中图法分类号:TP753　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)09-3153-18

论文引用格式:Tao C, Guo X, Hu K Y, Shen Y X and Wang H. 2025. Language-guided cross-spatiotemporal domain adaptation for remote sensing image semantic segmentation. Journal of Image and Graphics, 30(9):3153-3170(陶超, 郭鑫, 胡柯彦, 沈羽翔, 王昊. 2025. 以语言为媒介的遥感图像跨时空领域自适应语义分割. 中国图象图形学报, 30(9):3153-3170)[DOI:10.11834/jig.240640]

以语言为媒介的遥感图像跨时空领域自适应语义分割

陶超[1],郭鑫[1],胡柯彦[1],沈羽翔[1],王昊[1,2*]

1. 中南大学地球科学与信息物理学院,长沙　410083; 2. 内蒙古大学计算机学院,呼和浩特　010021

摘　要: **目的**　随着视觉大模型的发展,利用多源无标注遥感影像预训练学习全局视觉特征,并在局部目标任务上进行迁移微调,已成为遥感影像领域自适应的一种新范式。然而,现有的全局预训练策略主要聚焦于学习低级的通用视觉特征,难以捕捉复杂、高层次的语义关联。此外,微调过程中使用的少量标注样本往往只反映目标域的特定场景,无法充分激活全局模型中与目标域匹配的领域知识。因此,面对复杂多变的遥感影像跨时空领域偏移,现有方法得到的全局模型与目标任务之间仍然存在巨大的语义鸿沟。为应对这一挑战,本文提出一种语言文本引导的"全局模型预训练—局部模型微调"的领域自适应框架。**方法**　提出框架针对遥感数据的时空异质性特点,借助大型视觉语言助手LLaVA(large language and vision assistant)生成包含季节、地理区域及地物分布等时空信息的遥感影像文本描述。通过语言文本引导的学习帮助全局模型挖掘地物的时空分布规律,增强局部任务微调时相关领域知识的激活。**结果**　在对比判别式、掩码生成式和扩散生成式3种不同全局预训练策略上设置了3组"全局—局部"跨时空领域自适应语义分割实验来验证提出框架的有效性。以全局→局部(长沙)为例,使用语言文本引导相比于无文本引导在3种不同预训练策略上分别提升了8.7%、4.4%和2.9%。同样地,提出框架在全局→局部(湘潭)和全局→局部(武汉)上也都有性能提升。**结论**　证明了语言文本对准确理解跨时空遥感影像中的语义内容具有积极影响。与无文本引导的学习方法相比,提出框架显著提升了模型的迁移性能。

关键词: 遥感影像;语义分割;领域自适应;视觉语言模型;时空异质性

Language-guided cross-spatiotemporal domain adaptation for remote sensing image semantic segmentation

Tao Chao[1], Guo Xin[1], Hu Keyan[1], Shen Yuxiang[1], Wang Hao[1,2*]

1. School of Geosciences and Info-physics, Central South University, Changsha　410083, China;

2. College of Computer Science, Inner Mongolia University, Hohhot　010021, China

Abstract: **Objective** With the development of large-scale visual models, pre-training on multi-source unlabeled remote sensing images to learn global visual features and fine-tuning for target tasks has become a new paradigm for domain adaptation of remote sensing image semantic segmentation. Across spatiotemporal domains, joint distribution shifts, comprising

收稿日期:2024-11-07;**修回日期**:2025-01-15;**预印本日期**:2025-01-22

*通信作者:王昊　haowang7cc@gmail.com

基金项目:国家自然科学基金项目(42171376,42471419);湖南省杰出青年科学基金(2022JJ10072);内蒙古大学青年学术人才科研启动项目(10000-A25206020)

Supported by: National Natural Science Foundation of China (42171376, 42471419); Natural Science Foundation of Hunan for Distinguished Young Scholars (2022JJ10072); High-Level Talents Introduction Project of Inner Mongolia University (10000-A25206020)

feature and label distribution shifts, occur due to variations in lighting, weather, phenology, natural landscapes, and human-made environments. This spatiotemporal heterogeneity complicates the accurate assessment of domain relevance, challenging the applicability of most "local-local" domain adaptation methods. In contrast, "global-local" learning strategies, which extract general visual features from a broad spectrum of unlabeled data, enhance the relevance of knowledge across domains. However, current global pre-training approaches primarily focus on low-level feature learning, which limits the ability to capture complex, high-level semantic relationships. Furthermore, during the fine-tuning phase with limited annotated samples, these samples often reflect only specific scenarios within the target domain, making it insufficient to fully activate the relevant knowledge within the global model. Consequently, a major semantic gap persists between the globally trained models and the actual task requirements. This challenge manifests in two aspects: 1) a mismatch between the global pre-training objectives and the requirements of the target semantic segmentation task, as pre-trained features focused on low-level information may not align well with the need for deep semantic associations, thus limiting the effectiveness of model transfer; and 2) insufficient learning of target-specific semantic features during local fine-tuning due to the limited representativeness of the few annotated samples, which may fail to encompass the full range of variability within the target domain. To address these issues, this paper proposes a language-guided "global pre-training-local fine-tuning" framework for domain adaptation to overcome the challenges associated with cross-spatiotemporal domain shifts of remote sensing images. **Method** The proposed framework addresses the spatiotemporal heterogeneity of remote sensing data by leveraging a large-scale visual-language assistant called large language and vision assistant (LLaVA) to generate textual descriptions of remote sensing images that include information on season, geographical area, and distribution of ground objects. Rich in semantic and contextual information, these language texts, when combined with visual features, enable the model to better understand the deep semantic associations across different remote sensing images. For the generative pre-training strategy of the global model, the complex contextual information in long texts aids in the reconstruction and generation of detailed image content. For the discriminative pre-training strategy, clear and concise short texts are beneficial for contrastive learning optimization. Therefore, this paper proposes a method for generating long and short textual descriptions of remote sensing images, tailored to the different pre-training strategies of the global model. Within the global pre-training-local fine-tuning domain adaptation framework, the language text not only guides the global model in capturing and understanding the spatiotemporal distribution patterns within remote sensing images but also facilitates the rapid activation of associated domain knowledge in the local model: 1) during the global pre-training phase, textual descriptions that include information about the season, geographic region, and distribution of ground objects guide the model in learning associations between visual features and semantic information, thereby capturing and understanding the spatiotemporal patterns within the imagery. 2) During the local fine-tuning phase, similar textual descriptions assist in rapidly activating relevant domain knowledge embedded within the global model. **Result** Three sets of global-local cross-spatiotemporal domain adaptation experiments for semantic segmentation were conducted, comparing discriminative, masked generative, and diffusion generative pre-training strategies to validate the effectiveness of the proposed framework. Using the example of global-local (Changsha, CS), the employment of language text guidance, compared with no text guidance, has resulted in performance improvements of 8.7%, 4.4%, and 2.9% across three different pre-training strategies, with similar performance enhancements observed for global-local (Xingtan, XT) and global-local (Wuhan, WH). Compared with traditional local-local learning methods and global-local learning methods without text guidance, the proposed framework significantly enhances the model's transfer performance. **Conclusion** This paper pioneers the exploration and validation of the positive role of language text in mitigating spatiotemporal domain shifts in remote sensing imagery, introducing a language-guided global pre-training-local fine-tuning framework for domain adaptation. This framework uses textual descriptions of remote sensing images to facilitate the global model's learning of spatiotemporal distribution patterns of ground objects during pre-training and enhances the activation of relevant domain knowledge during local fine-tuning. Three multi-source, cross-spatiotemporal semantic segmentation experiments demonstrate that the proposed framework significantly improves model transfer performance compared with traditional local-local domain adaptation methods and global-local methods without text guidance. Future research will focus on two main directions: 1) the impact of language text on model transfer performance over larger spatiotemporal scales will be investigated. While this study conducted exploratory experiments using remote

sensing data sampled across four seasons in the Hunan and Hubei regions, future work will aim to extend the spatial coverage and temporal span to assess the feasibility of applying the proposed framework at national and even global levels. 2) More refined textual description methods for remote sensing images will be developed, potentially incorporating meteorological data (e. g. , temperature and precipitation) and topographic information to enrich the content of the descriptions.

Key words: remote sensing image; semantic segmentation; domain adaptation; visual-language model; spatiotemporal heterogeneity

0 引　言

遥感影像语义分割是遥感影像分析中很基础的研究，可以为许多遥感应用提供重要支持，如土地利用监测（董荣胜 等，2022）、环境监测和城区规划等（李林娟 等，2024）。在真实环境下的遥感影像语义分割任务中，训练数据（源域）和测试数据（目标域）往往采集自不同时相和不同地理空间区域，导致领域之间不满足独立同分布假设，传统的深度学习方法不再适用（Tuia 等，2016）。不同时相的遥感影像由于光照、天气和地类物候等视觉环境变化导致领域之间具有特征分布偏移，不同地理空间区域的遥感影像由于自然地貌和人文景观等地理场景变化导致领域之间具有标签分布偏移，即跨时空域的遥感影像具有特征分布偏移与标签分布偏移构成的联合分布偏移（Wang 等，2022a）。因此，通过减小领域偏移，允许源域学习模型在标注匮乏的目标域上也具有良好泛化能力的域自适应学习技术受到广泛关注。

