能下降。分析认为，这是由于对比方法中两种模态的信息融合不足，模态间的巨大差异导致了性能下降。相比之下，本文的预训练方法更好地促进了多模态融合，因此，引入更多模态进行训练可进一步带来提升。

3.3.2 SEN12MS实验结果

在BigEarthNet-MM数据集上进行预训练后，在SEN12MS数据集上进行了迁移实验，结果如表3所示。

结果表明，本文方法在SEN12MS数据集上展现了最优的性能，达到的整体准确率为77.2%、召回率为60.8%，F1分数为63.9%，超越了以往的预训练方法。我们分析认为，这主要归功于跨尺度一致性和融合模块的引入，使得预训练模型的特征提取和融合更加有效，能很好地将其能力转移到下游任务。

3.3.3 C2Seg-BW实验结果

C2Seg-BW数据集则更具有挑战性，因此进一步验证了BigEarthNet-MM预训练模型在该数据集上的有效性。

C2Seg-BW数据集上的迁移学习结果如表4所示，可看到本文方法显著优于其它对比方法。值得注意的是，由于预训练数据集BigEarthNet-MM包含多光谱和SAR两种模态，迁移到C2Seg-BW数据集也是基于这两种模态进行训练。然而，与那些额外利用了高光谱模态数据的方法相比，本文方法仍然展现出更优越的性能。具体而言，本文方法可达到mIoU值为14.79%、OA值为46.02%以及mF1值为20.94%。此外，本文方法在大多数类别上都有改进，尤其是其它对比方法难以分割的类别，如“P”（牧场）、“OS”（无植被的开放空间）和‘IW’（内陆湿

表2 不同算法在BigEarthNet-MM数据集上的定量结果

Table 2 Quantitative results of different algorithms on BigearthNet-MM dataset.

方法	mAP-100% labels			mAP-1% labels		
	S1	S2	S1+S2	S1	S2	S1+S2
Dino-MM[23]（2022）	69.7	83.9	84.6	52.7	58.7	60.3
SatViT[14]（2022）	75.4	85.6	85.5	52.4	58.5	58.0
Fus-MAE[22]（2024）	75.5	87.9	87.9	57.8	70.0	68.7
Ours	**78.6**	**89.8**	**90.4**	**59.3**	**70.6**	**71.4**

表3 不同算法在SEN12MS数据集上的定量结果

Table 3 Quantitative results of different algorithms on SEN12MS dataset.

方法	Top1-Acc	Top3-Acc	Precision	Recall	F1-score
Dino-MM[23]（2022）	58.8	91.0	71.4	58.8	59.7
SatViT[14]（2022）	59.1	91.8	75.2	59.1	59.9
Fus-MAE[22]（2024）	59.8	92.4	75.2	59.8	62.0
Ours	**60.9**	**93.4**	**77.2**	**60.8**	**63.9**

表 4　不同算法在 C2Seg-BW 数据集上的定量结果

Table 4　Quantitative results of different algorithms on C2Seg-BW dataset.

方法	模态	类别准确率													mIoU	OA	mF1
		SW	SN	UF	ICT	MDCS	AVA	AL	PC	P	F	S	OS	IW			
AdaptSeg[26]（2018）	MSI+SAR	59.57	**17.05**	25.06	32.77	1.94	8.99	9.11	0.30	0.00	32.98	0.00	0.00	0.00	8.92	29.26	14.44
DSAN[27]（2021）	HSI+MSI+SAR	0.00	0.00	**42.08**	25.68	1.22	9.55	**25.37**	0.00	0.00	46.85	0.00	0.00	0.00	7.09	18.55	11.60
DualHR[28]（2018）	HSI+MSI+SAR	60.51	0.29	0.76	24.19	0.26	4.19	0.00	0.00	0.00	32.70	0.22	0.33	0.01	6.17	31.97	9.53
SegFormer[25]（2021）	HSI+MSI+SAR	78.49	0.05	30.90	20.38	2.52	10.79	18.01	1.77	0.00	38.72	10.48	0.01	0.01	10.89	33.56	16.32
FastFCN[24]（2022）	HSI+MSI+SAR	45.39	2.38	38.44	27.63	0.86	8.83	0.00	0.00	0.00	4.14	0.00	0.00	0.00	5.96	21.22	9.82
HighDAN[20]（2023）	HSI+MSI+SAR	78.37	0.58	40.04	**43.67**	1.67	9.28	0.43	0.10	0.00	**47.22**	0.26	0.00	0.00	11.92	39.58	17.69
Ours	MSI+SAR	**79.75**	1.99	30.68	22.95	**3.91**	**11.58**	21.44	**1.83**	**0.06**	41.55	**10.80**	**1.12**	**0.64**	**14.79**	**46.02**	**20.94**