传统的域自适应学习方法遵循“局部到局部”的学习范式，其假设不同局部来源的领域之间具有一定的相关性（郑向涛 等，2024）和可共享迁移的隐含结构（Wang 和 Deng，2018）。根据“如何迁移”的策略机制的不同，将现有方法分为以下几类：1）领域不变特征自适应。该类方法关心的首要问题是如何度量源域和目标域之间的特征分布差异，其次是如何有效减小不同领域之间特征分布的差异来学习领域不变的稳定特征表示（Tahmoresnezhad 和 Hashemi，2017）。常用的特征分布差异度量方法包括最大均值差异（maximum mean discrepancy，MMD）（郑宗生 等，2020）、生成对抗网络（generative adversarial network，GAN）（Goodfellow 等，2014）等。其中，基于 MMD 的思路有多核最大均值差异匹配（Long 等，2015）、联合最大均值差异匹配（Long 等，2017）；基于 GAN 的思路有以对抗方式在输出空间对齐不同领域预测熵值图的特征匹配（Vu 等，2019）、顾及类别权重的对抗领域特征匹配（Xiao 和 Zhang，2021）、顾及全局—局部对齐的生成式特征匹配（Wang 等，2024）。2）图像风格迁移。该类方法关注的是如何有效减少源域和目标域之间的风格差异。当源域图像的风格可以转换为目标域图像的风格，使迁移样本在风格外观上与目标域更接近，那么这些风格迁移样本就可以被进一步纳入训练集，从而有效减少源域和目标域之间的分布差异，帮助模型更好地学习和识别跨域的共享内容。主要思路包括：基于循环一致生成对抗网络（Zhu 等，2017）生成不同季节不同光照风格影像的风格迁移（Zhao 等，2021）、对光谱特征统计量自适应加权匹配的风格迁移（Zhang 等，2022）、基于图像潜空间滤波分离关联特征进行匹配的风格迁移（Liu 等，2024）。3）基于伪标签的领域自适应。基于伪标签的领域自适应学习使用已经在标注源域数据上训练好的模型来为未标注的目标数据生成伪标签，然后将这些带有伪标签的数据重新纳入训练集，以此不断迭代提高模型的性能和泛化能力。这类方法的一个关键问题是如何有效地筛选和过滤噪声伪标签，主流的思路有置信度阈值过滤的伪标签训练（Zou 等，2018）、软伪标签来正则化领域自适应策略（Zou 等，2019）以及结合像素关系与类别特征的可靠伪标签学习策略（Zhao 等，2023）。

但是，由于遥感时空数据的复杂性，不同时空采集的源域和目标域之间的领域相关性难以准确评估。例如，假设源域是一组夏季在长江中下游平原采集的遥感影像，目标域是一组冬季在东北平原地区采集的遥感影像。虽然这两个领域都包含耕地、林草和建筑物等地类要素，但由于季节和地理环境的巨大差异，这些地类要素在图像中的表现形式可能会有很大不同：长江中下游平原夏季的耕地和林草绿色盎然；东北平原冬季的耕地和林草覆盖着厚厚的积雪。在目标领域标注匮乏的情况下，很难准

确评估源域和目标域之间各种地类特征的相关性程度。因此，面对复杂多变的遥感影像跨时空领域偏移，“局部到局部”的域自适应学习范式的相关性假设很难决策是否成立，这极大地限制了域自适应学习技术的应用推广。

近年来，随着视觉大模型的发展，“全局模型预训练—局部模型微调”已成为域自适应学习的一种新范式（陶超 等，2021）。与“局部到局部”的域自适应学习范式不同，“全局到局部”的学习范式在更广泛的范围内确保了相关性假设的适用性，其分为两个步骤：首先，在大规模无标注样本上训练全局模型，以学习遥感影像的通用特征；然后，使用少量标注样本对全局模型进行微调，以获得适用于下游任务（如语义分割）的局部模型。在无标注图像样本上预训练全局模型的学习方法一般可以分为两种类型：1）基于判别式的全局预训练模型，如SimCLR（Chen等，2020）、TOV（Tao等，2023b）；2）基于生成式的全局预训练模型，如MAE（masked autoencoders）（He等，2022）、DDPM（denoising diffusion probabilistic models）（Ho等，2020）、DDIM（denoising diffusion implicit models）（Song等，2022）。尽管这些学习方法摆脱了对图像样本标注的依赖，但是仍然存在“语义鸿沟”（李德仁 等，2014）挑战：难以将低级的视觉特征与高级的语义概念建立联系。在域自适应学习过程中，这一挑战表现在两个方面：1）全局预训练时与目标任务需求不匹配。全局模型在预训练过程中主要依赖低级视觉特征挖掘领域间的相关信息，难以捕捉到深层次的语义关联信息。因此，预训练特征可能存在与目标语义分割任务需求不匹配的问题，限制了全局模型迁移到目标任务的有效性。2）局部微调时目标语义特征学习不充分。微调所使用的少量标注样本可能仅仅反映目标域内某些特定的场景、地物或地貌，不能完全反映目标域中所有可能出现的样本变化模式。因此，局部模型在微调时存在目标语义特征学习不充分的问题。

语言文本是一种不受图像领域偏移影响的信息，同时语言文本具有丰富的语义和上下文信息，通过将语言文本与视觉特征相结合，学习模型能够更好地理解不同遥感影像中深层次的语义关联信息。最近，视觉语言预训练模型（张浩宇 等，2022）引起研究人员的广泛关注，如基于判别式训练的CLIP（contrastive language-image pre-training）（Radford等，2021）和基于生成式训练的BEiT（bidirectional encoder representation from image transformers）（Wang等，2022b）、Stable Diffusion（Rombach等，2022）。视觉语言模型的不同预训练策略使用的文本描述并不相同：1）判别式视觉语言模型通过对比学习建立图像与文本之间的对齐关系，其主要关注于实例级别的特征对齐，如CLIP。有研究表明（Gou等，2024），尽管长文本描述包含丰富的细节信息，但是对比学习在图文实例特征对齐过程中难以捕捉这些文本细节与图像特征之间的关联，可能阻碍图像与文本之间对齐关系的建立。相比之下，短文本描述中的信息要更加明确，有利于对比学习建立图像与文本的稳定对齐关系。因此，大多数基于对比学习的判别式预训练模型普遍使用短文本描述（Yang等，2022）。2）生成式视觉语言模型一般是将图像特征与文本特征组合构建联合表征，通过图像掩码生成（Wang等，2022b）或扩散去噪生成（Rombach等，2022）来学习图像与文本之间的关联信息，使用长文本更有助于对图像细节内容的重建生成（Yang等，2024）。此外，由于视觉语言预训练模型通常是在自然图像数据上进行训练的，这些模型并未考虑遥感影像中时空变化对模型迁移的影响，这导致模型在遥感领域的性能会明显下降。因此，研究针对遥感影像的文本描述方法，有助于充分发挥视觉语言预训练模型在遥感领域的应用价值。

本文提出一种顾及时空信息的语言文本描述方法，在3种不同的全局模型预训练策略下探索并验证了语言文本对于减小跨时空遥感影像领域偏移有积极影响，其中全局模型预训练策略包括图像掩码生成、扩散去噪生成和判别式对比学习3种。语言文本的积极作用表现在两个方面：1）在全局模型预训练过程中，包含有遥感影像季节、地理区域、地物分布等信息的文本描述能够引导全局模型学习视觉特征与文本描述中语义信息的关联，捕捉和理解遥感影像中的时空分布规律。2）在局部模型微调过程中，类似的包含时空信息的文本描述又能够帮助局部模型快速激活全局模型中与之关联的领域知识。在语言文本描述设计上，对于全局模型的生成式预训练策略（包括图像掩码、扩散去噪两种），长文本中复杂的上下文信息有助于对图像细节内容的重建生成，而对于全局模型的判别式预训练策略，信息明确的短文本有助于对比学习优化。因此，为适应不同

的预训练策略，本文提出顾及时空信息的遥感影像长短文本描述方法。

总体来说，本文的贡献包括以下几点：1）探索并验证了语言文本在减少跨时空遥感影像领域偏移中的积极作用，在“全局模型预训练—局部模型微调”的领域自适应学习范式中，语言文本既引导全局模型捕捉和理解遥感影像中的时空分布规律，又帮助局部模型快速激活全局模型中与之关联的领域知识。2）提出顾及时空信息的遥感影像的长短文本描述方法，该方法能够适应不同的预训练策略，有助于挖掘视觉语言模型在跨时空遥感影像领域自适应学习上的应用价值。3）通过对3个多源跨时空遥感影像领域自适应语义分割任务进行的大量实验，本研究证实了相比于传统的纯视觉领域自适应方法，采用语言文本引导的领域自适应学习策略显著提升了模型的迁移性能。此外，研究结果还揭示了相较于语言引导的全局模型判别式预训练策略，语言引导的全局模型生成式预训练方法展现出更优异的性能。

1 本文方法

“全局到局部”的领域自适应学习范式在更广泛的范围内确保了相关性假设的适用性，有效地避免了“局部到局部”的领域相关性不足的问题。但是，“全局模型预训练—局部模型微调”仍然存在“语义鸿沟”挑战：1）全局预训练时与目标任务需求不匹配；2）局部微调时目标语义特征学习不充分。因此，本文提出利用语言文本作为全局预训练模型和局部微调模型的语义信息连接桥梁。在1.1小节中，阐述了方法的总体框架。在1.2小节中，阐述了面向“全局—局部”领域自适应学习的遥感影像文本描述方法。在1.3小节中介绍了语言文本引导的全局模型预训练。在1.4小节中介绍了语言文本引导的局部模型微调。

1.1 总体框架

给定采样自多时空来源的大规模无标注遥感影像组成源域空间 $\mathcal{D}_s=\{\mathcal{D}_1,\mathcal{D}_2,\cdots,\mathcal{D}_m\}$ 以及采样自与源域空间具有时空相关性的但从未见过的目标域 $\mathcal{D}_t$。定义 $\mathcal{Z}_{te}$ 为时序演化潜变量空间，它包含了一系列在时间维度上对遥感影像变化产生影响的潜在变量。定义 $\mathcal{Z}_{sp}$ 为地理环境潜变量空间，它包含了一系列在地理空间维度上对遥感影像变化产生影响的潜在变量。用 $\mathcal{Z}^{(t)}=\{z_{te}^{(t)},z_{sp}^{(t)}\}$ 表示目标域的时序演化潜变量和地理环境潜变量构成的潜变量集合，用 $\mathcal{Z}^{(s)}=\left\{\left(z_{te}^{(i)},z_{sp}^{(i)}\right)\right\}_{i=1}^{m}$ 表示多源域的时序演化潜变量和地理环境潜变量构成的潜变量集合，满足 $\mathcal{Z}^{(t)}\subset\mathcal{Z}^{(s)}$。

如图1所示，本文将语言文本作为引导信息帮助模型捕捉低级视觉特征与高级语义概念的联系，以解决“全局模型预训练—局部模型微调”范式存在的语义鸿沟问题，所提框架包含两个学习过程。

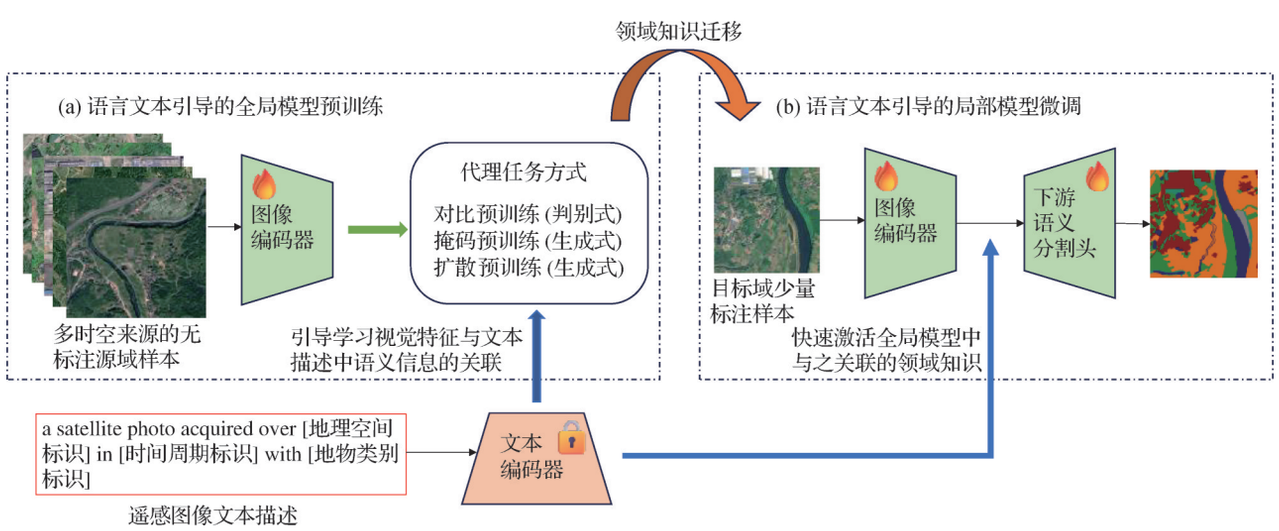


图1 语言文本引导的“全局模型预训练—局部模型微调”领域自适应学习框架图

Fig. 1 Language-guided “global pre-training - local fine-tuning” framework for domain adaptation learning

1)全局模型M_{global}预训练过程中,使用多时空来源的无标注遥感影像样本学习视觉表征,并且使用文本描述c_{Text}引导全局模型,从源域数据集中学习遥感影像视觉特征与文本描述中语义信息的关联,具体为

$$M_{\text{global}} = \arg\min_{M} \mathcal{L}_{\text{pretrain}}(\boldsymbol{D}_s, c_{\text{Text}}) \tag{1}$$

式中,$\mathcal{L}_{\text{pretrain}}$表示基于判别式或基于生成式的全局模型预训练损失函数,M_{global}由开源的图像编码器和文本编码器组成,训练时冻结文本编码器,仅更新图像编码器参数。2)局部模型微调过程中,使用目标域少量标注样本$\boldsymbol{\mathcal{D}}_{\text{t}}^{\#} \subset \boldsymbol{\mathcal{D}}_{\text{t}}$对已训练好的全局模型微调学习局部模型$M_{\text{local}}$,并且使用类似的文本描述$c_{\text{Text}}^{\#}$帮助局部模型快速激活全局模型中与之关联的领域知识,具体为

$$M_{\text{local}} = \arg\min_{M} \mathcal{L}_{\text{finetune}}\left(M_{\text{global}}, \boldsymbol{D}_{\text{t}}^{\#}, c_{\text{Text}}^{\#}\right) \tag{2}$$

式中,$\mathcal{L}_{\text{finetune}}$表示模型预测概率与样本标注之间的交叉熵损失函数。

1.2 遥感影像文本描述方法

本小节提出遥感影像文本描述方法,以构建面向“全局模型预训练—局部模型微调”领域自适应学习的遥感影像文本配对数据集,挖掘视觉语言模型在遥感影像跨时空领域自适应学习上的应用价值。跨时空遥感影像领域自适应学习所关注的是地物特征变化与时空变化之间的稳定模式。因此,遥感影像中的信息可以描述为“a satellite photo acquired over [地理空间标识] in [时间周期标识] with [地物类别标识]”,其中,[地理空间标识]是遥感影像采集的地理空间位置(如:湖南省长沙市),[时间周期标识]是遥感影像采集的季节(如:春、夏、秋、冬),而[地物类别标识]是对遥感影像中地物类别的描述。本文构建了长短两种不同的文本描述,长文本与短文本的主要区别体现在[地物类别标识]的详细程度,短文本简单陈述遥感影像包含的主要地物类别是什么;长文本进一步描述这些地物类别的空间依赖关系。此外将仅包含地理空间标识和时间周期标识的文本称为简单文本。表1对比了不同类型文本描述的差异。

给定采样裁剪大小为512 × 512像素的遥感影像,“地理空间标识”可以根据其采样的地理坐标获取,“时间周期标识”可以根据采样的元数据文件获取。对于“地物类别标识”的描述,本文先使用大型视觉语言助手LLaVA(large language and vision assistant)(Liu等,2023)根据提示指令自动生成,随后再由人工检查纠正具有严重认知错误的描述。“地理空间标识”、“时间周期标识”和“地物类别标识”共同构成遥感影像的文本描述。图2给出了遥感影像长文本和短文本描述的生成示例。

1)长文本的“地物类别标识”除了需要描述遥感影像中有什么地物,还需要描述这些地物类别的空间依赖关系。本文将512 × 512像素的遥感影像划分为5个方位:左上(upper left)、右上(upper right)、中间(center)、左下(lower left)和右下(lower right)。给定地物类别的名称集合,如('farmland', 'forest', 'water', 'artificial facilities')。对于任意遥感影像,使用LLaVA对5个方位的局部场景信息分别按照以下指令模板进行文本描述生成:“You are an excellent visual language assistant who is able to describe scenes and objects ('farmland', 'forest', 'water', 'artificial facilities') based on the content of an image

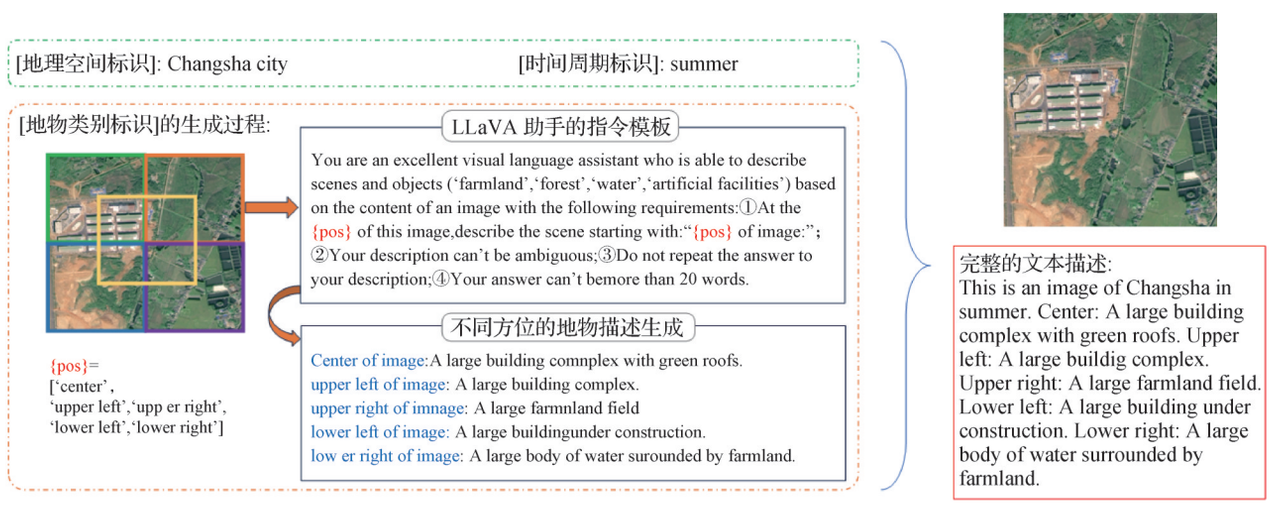
表1 不同文本类型的差异

Table 1 Differences among various text types

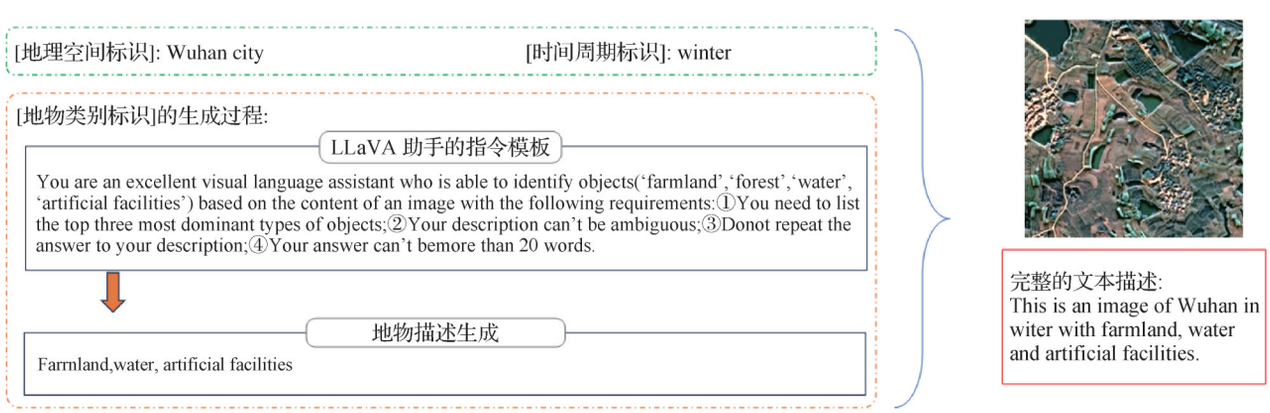
文本类型	地理时空标识	地物语义描述	语言文本示例
无文本	无	无	无
简单文本	有	无	This is an image of Wuhan in winter.
短文本	有	仅陈述遥感影像包含的主要地物类别有什么	This is an image of Wuhan in winter with farmland, water and artificial facilities.
长文本	有	不仅描述遥感影像中有什么地物,还需要描述这些地物类别的空间依赖关系	This is an image of Changsha in summer. Center: A large building complex with green roofs. Upper left: A large building complex. Upper right: A large farmland field. Lower left: A large building under construction. Lower right: A large body of water surrounded by farmland.