地),在这些类别上本文方法也取得了良好结果。此外,对于“SN”、“UF”和“ICT” 等类别,其分割精度受到不同城市间语义差距的限制。与考虑了域差异的方法如 HighDAN 和 AdaptSeg 等相比,由于本文方法未考虑到不同城市的域差距,因此在这些类别上准确率持平或略低,但总体性能上仍显著优于其它对比方法,验证了我们预训练模型迁移到不同下游任务的有效性。

3.3.4　Globe230k 实验结果

为了验证预训练方法的可扩展性,在 Globe230k 数据集上进行了实验,并在表 5 中展示了与其它方法的对比结果。

首先,为了验证不同模态对分割任务的重要性,评估了实现的基线模型与经典的单模态语义分割算法在输入不同单模态图像时的表现。实验结果表明,所实现的基线模型在输入为 RGB 图像时取得了很好的结果,相比之下,其它单模态训练的模型都有不同程度的性能下降,尤其是,使用 DEM 训练的模型,表现最差。这主要是由不同模态的特性和所包含的信息决定的。DEM 主要包含高程信息,在分割任务中的相关性有限,因此表现明显不佳。

其次,我们分别研究了三种(RGB、SAR、NDVI)和四种(RGB、SAR、NDVI、DEM)模态输入的多模态场景。与仅使用 RGB 模态作为输入相比,基线模型中的简单拼接来融合多模态特征并没有提高分割精度,反而导致了性能下降。这应是因为简单拼接可能导致差异较大的特征间冲突,阻碍了模型学习。相比之下,得益于跨尺度一致性约束和设计的两个融合模块的引入,本文提出的方法超越了开源的对比模型和所实现的基线模型,在大多数类别的分割准确率上均取得了最好结果。与比较方法中的最佳结果相比,本文模型的 mIoU 提高了 1.73%,OA 提高了 0.63%,mF1 提高了 2.38%。此外,与基于拼接的基线模型观察到的情况不同,即使引入了

分辨率较低、在语义分割任务表现不佳的DEM模态输入，分割认任务上的效果仍然可以进一步改进，为“森林”类别带来了0.93%的收益，并且其它类别上也取得了很好的结果。这表明本文的方法能从新模态中进一步提取互补信息，显著提高下游任务的性能。

另外，从表5的实验结果可观察到，在冰/雪类别效果相对较差，这是因为该类别样本量较少，不足以使得预训练模型很好地理解该类别特征；草地和不透水表面准确率也略低于对比方法，这是因为这两个类别特征明显，相对简单，因此不同模型均能取得较好的结果，我们的方法未展现出显著的差异，这一点可从各模型在这些类别的分类准确率差异较小得到验证；总体来看，我们方法的效果显著优于其它对比方法，证明了我们方法的有效性。

3.4 消融实验分析

3.4.1 模型结构对结果的影响

为了验证不同模型结构对最终结果的影响，在BigEarthNet-MM数据集上进行了实验评估，表6给出了实验结果。首先，从完整模型结构中分别移除了CAF和CMF，观察到在移除它们后最终结果有所下降。具体来说，移除CAF导致性能下降了3.5%；而移除CMF导致性能下降了1.7%。分析表明，编码器中缺少CAF阻碍了跨模态信息的早期显式整合，限制了编码器处理多模态输入时互补信息的聚合，导致信息提取不充分，显著影响下游任务；另一方面，解码器中缺少CMF会影响单模态标记从所有标记中聚合补充信息的能力。此外，从完整方法中移除跨尺度对比学习，导致性能下降了1.5%，证明了跨尺度一致性约束的有效性。这种约束确保了对遥感图像内部信息的深层次挖掘，从而产生了更强的图像表示特征。总之，本文提出的融合模块和跨尺度对比学习分支对于获得鲁棒的预训练表示至关重要，有效促进了多模态输入交互和补充信息的整合。