with the following requirements: ① At the {pos} of this image, describe the scene starting with: "{pos} of image:"; ② Your description can't be ambiguous; ③ Do not repeat the answer to your description; ④ Your answer can't be more than 20 words.", 其中{pos}表示方位。5个方位的文本描述组合构成长文本的[地物类别标识]。

2)短文本的"地物类别标识"只需要陈述遥感影像包含的主要地物类别有什么。给定地物类别的名称集合,比如('farmland', 'forest', 'water', 'artificial facilities')。对于任意遥感影像,使用LLaVA按照以下指令模板进行文本描述生成:"You are an excellent visual language assistant who is able to identify objects ('farmland', 'forest', 'water', 'artificial facilities') based on the content of an image with the following requirements: ① You must list the top three most dominant types of objects; ② Your answer can't be ambiguous."。



(a) 遥感影像长文本描述生成



(b) 遥感影像短文本描述生成

图2　遥感影像长短文本描述的生成示例

Fig. 2　Examples of generating long and short text descriptions for remote sensing images

((a) long text description generation for remote sensing images; (b) short text description generation for remote sensing images)

1.3　语言文本引导的全局模型预训练

根据代理任务的差异,全局模型预训练策略可以大致分为判别式与生成式两种:1)判别式预训练的一般思路是利用对比学习拉近包含相同语义信息的图像和文本之间的距离,同时推远包含不同语义信息的图像和文本之间的距离,从而学习图像特征与文本特征在实例级别的对齐关系。尽管长文本描述包含有丰富的细节信息,但是对比学习在图文实例特征对齐过程中难以捕捉这些文本细节与图像特征之间的关联,可能阻碍图像与文本之间对齐关系的建立。因此大多数判别式预训练使用信息更加明确的短文本(Yang等,2022)。2)生成式预训练的一

般思路是从掩码图像或加噪图像预测生成原始图像数据，从而学习图像中通用的视觉特征。语言文本可以作为条件引导，帮助模型更好地理解图像与文本之间的关联语义信息。特别是包含有丰富上下文信息的长文本描述，可能有助于模型对图像细节内容的理解，从而生成与长文本描述相匹配的图像细节特征。

1.3.1 语言文本引导的全局模型判别式预训练

全局模型判别式对比学习预训练（简称对比预训练）的基本思路是通过数据增广构建正负样本对，利用对比学习拉近正样本对之间的距离，同时推远负样本对之间的距离，从而学习图像的一般通用特征。在语言文本引导的全局模型判别式预训练中，通常是利用对比学习拉近匹配的图像和文本，推远不匹配的图像和文本，从而实现图像特征与文本特征之间的对齐，比如CLIP。

给定遥感影像样本$\boldsymbol{x}$及其对应的文本描述c_T。用$\boldsymbol{f}$表示图像输入图像编码器得到的图像特征，$\boldsymbol{f}_T$表示文本输入经过文本编码器得到的文本特征。$\boldsymbol{f}_i$与$\boldsymbol{f}_T^i$表示源数据集中第i个图像—文本对。在预训练过程中，CLIP通过一个实例判别任务优化InfoNCE目标函数，具体为

$$L_{\text{InfoNCE}}=-\frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\boldsymbol{f}^i\cdot\boldsymbol{f}_T^i}{\tau_{\text{CLIP}}}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\boldsymbol{f}^i\cdot\boldsymbol{f}_T^j}{\tau_{\text{CLIP}}}\right)}+\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\boldsymbol{f}_T^i\cdot\boldsymbol{f}^i}{\tau_{\text{CLIP}}}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\boldsymbol{f}_T^i\cdot\boldsymbol{f}^j}{\tau_{\text{CLIP}}}\right)}\right) \tag{3}$$

式中，第1项表示最大化图像特征和正确文本特征之间的相似度，第2项表示最大化文本特征和正确图像特征之间的相似度，“·”代表向量点乘，N表示批量大小，τ是一个可学习参数。

1.3.2 语言文本引导的全局模型生成式预训练

1）语言文本引导的图像掩码生成式预训练。图像掩码生成式预训练（简称掩码预训练）的基本思想是随机遮蔽一部分图像块，然后预测与这些遮蔽块对应的原始像素，从而学习图像中的通用特征表示。在语言文本引导的图像掩码生成式预训练中，文本编码被作为额外的输入帮助预测被遮蔽部分的原始像素，从而学习图像与文本之间的语义关系。BEiT（Wang等，2022b）是首次使用视觉Transformer架构作为骨干网络的掩码图像模型，其首先将图像表示为一系列离散的视觉标记（visual tokens），然后随机遮蔽一部分图像块，通过图像重建损失来预测这些遮蔽图像块的视觉标记。

给定一个视觉Transformer作为图像编码器，给定遥感影像样本$\boldsymbol{x}$及其对应的长文本描述c_T（具体的文本描述生成细节见2.2）。图像样本可以被切分成N个图像块$\{\boldsymbol{x}_i^p\}_{i=1}^N$，并且利用离散变分自编码器将其转换为$N$个视觉标记$\{h_i\}_{i=1}^N$。如果随机遮蔽将近40%的图像块，掩码位置表示为$M\in\{1,\cdots,N\}^{0.4N}$。被遮蔽后的掩码图像样本$\boldsymbol{x}^M=\{\boldsymbol{x}_i^p:i\notin M\}_{i=1}^N\cup\{\boldsymbol{e}_{[M]}:i\in M\}_{i=1}^N$，其中$\boldsymbol{e}_{[M]}\in\mathbf{R}^d$表示一个可学习的嵌入表征，用于替换被遮蔽的图像块。随后将$\boldsymbol{x}^M$输入视觉Transformer，其预训练目标函数可以定义为最大化给定掩码图像时预测正确视觉标记h_i的似然概率，具体为

$$\max\sum_{\boldsymbol{x}\in\mathcal{D}_x}\mathbb{E}_M\left[\sum_{i\in M}\log p\left(h_i\left|\boldsymbol{x}^M\right.\right)\right] \tag{4}$$

给定一个预训练的文本编码器，将c_T输入文本编码器得到文本特征，随后文本特征与掩码图像的图像特征进行拼接一起作为Transformer的输入，此时预训练目标函数可以定义为最大化给定掩码图像—文本对时预测正确视觉标记h_i的似然概率，具体为

$$\max\sum_{\boldsymbol{x}\in\mathcal{D}_x}\mathbb{E}_M\left[\sum_{i\in M}\log p\left(h_i\left|\left(\boldsymbol{x}^M,c_T\right)\right.\right)\right] \tag{5}$$

2）语言文本引导的扩散去噪生成式预训练。扩散去噪生成式预训练（简称扩散预训练）的基本思想是对一个从高斯分布上采样的随机变量逐渐去噪来学习图像数据的概率分布。在语言文本引导的扩散去噪生成式预训练中，文本编码作为条件输入用于指导加噪图像中噪声的去除。这种方法使得模型能够理解并利用图像与文本之间的语义关系，从而生成与给定文本描述相匹配的图像特征。Stable Diffusion是一种典型的扩散模型，其包含一个已训练的文本编码器、一个已训练的自编码器和一个U-Net的去噪网络ϵ_θ。这个自编码器和去噪网络的组合可以看做是图像编码器。

给定遥感影像样本$\boldsymbol{x}$及其对应的文本描述c_T，

可以使用自编码器 ε 得到编码潜变量 $v_0 = \varepsilon(\boldsymbol{x})$，通过逐步地向这个编码潜变量添加噪声可以生成一个马尔可夫链的隐变量 $v_1, \cdots, v_T$。具体为

$$q\left(v_t \middle| v_{t-1}\right) = N\left(v_t; \sqrt{1-\beta_t}\, v_{t-1}, \beta_t \boldsymbol{I}\right) \tag{6}$$

式中，β_t 决定了在时间步 t 添加噪声的方差。当通过链式添加的总噪声足够大的时候，v_T 就近似于 $N(0, \boldsymbol{I})$。从 v_0 直接 v_t 的边缘结果可以表示为

$$q\left(v_t \middle| v_0\right) = N\left(v_t; \sqrt{\alpha_t}\, v_0, (1-\alpha_t)\boldsymbol{I}\right) \tag{7}$$

式中，$\alpha_t = \prod_{i=1}^{t} (1-\beta_i)$。换言之，$v_t = \sqrt{\alpha_t}\, v_0 + \sqrt{1-\alpha_t}\, \epsilon$，$\epsilon$ 表示标准高斯噪声。训练时，去噪网络 $\epsilon_\theta(v_t, t, c_T)$ 以时间步 t 的噪声潜变量 v_t 和文本描述 c_T 的编码特征作为输入，并且使用平方误差损失预测添加的噪声 ϵ，具体为

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, \varepsilon(\boldsymbol{x}), c_T, \epsilon \sim N(0,I)}\left[\left\| \epsilon_\theta(v_t, t, c_T) - \epsilon \right\|_2^2\right] \tag{8}$$

1.4 语言文本引导的局部模型微调

全局模型训练时通过文本描述的引导学习遥感影像视觉特征与文本描述中语义信息的关联。在局部模型微调过程中，利用类似的包含时空信息的文本描述帮助局部模型快速激活全局模型中与之关联的领域知识。在 1.2 小节中，已详细介绍了遥感影像文本描述生成的方法。本文使用的文本描述通过大型语言和视觉助手 LLaVA 根据提示指令自动生成，其中直接生成的长文本包含有丰富上下文信息，但是可能存在严重认知错误，还需要经过人工检查和纠正。相比之下，生成短文本发生认知错误的概率较低，但是信息会有所缺失。为平衡模型性能和部署成本，本文在全局模型训练时使用信息丰富但需要人工检查纠正的长文本描述，在局部模型微调时使用信息简单但无需人工干预的短文本描述。在算法 1 中展示了语言文本引导的“全局—局部”域自适应学习过程。