3.4.2 图像裁切比例对结果的影响

在本节中，探讨了在动态生成子图进行对比学习时，原始图像的不同裁切比例对最终结果的影响。我们尝试了各种裁切比例，其实验结果如表7所示。

结果表明，当裁剪比例较小，设定为0.3时，性能显著下降。随着裁切比例的逐渐增加，分类性能明显提高。然而，进一步将裁切比例增加到0.7时，反而会导致性能下降。这一趋势与关于跨尺度一致性的假设相符。较小的裁切比例在子图像和原始图像之间引入了较大的差异，破坏了跨尺度一致性，并通过对比学习引入了分布噪声，影响了预训练的模型性能；相反，较大的裁切比例导致子图像与原始图像非常相似，降低了跨尺度一致性和对比学习的效果。在最终的实验中，选择为每个样本采用0.4到0.6之间的随机裁切比例。这种策略旨在更好地利用跨尺度信息，同时保持图像尺度的多样性，进一步提高模型的性能。

3.4.3 对比损失比例对结果的影响

另外，对对比损失权重α进行了消融研究，该权重调节对比损失的比例。分析表明，当对比损失权重较小时，模型倾向于更多地关注重建任务。这对解码器中引入的跨模态融合机制是有益的，因为它有助于更好地整合多模态信息，对下游应用更有利；然而，在某种程度上，这种设置可能会忽视对比学习引入的一致性约束，可能导致单模态信息提取不足。相反，当对比损失权重较大时，模型可能会过分强调编码器对单模态输入的跨尺度特征提取，忽略了跨模态特征的融合。这些分析在表8所示的消融研究中得到了进一步验证。因此，在所有实验中，选择对比损失权重为0.5。

表8 不同对比损失权重在BigEarthNet-MM数据集上的结果

Table 8 Quantitative results of different contrastive loss weights on the BigEarthNet-MM dataset.

对比损失权重	mAP-100% labels
0.3	89.8
0.5	**90.4**
0.7	90.3
1.0	90.0
2.0	88.4

3.5 可视化结果

在本节中，将对本文方法在下游任务中的效果进行可视化展示和分析。图3展示了在Globe230k数据集上10个语义类别的分割结果。

图3的前四列展示了原始图像输入，包括RGB、SAR、DEM和NDVI图像；第五列展示了样本的语义分割标签；第六和第七列展示了两个对比方法，即Segformer和Swin Transformer的可视化结果；最后一列展示了本文方法获得的结果。通过实验结果对比可看到，本文方法可得到噪声更少、更真实、类别间过渡更平滑的分割结果，它在大多数前景类别中

表5　不同算法在Globe230k数据集上的定量结果

Table 5　Quantitative results of different algorithms on Globe230k dataset.