算法 1　语言文本引导的“全局—局部”域自适应学习过程

1）语言文本引导的全局模型预训练

输入：多源域数据集 $\mathcal{D}_s$、全局模型 M_{global}、LLaVA 助手

输出：全局模型 M_{global}

（1）使用 LLaVA 助手根据 1.2 节方法对 $\mathcal{D}_s$ 生成长文本描述 c_{Text}

（2）对 c_{Text} 进行人工检查并纠正具有严重认知错误的描述

（3）for $m \leftarrow 1$ to N_s do

（4）从 $\{\mathcal{D}_s, c_{\text{Text}}\}$ 中进行采样

（5）根据式（1）更新 M_{global}

（6）end for

2）语言文本引导的局部模型微调

输入：目标域数据集 $\mathcal{D}_t^{\#}$、局部模型 M_{local}、LLaVA 助手

输出：局部模型 M_{local}

（1）使用 LLaVA 助手根据 1.2 节方法对 $\mathcal{D}_t^{\#}$ 生成短文本描述 $c_{\text{Text}}^{\#}$

（2）for $m \leftarrow 1$ to $N_t^{\#}$ do

（3）从 $\{\mathcal{D}_t^{\#}, c_{\text{Text}}^{\#}\}$ 中进行采样

（4）根据式（2）更新 M_{local}

（5）end for

2 实验结果与分析

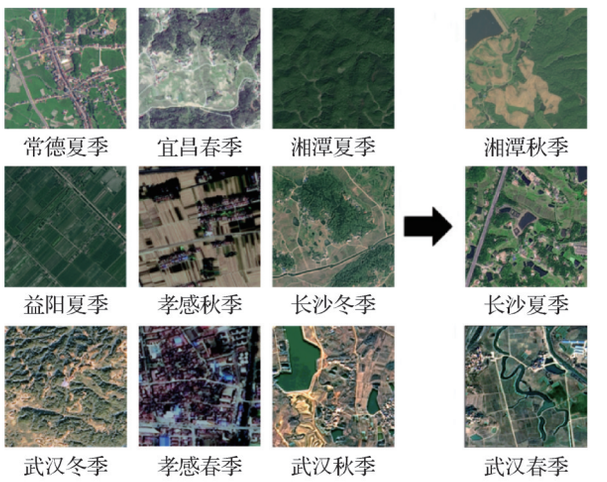
2.1 实验设置

2.1.1 数据集

数据集样本如图 3 所示。预训练多源域数据集收集自湖南、湖北多个地区的四季 RGB 遥感影像，共覆盖湘潭、长沙、益阳、常德、武汉、宜昌、孝感 7 个城市，包括 79 648 幅尺寸为 512 × 512 像素的无标注影像，原始影像数据均来自高分二号（Gaofen-2）卫星，并将全色影像（分辨率 0.8 m）和多光谱影像（分辨率 3.2 m）经过融合和下采样处理后得到的数据，处理后影像空间分辨率为 2 m。

目标域数据集包括湘潭数据集（秋季）、长沙数据集（夏季）和武汉数据集（春季）3 个城市的数据集，其中，湘潭数据集包含 1 536 幅 512 × 512 像素尺寸的 RGB 遥感影像，长沙数据集包含 768 幅同样尺寸的影像，武汉数据集包含 5 352 幅同样尺寸的影像。目标域数据集标注了 4 种地表覆盖类别：耕地、林草、水体和人工地表。3 个不同目标域数据集的类别分布如图 4 所示，可以看出不同领域的类别分布不一定一致。其中，湘潭数据集和长沙数据集的类别分布相似，而武汉数据集与前两者具有显著差异。

本节使用预训练多源域数据集训练全局模型（记做全局），并设置了 3 组“全局—局部”跨时空领域自适应语义分割任务，分别为全局→局部（湘潭）、



(a) 预训练多源域数据集　(b) 目标域数据集

图3　数据集样本示例

Fig. 3　Dataset samples

((a) multi-source dataset; (b) target dataset)

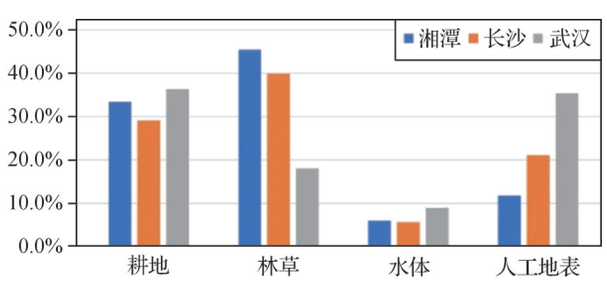


图4　目标域数据集类别分布

Fig. 4　Category distribution of the target dataset

全局→局部(长沙)、全局→局部(武汉)。每组任务中,局部模型微调时从对应目标域数据集中随机抽取包含50个样本的子集进行微调训练,表2提供了目标域数据集的详细描述。

表2　目标域数据集描述

Table 2　Description of the target dataset

目标域数据集	传感器	季节分布	空间分辨率/m	样本数量
湘潭数据集	GaoFen-2	秋季	2	1 536
长沙数据集	GaoFen-2	夏季	2	768
武汉数据集	GaoFen-2	春季	2	5 352

2.1.2　实现细节

在3种不同的全局模型预训练策略下,本文探究了语言文本引导对遥感图像跨时空领域自适应语义分割任务的影响。这3种预训练策略分别为对比预训练、掩码预训练和扩散预训练:1)对比预训练方法采用ResNet-50 (He等,2016)作为图像编码器,CLIP(Radford等,2021)作为文本编码器,其中无文本引导实验使用CMID(contrastive mask image distillation)(Muhtar等,2023)方法,有文本引导实验使用denseCLIP(Rao等,2022)框架,该框架将文本编码器生成的文本嵌入向量与图像编码器产生的特征图通过对比学习进行对齐,并在推理时通过像素—文本对齐匹配进行语义分割;2)掩码预训练方法使用BEIT3(Wang等,2022b)作为backbone,该模型通过多模态Transformer层实现文本—影像融合,其中文本嵌入作为额外的token加入到视觉token序列中,训练时对token序列进行随机掩码,并训练模型恢复完整token序列,最后连接单组投影层和上采样层组成的语义分割头,将融合后的特征转换为最终的分割结果;3)扩散预训练方法使用stable diffusion-1.5(Rombach等,2022)架构,该架构在条件输入部分引入文本特征,利用交叉注意力机制,让图像特征和文本特征相互作用,允许文本信息在整个扩散去噪过程中持续影响图像特征,本文实验中将模型的最后一层输出通道调整为对应语义分割类别数,之后连接一个上采样层以恢复原始分辨率分割图。

每种预训练策略都设置了有文本引导和无文本引导的对照实验,在实验结果中通过文本引导标记来说明预训练时使用的文本类型,如"无文本"标记的实验中预训练不使用文本引导,"长文本"标记的实验中预训练使用长文本引导,"短文本"标记的实验中预训练使用短文本引导。而在局部模型微调阶段中,无文本引导的实验不使用文本引导,有文本引导的实验统一使用短文本进行引导。

所有模型训练过程中,优化器的选择和超参数的设置均采用各个方法原论文推荐的配置,所有实验均在配备1张NVIDIA GeForce RTX A6000的计算机上进行。在精度统计中,使用交并比(intersection over union, IoU)评估每个类别的预测精度,并通过平均F1分数(mean F1 score, mF1)、Kappa系数、频率加权交并比(frequency weighted intersection over union, FWIoU)和总体准确率(overall accuracy, OA)等指标对模型的整体性能进行综合评估。

2.2　多源遥感图像跨时空领域自适应实验结果

本小节在对比判别式、掩码生成式和扩散生成式3种不同预训练策略上设置了3组"全局—局部"跨时空领域自适应语义分割实验来验证所提出的框

架的有效性，每一组实验中包括无文本引导和有文本引导对照实验，其中文本引导实验的标记含义为预训练使用的文本类型，如“长文本”代表预训练使用长文本。

在表3所示的全局→局部（湘潭）定量实验结果上，可以观察到两个有趣的现象。

1）从整体评价指标来看，有文本引导的方法对比无文本引导的方法在3种不同的学习策略上均表现出明显的性能提升。以OA为例，在对比预训练方法中，使用文本引导后精度从61.5%提升至71.8%；在掩码预训练方法中，使用文本引导后精度从72.0%提升至72.9%；在扩散预训练方法中，使用文本引导后精度从78.4%提升至80.6%。其他3个整体指标也可以观察到相同的现象，这表明语言文本在减少跨时空遥感影像领域偏移中起到了积极作用。此外，在对比预训练方法中，与短文本引导相比，长文本引导的对比预训练方法在所有整体指标上都表现出性能下降。也就是说，基于对比学习的模型训练不适合使用长文本引导。一个可能的原因是长文本虽然包含丰富的细节信息，但对比学习在实例特征对齐中很难捕捉到这些细节与图像特征之间的关联。相比于信息明确的短文本，基于对比学习的模型训练中使用长文本引导大概率会对图像与文本之间对齐关系的建立造成干扰。从图5定性结果来看，相比于无文本引导方法，有文本引导的方法在整体上减少了错误分类。

2）从类别评价指标来看，有文本引导的图像掩码和扩散去噪策略在各类别均表现出性能提升，但

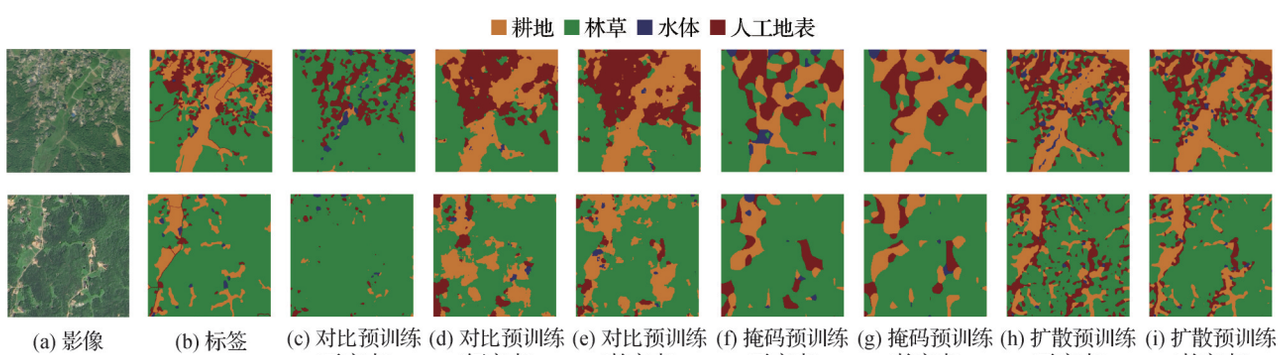
表3　全局→局部（湘潭）的定量实验结果

Table 3　Quantitative experimental results for global→local (Xiangtan, XT)

/%

预训练策略	文本引导	IoU				mF1	Kappa	FWIoU	OA
		耕地	林草	水体	人工地表				
对比预训练	无	17.3	59.4	**25.7**	27.3	47.0	27.4	40.6	61.5
	短文本	**57.1**	**67.1**	23.4	27.6	**58.5**	**54.6**	**58.3**	**71.8**
	长文本	56.2	64.7	21.3	**27.7**	57.2	52.9	56.6	70.6
掩码预训练	无	54.5	**70.0**	18.6	26.2	56.4	53.9	58.6	72.0
	长文本	**56.9**	**70.0**	**19.9**	**26.7**	**57.5**	**55.4**	**59.5**	**72.9**
扩散预训练	无	68.9	74.9	39.6	30.9	67.8	65.5	67.6	78.4
	长文本	**70.0**	**77.4**	**42.5**	**33.2**	**69.8**	**68.3**	**69.6**	**80.6**

注：加粗字体表示各项指标在每种预训练策略上的最优结果。



(a) 影像　(b) 标签　(c) 对比预训练（无文本）　(d) 对比预训练（短文本）　(e) 对比预训练（长文本）　(f) 掩码预训练（无文本）　(g) 掩码预训练（长文本）　(h) 扩散预训练（无文本）　(i) 扩散预训练（长文本）

图5　全局→局部（湘潭）的定性实验结果图

Fig. 5　Qualitative experimental results for global→local (XT) ((a) images; (b) labels; (c) contrastive pre-training (no text); (d) contrastive pre-training (short text); (e) contrastive pretraining (long text); (f) masked pre-training (no text); (g) masked pre-training (long text); (h) diffusion pre-training (no text); (i) diffusion pre-training (long text))

有文本引导的判别式对比学习策略仅在部分类别上表现出性能提升。如图5所示湘潭数据集中的主导类别是耕地和林草，人工地表和水体占比相对较少。相比于无文本引导的对比预训练方法，加入文本引导后在整体上的性能提升主要来源于主导类别的精度提升，在人工地表上表现相近，但是在水体上有明显下降。一个可能的原因是判别式对比学习聚焦于实例级别的特征对齐，其优化过程更加倾向于学习数据集中主导类别的特征对齐，容易忽略稀少类别的特征对齐。

全局→局部（长沙）和全局→局部（武汉）的实验结果分别见表4和表5，可以看到在3种不同的全局模型预训练策略中，有文本引导的结果均优于无文本引导的结果。全局→局部（长沙）实验结果中，以OA指标为例，在对比预训练方法中，使用文本引导后精度从62.7%提升至71.4%；在掩码预训练方法中，使用文本引导后精度从62.2%提升至66.6%；在扩散预训练方法中，使用文本引导后精度从71.9%提升至74.8%。全局→局部（武汉）实验结果中，以OA指标为例，在对比预训练方法中，使用文本引导后精度从64.3%提升至67.5%；在掩码预训练方法中，使用文本引导后精度从68.1%提升至69.6%；在扩散预训练方法中，使用文本引导后精度从69.0%提升至70.1%。从图6和图7的定性结果来看，有

表4　全局→局部（长沙）的定量实验结果

Table 4　Quantitative experimental results for global→local (Changsha, CS)

/%

预训练策略	文本引导	IoU				mF1	Kappa	FWIoU	OA
		耕地	林草	水体	人工地表				
对比预训练	无	36.9	54.6	23.9	43.6	55.9	44.8	44.3	62.7
	短文本	**55.6**	62.6	**39.4**	**50.6**	**68.0**	**59.1**	**55.7**	**71.4**
	长文本	54.4	**62.7**	33.3	47.2	65.4	57.0	54.0	70.4
掩码预训练	无	36.0	54.5	18.2	**47.4**	54.7	45.4	44.3	62.2
	长文本	**44.5**	**58.9**	**36.2**	47.0	**63.2**	**51.9**	**49.8**	**66.6**
扩散预训练	无	56.7	65.4	29.5	50.0	65.9	59.5	56.1	71.9
	长文本	**61.7**	**67.5**	**46.0**	**50.9**	**71.8**	**63.9**	**60.0**	**74.8**

注：加粗字体表示各项指标在每种预训练策略上的最优结果。

表5　全局→局部（武汉）的定量实验结果

Table 5　Quantitative experimental results for global→local (Wuhan, WH)

/%

预训练策略	文本引导	IoU				mF1	Kappa	FWIoU	OA
		耕地	林草	水体	人工地表				
对比预训练	无	51.4	**33.3**	39.3	49.7	60.2	47.0	46.9	64.3
	短文本	**53.9**	30.0	**51.3**	**52.4**	**63.2**	**52.6**	**50.7**	**67.5**
	长文本	50.9	23.9	48.9	51.8	60.0	49.5	48.4	65.6
掩码预训练	无	54.1	34.0	47.4	54.9	64.1	53.0	51.4	68.1
	长文本	**55.0**	**41.0**	**50.4**	**56.0**	**67.0**	**55.4**	**53.3**	**69.6**
扩散预训练	无	**55.5**	30.9	51.0	55.1	64.3	54.4	52.3	69.0
	长文本	55.3	**42.2**	**51.7**	**56.3**	**67.7**	**56.1**	**53.8**	**70.1**

注：加粗字体表示各项指标在每种预训练策略上的最优结果。

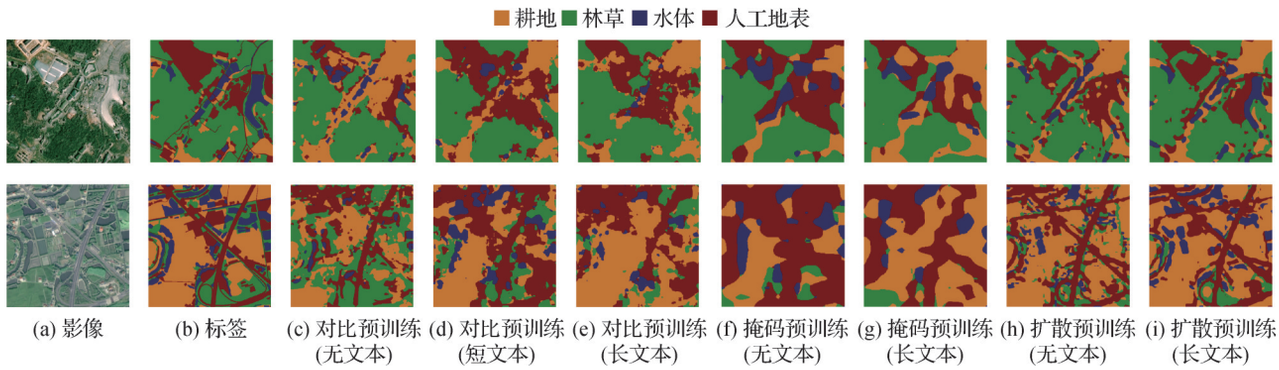


图 6　全局→局部(长沙)的定性实验结果图

Fig. 6　Qualitative experimental results for global→local (CS) ((a) images; (b) labels; (c) contrastive pre-training (no text); (d) contrastive pre-training (short text); (e) contrastive pretraining (long text); (f) masked pre-training (no text); (g) masked pre-training (long text); (h) diffusion pre-training (no text); (i) diffusion pre-training (long text))

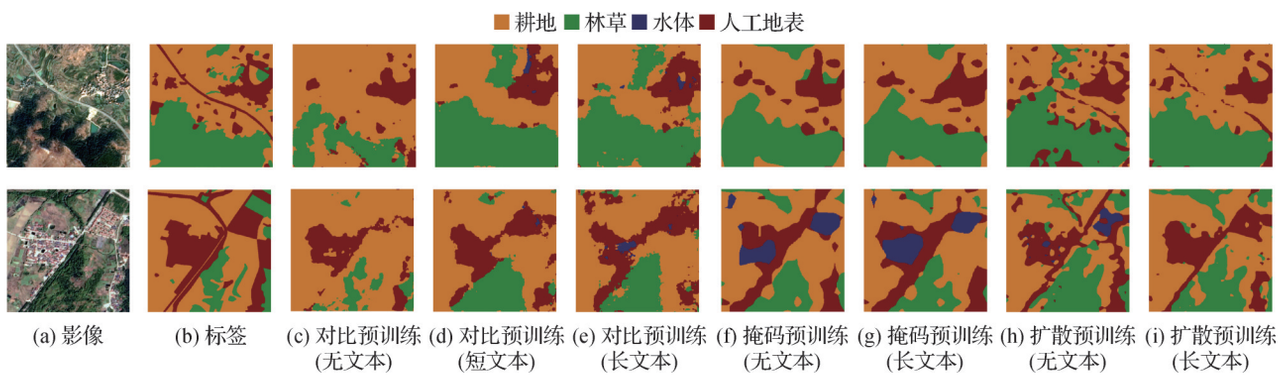


图 7　全局→局部(武汉)的定性实验结果图

Fig. 7　Qualitative experimental results for global→local (WH) ((a) images; (b) labels; (c) contrastive pre-training (no text); (d) contrastive pre-training (short text); (e) contrastive pretraining (long text); (f) masked pre-training (no text); (g) masked pre-training (long text); (h) diffusion pre-training (no text); (i) diffusion pre-training (long text))

文本引导的结果相比于无文本引导的结果，类别错分的现象有明显改善。这些实验结果同样都证明了语言文本在减少跨时空遥感影像领域偏移中起到了积极作用。

2.3　性能分析

本节设计了“局部—局部”与“全局—局部”学习范式的对比实验、不同微调标注样本量的对比实验和不同文本类型微调的对比实验，以进一步验证本文提出框架的有效性，分析微调标注量和微调时不同文本类型的影响。