方法	模态	类别准确率										mIoU	OA	mF1
		农田	森林	草地	灌木	湿地	水域	苔原	不透水表面	荒地	冰/雪			
FCN[31] (2015)	RGB	90.73	95.37	71.90	70.20	68.23	88.34	0.30	94.03	59.59	96.26	67.97	89.12	/
OCNet[32] (2018)	RGB	89.98	89.71	75.34	57.58	49.62	82.67	27.91	80.11	74.78	91.45	70.12	87.45	/
DeepLabv3+[30] (2018)	RGB	90.89	95.10	77.46	75.74	70.35	95.33	0.00	**94.29**	90.77	93.43	70.56	90.46	/
PSPNet[33] (2017)	RGB	91.41	94.91	77.32	74.32	77.62	93.93	34.21	90.47	90.35	97.51	73.12	90.49	/
PointRend[34] (2020)	RGB	92.10	95.78	64.56	72.49	71.16	89.07	4.22	86.67	87.49	97.72	67.08	88.64	/
ViT[8] (2020)	RGB	78.83	89.12	55.84	53.81	47.63	82.87	0.00	80.59	73.19	91.19	65.31	87.70	/
Segformer[25] (2021)	RGB	92.65	94.64	55.14	56.13	41.96	92.95	0.00	92.36	90.05	95.27	62.64	86.69	/
SwinTransformer[35] (2021)	RGB	89.42	96.52	**79.33**	71.73	77.28	94.99	42.64	93.06	90.74	98.06	75.72	90.90	/
Baseline	RGB	92.27	97.02	76.32	79.49	80.62	94.40	29.89	92.32	94.28	**98.32**	77.28	91.93	85.67
	NDVI	87.63	89.98	31.00	25.35	29.77	89.13	0.00	85.46	77.41	75.11	49.56	77.83	68.07
	SAR	82.32	88.57	47.72	34.70	45.88	85.61	0.00	83.05	75.74	51.85	48.94	77.33	68.51
	DEM	69.26	78.14	2.12	1.07	4.06	76.26	0.00	1.34	48.72	40.36	23.57	57.33	36.18
Baseline (Concat)	RGB+NDVI+SAR	91.73	95.98	74.36	75.72	78.34	92.32	3.27	91.21	92.61	96.52	72.28	88.92	83.07
	RGB+NDVI+SAR+DEM	90.90	96.02	73.48	75.01	77.65	92.59	1.01	91.25	93.50	96.15	70.43	87.98	82.24
Ours	RGB+NDVI+SAR	94.38	96.45	76.38	79.98	80.03	96.12	59.64	93.60	94.89	96.99	78.96	92.15	87.69
	RGB+NDVI+SAR+DEM	**94.76**	**97.38**	76.42	**79.66**	**80.98**	**96.36**	**60.68**	93.56	**95.42**	97.55	**79.01**	**92.56**	**88.05**

实现了更好的分割结果,整体分布与真实标签更加接近。例如,在第一行样本中,本文方法可有效区分“农田”和“草地”等类似类别,而其它方法将“草地”误识别为“农田”。这可归因于本文方法中提出的两个融合模块,很好地促进了多模态互补信息的有效融合,而其它方法则表现相对较差。此外,本文方法提供了不同类别间更清晰的边界划分。在第二个例子中,本文方法实现了与标签相似的边界预测结果,而其它方法在类别间展示了更粗糙的过渡,并且在处理图像细节方面存在困难,这得益于我们引入的跨尺度一致性实现了更鲁棒的预训练,对于图像特征提取更加有效。上述结果与分析验证了本文方法的有效性。

3.6　小结

通过对所提出方法与其它方法的实验结果对比以及消融实验的分析,所提出方法具有如下主要优点:

精度高:本文方法在多模态遥感图像融合的各种下游任务中都展示出了优异的实验结果,在实验结果的精度上超过了目前的SOTA多模态遥感图像融合方法。我们认为,这一点主要得益于所设计的CAF和CMF融合机制,在编码器和解码器中都有效地促进了多模态互补信息的交互整合;其次,跨尺度一致性约束的引入则进一步提升了单模态图像信息提取的效果,因此本文的多模态遥感图像融合方法在各种下游任务上表现出色。

可扩展性强:提出的方法对不同数目和种类模态的输入都适用,在包括RGB图像、SAR、多光谱图像等多种模态的数据集上都取得了相比其它方法更好的结果,在场景分类、语义分割下游任务上表现了出色的迁移效果,展现出了强大的可扩展能力。这主要是因为设计的多模态融合结构能够简单方便地适用于更多模态数目和种类的输入,预训练的鲁棒特征能够迁移到各种不同的下游任务,从而实现了有效的多模态遥感图像融合。

超参数易调节:在所提出方法中,图

表6 不同结构在BigEarthNet-MM数据集上的结果

Table 6 Quantitative results of different architectures on the BigEarthNet-MM dataset.

模型结构	mAP-100% labels
w/o CAF	86.9
w/o CMF	88.7
w/o Contrastive Learning	88.9
Ours	**90.4**

表7 不同裁切比例在BigEarthNet-MM数据集上的结果

Table 7 Quantitative results of different cropping ratios on the BigEarthNet-MM dataset.