2.3.1　“局部—局部”与“全局—局部”领域自适应学习范式的对比实验分析

本小节设置了两组局部到局部领域自适应语义分割任务对比实验来验证所提出框架的有效性，分别为长沙↔武汉和长沙↔湘潭。长沙、湘潭和武汉数据集之间都存在特征分布和类别分布的差异，从图 4 所示的类别分布可以得出：长沙和湘潭的主导类别一致，都是耕地、林草；而长沙和武汉的主导类别不一致，武汉的主导类别是耕地、人工地表。因此长沙↔武汉的领域相关性比长沙↔湘潭更弱。

对比方法包括 Src(local)-FT、CBST(class-balanced self-training)(Zou 等，2018)、UDA-NAS(unsupervised domain adaptation neural architecture search)方法(Broni-Bediako，2024)和本文方法 4 种，其中，前 3 种方法都是局部—局部方法。Src(local)-FT 表示在单源域标注数据集上训练的源模型在 50 个目标标注样本上微调的策略。局部—局部方法使用的训练数据包括单源域的标注数据、50 个目标标注样本和所有目标无标注样本。实验结果中迁移方向记做“源域→目标域”，例如，使用长沙数据集作为源域、武汉数据集作为目标域，记做长沙→武汉。

实验结果表明，“局部—局部”的域自适应策略

在跨时空环境下无法确保满足相关性假设，而“全局—局部”的学习范式在更广泛的范围内保证了相关性假设的适用性，从而在长沙↔武汉和长沙↔湘潭实验中都获得了较好的性能提升。

1）在领域之间相关性较弱的情况下，局部—局部方法具有较高的负迁移风险。如表6所示，从长沙↔武汉实验结果可以观察到局部—局部的方法产生了明显的负迁移。以OA指标为例，相比于Src(local)-FT，CBST在长沙→武汉实验中下降了6.2%，在武汉→长沙实验中下降了9.1%。相比于Src(local)-FT，UDA-NAS的精度下降更为剧烈，在长沙→武汉实验中下降了26.9%，在武汉→长沙实验中下降了47.5%。这是因为局部—局部方法都是基于强相关假设进行算法设计，如果源域和目标域之间的差异较大，这类方法就难以保证有效迁移。而本文提出的语言文本引导的全局—局部方法在长沙→武汉和武汉→长沙上性能均领先于Src(local)-FT策略。

2）在领域之间相关性较强的情况下，局部—局部方法在性能提升上要略高于全局—局部方法。如表7所示，从长沙↔湘潭实验结果中可以观察到，以OA指标为例，相比于Src(local)-FT，UDA-NAS在长沙→湘潭和湘潭→长沙实验中分别提升了5.6%和

表6　长沙↔武汉的定量实验结果

Table 6　Quantitative experimental results for CS↔WH

/%

迁移方向	方法	IoU				mF1	Kappa	FWIoU	OA
		耕地	林地	水体	人造地表				
长沙→武汉	Src(local)-FT	51.2	28.9	35.9	45.0	56.9	43.6	44.2	62.4
	CBST	36.5	2.0	41.9	46.4	45.0	37.3	37.9	56.2
	UDA-NAS	41.0	3.6	0.6	20.8	25.2	9.0	23.0	35.5
全局→局部（武汉）	本文	**55.3**	**42.2**	**51.7**	**56.3**	**67.7**	**56.1**	**53.8**	**70.1**
武汉→长沙	Src(local)-FT	57.4	66.3	5.4	47.2	56.8	56.8	54.0	70.0
	CBST	43.1	63.1	7.9	26.4	48.5	42.8	43.5	60.9
	UDA-NAS	23.9	3.4	0.2	15.3	18.0	−2.9	12.3	22.5
全局→局部（长沙）	本文	**61.7**	**67.5**	**46.0**	**50.9**	**71.8**	**63.9**	**60.0**	**74.8**

注：加粗字体表示各项指标在每种局部域上的最优结果。

表7　长沙↔湘潭的定量实验结果

Table 7　Quantitative experimental results for CS↔XT

/%

迁移方向	方法	IoU				mF1	Kappa	FWIoU	OA
		耕地	林地	水体	人造地表				
长沙→湘潭	Src(local)-FT	62.8	69.6	32.4	32.3	64.2	60.3	62.4	75.3
	CBST	62.1	74.8	33.1	22.6	62.2	61.2	64.2	77.7
	UDA-NAS	**70.3**	77.2	**51.2**	**35.4**	**72.4**	**69.2**	**70.3**	**80.9**
全局→局部（湘潭）	本文	70.0	**77.4**	42.5	33.2	69.8	68.3	69.6	80.6
湘潭→长沙	Src(local)-FT	50.5	60.9	51.5	50.2	69.4	57.8	54.4	70.9
	CBST	53.6	65.3	44.0	53.8	70.0	61.2	57.2	72.3
	UDA-NAS	**62.1**	**70.7**	**55.2**	**56.3**	**75.7**	**67.9**	**63.5**	**77.3**
全局→局部（长沙）	本文	61.7	67.5	46.0	50.9	71.8	63.9	60.0	74.8

注：加粗字体表示各项指标在每种局部域上的最优结果。

6.4%，CBST 在长沙→湘潭和湘潭→长沙实验中分别提升了 2.4% 和 1.4%，本文方法分别提升了 5.3% 和 3.9%，精度提升略低于局部—局部方法 UDA-NAS，但是要优于经典方法 CBST。

2.3.2　不同微调标注样本量对算法性能的影响分析

本小节设计并进行了 3 组对比实验，以探究微调标注样本量对本文提出算法有效性的影响。这 3 组实验分别采用了不同数量的微调标注样本进行模型训练，其中微调标注样本量分别为 10 个、50 个和 100 个。实验结果如图 8 所示，结果表明：扩散预训练策略在不同微调标注样本量下，文本引导始终对模型具有积极的作用。

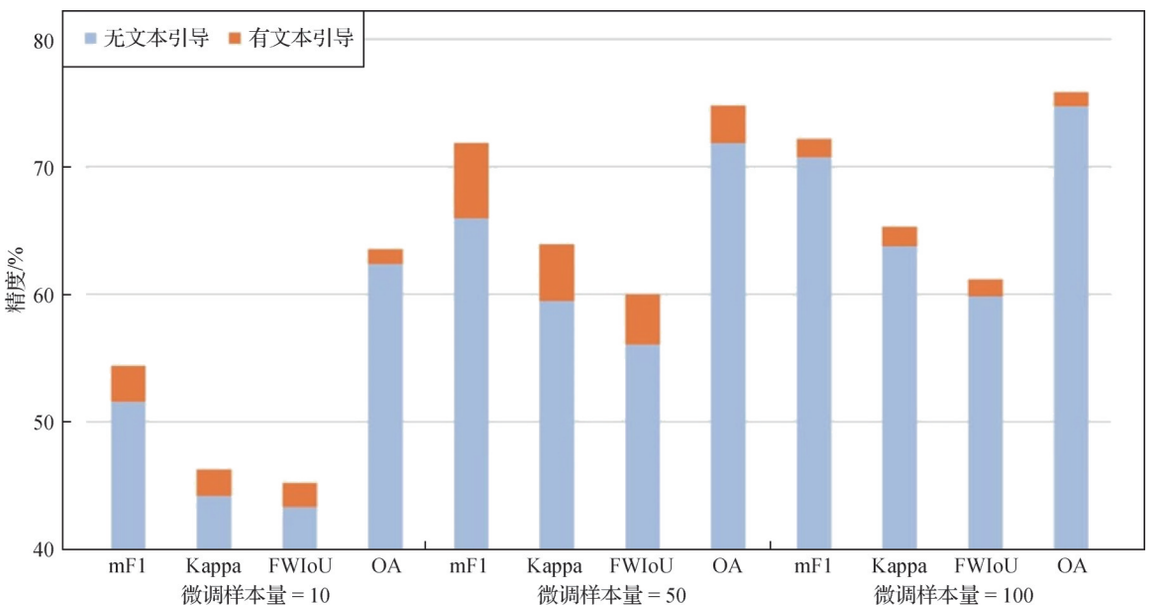


图 8　扩散预训练在不同微调样本量下的定量结果

Fig. 8　Quantitative results of diffusion pre-training with varying fine-tuning sample sizes

2.3.3　微调时使用不同文本类型的消融实验

本小节分析了局部模型微调时使用短文本和简单文本以及不使用文本的消融实验。其中，短文本是包含地理空间标识、时间周期标识和地物类别标识的短文本，简单文本是只包括地理空间标识和时间周期标识的短文本。在全局→局部（长沙）上的实验结果（见表 8）表明，微调时有文本引导的结果均优于无文本引导的结果，其中使用短文本进行微调的效果更好。结果表明，本文所提出语言文本引导对提升模型跨时空语义分割效果具有积极效果。

表 8　微调文本类型消融实验结果

Table 8　Ablation experimental results of fine-tuning on text types

/%

微调时使用的文本类型	mF1	Kappa	FWIoU	OA
无文本	65.9	59.5	56.1	71.9
短文本	**71.8**	**63.9**	**60.0**	**74.8**
简单文本	68.3	62.6	58.8	73.8

注：加粗字体表示各列最优结果。

3　结　论

本文探索并验证了语言文本在减少跨时空遥感影像领域偏移中的积极作用，提出一种语言文本引导的“全局模型预训练—局部模型微调”的领域自适应学习框架。该框架使用遥感影像的文本描述，不仅促进了全局模型预训练时捕捉遥感影像中地物的时空分布规律，还增强了局部模型微调时相关领域知识的激活。通过 3 个多源跨时空语义分割实验表明，与传统“局部—局部”的域自适应方法和无文本引导的“全局—局部”域自适应方法相比，本文提出的框架显著提升了模型迁移性能。后续研究将从以下两个方面考虑：1）探究在更大的时空范围下语言文本对模型迁移性能提升的影响。本文仅在湖南、

湖北地区4个季节采样的遥感数据上进行探索性实验，后续研究考虑拓展至更广的空间范围和更长的时间跨度，以评估所提框架在全国乃至全球范围应用的可行性。2）研究更精细化的遥感影像文本描述方法，考虑根据气象数据（温度、降水量等）和地形数据来丰富描述内容。