裁切比例	mAP-100% labels
0.3	86.7
0.4	90.1
0.5	90.0
0.6	90.3
0.7	89.7
random (0.4-0.6)	**90.4**

像的裁切比例、不同损失函数的比例等是需要调节的超参数。在本文所有实验中，在多个不同的数据集上均采用了相同的超参数设置，且在下游任务上都取得了很好的实验结果，而无需针对不同的数据集和下游任务重新进行超参数设置。因此，所提出方法还具有超参数容易调节的优点。

尽管提出的方法展现出了显著的优势，但它仍然存在某些局限性。虽然本文方法借助跨尺度对比学习，在各种下游任务上取得了显著的收益，但本文目前的方法是通过引入额外的对比学习分支来利用遥感图像的跨尺度一致性，在多模态融合编码器中互补信息聚合和提取的过程中，这一特性仍然没有得到充分利用。另外，本文方法没有考虑到分割等任务中的类间差异和域间差异，可能导致在不同类别或者不同领域数据得到差异化的结果。因此，如何将跨尺度信息直接集成到预训练模型的特征抽取和融合阶段中，并缓解类间差异和域间差异对结果的影响，是我们未来工作的一个方向。

4 结论

本文提出了一种基于自监督预训练与跨尺度对比学习的多模态遥感图像融合方法，以克服多模态遥感图像融合中模态特征聚合不足和忽视跨尺度信息的问题。为了增强多模态互补信息的整合，在多模态融合编码器中引入了CAF机制，用于多模态特征抽取过程中的信息聚合；在模态特定解码器中引入了CMF机制，促进单模态解码过程中对其余模态互补信息的整合；基于遥感图像内部的跨尺度信息一致性先验，在预训练过程中引入了跨尺度对比学习，以确保多模态遥感图像跨尺度特征的有效提取，为后续任务提供鲁棒的特征，最终实现了多模态遥感图像融合性能的显著提升。在多个公开数据集上的实验结果表明，所提议方法具有精度高、可扩展性强、超参数易调节等优点，这些特点对实际应用有重要意义。

在未来工作中，我们将通过深入研究预训练过程中跨尺度特征抽取和融合技术，以进一步增强多模态特征的学习，实现更有效的多模态遥感图像融合。

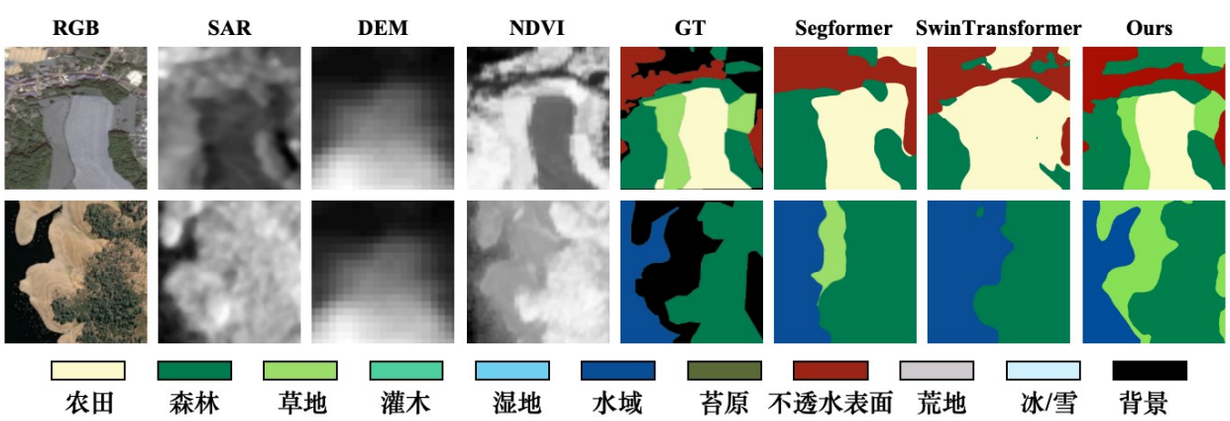

图3 Globe230k数据集的可视化结果

Fig. 3 Visual results in Globe230k dataset

References

[1] Kucharczyk M, Hugenholtz CH. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities [J]. Remote Sensing of Environment, 2021, 264: 112577.

[2] Yuan Q, Shen H, Li T, et al. Deep learning in environmental remote sensing: Achievements and challenges [J]. Remote Sensing of Environment, 2020, 241: 111716.

[3] Zhong Y, Su Y, Wu S, et al. Open-source data-driven urban land-use mapping integrating point-line-polygon semantic objects: A case study of Chinese cities [J]. Remote Sensing of Environment, 2020, 247: 111838.

[4] Li J, Hong D, Gao L, et al. Deep Learning in Multimodal Remote Sensing Data Fusion: A Comprehensive Review. International Journal of Applied Earth Observation and Geoinformation 112 (2022): 102926.

[5] Hong D, Gao L, Hang R, et al. Deep encoder – decoder networks for classification of hyperspectral and LiDAR data [J]. IEEE Geoscience and Remote Sensing Letters, 2020, 19(1): 1-5.

[6] Hong D, Gao L, Yokoya N, et al. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(5): 4340-4354.

[7] Mohla S, Pande S, Banerjee B, et al. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 92-93.

[8] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929, 2020.

[9] Roy SK, Deria A, Hong D, et al. Multimodal fusion transformer for remote sensing image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-20.

[10] Yao J, Zhang B, Li C, et al. Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-15.

[11] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16000-16009.

[12] Chen Y, Bruzzone L. Self-supervised SAR-optical data fusion of Sentinel-1/-2 images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-11.

[13] Montanaro A, Valsesia D, Fracastoro G, et al. Semi-supervised learning for joint SAR and multispectral land cover classification [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[14] Fuller A, Millard K, Green JR. Satvit: Pretraining Transformers for Earth Observation. IEEE Geoscience and Remote Sensing Letters 19 (2022): 1-5.

[15] Zhang L, Zhang Z, Guo W, et al. 3DMAE: Joint SAR and Optical Representation Learning with Vertical Masking. IEEE Geoscience and Remote Sensing Letters (2023).

[16] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning [C]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[17] Tang M, Cozma A, Georgiou K, Qi H. Cross-Scale MAE: A tale of multiscale exploitation in remote sensing [C]. Proceedings of the Advances in Neural Information Processing Systems, 2024, 36.

[18] Sumbul G, De Wall A, Kreuziger T, et al. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets] [J]. IEEE Geoscience and Remote Sensing Magazine, 2021, 9(3): 174-180.

[19] Schmitt M, Hughes LH, Qiu C, et al. SEN12MS--A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion [J]. arXiv preprint arXiv:1906.07789, 2019.

[20] Hong D, Zhang B, Li H, et al. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks [J]. Remote Sensing of Environment, 2023, 299: 113856.

[21] Shi Q, He D, Liu Z, et al. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping [J]. Journal of Remote Sensing, 2023, 3: 0078.

[22] Chan-To-Hing H, Veeravalli B. Fus-MAE: A cross-attention-based data fusion approach for Masked Autoencoders in remote sensing [J]. arXiv preprint arXiv: 2401.02764, 2024.

[23] Scheibenreif L, Mommert M, Borth D. Contrastive self-supervised data fusion for satellite imagery [J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2022, 3: 705-711.

[24] Wu H, Zhang J, Huang K, et al. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation [J]. arXiv preprint arXiv:1903.11816, 2019.

[25] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers [C]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[26] Tsai YH, Hung WC, Schulter S, et al. Learning to adapt structured output space for semantic segmentation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7472-7481.

[27] Zhu Y, Zhuang F, Wang J, et al. Deep subdomain adaptation network for image classification [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32 (4): 1713-1722.

[28] Ren B, Ma S, Hou B, et al. A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification [J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 112: 102896.

[29] Zhang B, Chen T, Wang B. Curriculum-style Local-to-Global Adaptation for Cross-Domain Remote Sensing Image Segmentation. IEEE Transactions on Geoscience and Remote Sensing 60 (2021): 1-12.

[30] Cheng B, Collins MD, Zhu Y, et al. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

2020: 12475-12485.
[31] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.
[32] Yuan Y, Huang L, Guo J, et al. OCNet: Object context for semantic segmentation [J]. International Journal of Computer Vision, 2021, 129(8): 2375-2398.
[33] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2881-2890.
[34] Kirillov A, Wu Y, He K, et al. Pointrend: Image segmentation as rendering [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9799-9808.
[35] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022. https://bigearth.net/