参考文献（References）

Broni-Bediako C, Xia J S and Yokoya N. 2024. Unsupervised domain adaptation architecture search with self-training for land cover mapping//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR). Seattle, USA: IEEE: 543-553 [DOI: 10.1109/CVPRW63382.2024.00059]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: 1597-1607 [DOI: 10.5555/3524938.3525087]

Dong R S, Ma Y Q, Liu Y and Li F Y. 2022. CRNet: class relation network for crop remote sensing image semantic segmentation. Journal of Image and Graphics, 27(11): 3382-3394 (董荣胜, 马雨琪, 刘意, 李凤英. 2022. 加强类别关系的农作物遥感图像语义分割. 中国图象图形学报, 27(11): 3382-3394) [DOI: 10.11834/jig.210760]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680 [DOI: 10.5555/2969033.2969125]

Gou C H, Felemban A, Khan F F, Zhu D Y, Cai J F, Rezatofighi H and Elhoseiny M. 2024. How well can vision language models see image details? [EB/OL]. [2024-08-08]. https://arxiv.org/pdf/2408.03940.pdf

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: #574 [DOI: 10.5555/3495724.3496298]

Li D R, Zhang L P and Xia G S. 2014. Automatic analysis and mining of remote sensing big data. Acta Geodaetica et Cartographica Sinica, 43(12): 1211-1216 (李德仁, 张良培, 夏桂松. 2014. 遥感大数据自动分析与数据挖掘. 测绘学报, 43(12): 1211-1216) [DOI: 10.13485/j.cnki.11-2089.2014.0187]

Li L J, He Y, Xie G, Zhang H X and Bai Y H. 2024. Cross-layer detail perception and group attention-guided semantic segmentation network for remote sensing images. Journal of Image and Graphics, 29(5): 1277-1290 (李林娟, 贺赟, 谢刚, 张浩雪, 柏艳红. 2024. 跨层细节感知和分组注意力引导的遥感图像语义分割. 中国图象图形学报, 29(5): 1277-1290) [DOI: 10.11834/jig.230653]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1516 [DOI: 10.5555/3666122.3667638]

Liu Y K, Wang K K, Li M S, Huang Y W and Yang G P. 2024. Exploring the cross-temporal interaction: feature exchange and enhancement for remote sensing change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 11761-11776 [DOI: 10.1109/JSTARS.2024.3413715]

Long M S, Cao Y, Wang J M and Jordan M I. 2015. Learning transferable features with deep adaptation networks//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 97-105 [DOI: 10.5555/3045118.3045130]

Long M S, Zhu H, Wang J M and Jordan M I. 2017. Deep transfer learning with joint adaptation networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 2208-2217 [DOI: 10.5555/3305890.3305909]

Muhtar D, Zhang X L, Xiao P F, Li Z S and Gu F. 2023. CMID: a unified self-supervised learning framework for remote sensing image understanding. IEEE Transactions on Geoscience and Remote Sensing, 61: #5607817 [DOI: 10.1109/TGRS.2023.3268232]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. ACM: PMLR: 8748-8763

Rao Y M, Zhao W L, Chen G Y, Tang Y S, Zhu Z, Huang G, Zhou J and Lu J W. 2022. DenseCLIP: language-guided dense prediction with context-aware prompting//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 18061-18070 [DOI: 10.1109/CVPR52688.2022.01755]

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 10674-

10685［DOI：10.1109/CVPR52688.2022.01042］

Song J M，Meng C L and Ermon S. 2022. Denoising diffusion implicit models［EB/OL］.［2022-10-05］. https://arxiv.org/pdf/2010.02502.pdf

Tahmoresnezhad J and Hashemi S. 2017. Visual domain adaptation via transfer feature learning. Knowledge and Information Systems，50(2)：585-605［DOI：10.1007/s10115-016-0944-x］

Tao C，Qi J，Guo M N，Zhu Q and Li H F. 2023a. Self-supervised remote sensing feature learning：learning paradigms，challenges，and future works. IEEE Transactions on Geoscience and Remote Sensing，61：#5610426［DOI：10.1109/TGRS.2023.3276853］

Tao C，Qi J，Zhang G，Zhu Q，Lu W P and Li H F. 2023b. TOV：the original vision model for optical remote sensing image understanding via self-supervised learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing，16：4916-4930［DOI：10.1109/JSTARS.2023.3271312］

Tao C，Yin Z W，Zhu Q and Li H F. 2021. Remote sensing image intelligent interpretation：from supervised learning to self-supervised learning. Acta Geodaetica et Cartographica Sinica，50(8)：1122-1134（陶超，阴紫薇，朱庆，李海峰. 2021. 遥感影像智能解译：从监督学习到自监督学习. 测绘学报，50(8)：1122-1134）［DOI：10.11947/j.AGCS.2021.20210089］

Tuia D，Persello C and Bruzzone L. 2016. Domain adaptation for the classification of remote sensing data：an overview of recent advances. IEEE Geoscience and Remote Sensing Magazine，4(2)：41-57［DOI：10.1109/MGRS.2016.2548504］

Vu T H，Jain H，Bucher M，Cord M and Pérez P. 2019. ADVENT：adversarial entropy minimization for domain adaptation in semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Long Beach，USA：IEEE：2512-2521［DOI：10.1109/CVPR.2019.00262］

Wang H，Guo M N，Li S X，Li H F and Tao C. 2024. Global-local coupled style transfer for semantic segmentation of bitemporal remote sensing images. IEEE Transactions on Geoscience and Remote Sensing，62：#4410615［DOI：10.1109/TGRS. 2024. 3425672］

Wang H，Tao C，Qi J，Xiao R and Li H F. 2022a. Avoiding negative transfer for semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing，60：#4413215［DOI：10.1109/TGRS.2022.3201688］

Wang M and Deng W H. 2018. Deep visual domain adaptation：a survey. Neurocomputing，312：135-153［DOI：10.1016/j. neucom. 2018.05.083］

Wang W H，Bao H B，Dong L，Bjorck J，Peng Z L，Liu Q，Aggarwal K，Mohammed O K，Singhal S，Som S and Wei F R. 2022b. Image as a foreign language：BEiT pretraining for all vision and vision-language tasks［EB/OL］.［2022-08-23］. https://arxiv.org/pdf/2208.10442.pdf

Xiao N and Zhang L. 2021. Dynamic weighted learning for unsupervised domain adaptation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Nashville，USA：IEEE：15237-15246［DOI：10.1109/CVPR46437.2021.01499］

Yang J W，Li C Y，Zhang P C，Xiao B，Liu C，Yuan L and Gao J F. 2022. Unified contrastive learning in image-text-label space//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. New Orleans，USA：IEEE：19141-19151［DOI：10.1109/CVPR52688.2022.01857］

Yang L，Yu Z C，Meng C L，Xu M K，Ermon S and Cui B. 2024. Mastering text-to-image diffusion：recaptioning，planning，and generating with multimodal LLMs［EB/OL］.［2024-01-26］. https://arxiv.org/pdf/2401.11708.pdf

Zhang H Y，Wang T B，Li M Z，Zhao Z，Pu S L and Wu F. 2022. Comprehensive review of visual-language-oriented multimodal pre-training methods. Journal of Image and Graphics，27(9)：2652-2682（张浩宇，王天保，李孟择，赵洲，浦世亮，吴飞. 2022. 视觉语言多模态预训练综述. 中国图象图形学报，27(9)：2652-2682）［DOI：10.11834/jig.220173］

Zhang T G，Gao F，Dong J Y and Du Q. 2022. Remote sensing image translation via style-based recalibration module and improved style discriminator. IEEE Geoscience and Remote Sensing Letters，19：#8009805［DOI：10.1109/LGRS.2021.3068558］

Zhao D，Wang S，Zang Q，Quan D，Ye X T，Yang R and Jiao L C. 2023. Learning pseudo-relations for cross-domain semantic segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE：19134-19146［DOI：10.1109/ICCV51070.2023.01758］

Zhao W，Yamada W，Li T X，Digman M and Runge T. 2021. Augmenting crop detection for precision agriculture with deep visual transfer learning—a case study of bale detection. Remote Sensing，13(1)：#23［DOI：10.3390/rs13010023］

Zheng X T，Xiao X L，Chen X M，Lu W X，Liu X Y and Lu X Q. 2024. Advancements in cross-domain remote sensing scene interpretation. Journal of Image and Graphics，29(6)：1730-1746（郑向涛，肖欣林，陈秀妹，卢宛萱，刘小煜，卢孝强. 2024. 跨域遥感场景解译研究进展. 中国图象图形学报，29(6)：1730-1746）［DOI：10.11834/jig.240009.］

Zheng Z S，Hu C Y and Jiang X Y. 2020. Deep transfer adaptation network based on improved maximum mean discrepancy algorithm. Journal of Computer Applications，40(11)：3107-3112（郑宗生，胡晨雨，姜晓轶. 2020. 基于改进的最大均值差异算法的深度迁移适配网络. 计算机应用，40(11)：3107-3112）［DOI：10.11772/j.issn.1001-9081.2020020263］

Zhu J Y，Park T，Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision（ICCV）. Venice，Italy：IEEE：2242-2251［DOI：10.1109/ICCV.

2017.244]

Zou Y, Yu Z D, Kumar B V K V and Wang J S. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 297-313 [DOI: 10.1007/978-3-030-01219-9_18]

Zou Y, Yu Z D, Liu X F, Kumar B V K V and Wang J S. 2019. Confidence regularized self-training//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 5981-5990 [DOI: 10.1109/ICCV.2019.00608]

作者简介

陶超,男,教授,主要研究方向为遥感数据视觉表征和多模态遥感视觉基础模型。E-mail:kingtaochao@csu.edu.cn

王昊,通信作者,男,讲师,主要研究方向为高分遥感影像智能解译。E-mail:haowang7cc@gmail.com

郭鑫,男,硕士研究生,主要研究方向为高分遥感影像智能解译。E-mail: gethin241@gmail.com

胡柯彦,男,硕士研究生,主要研究方向为多模态遥感影像语义分割。E-mail: phycheor@gmail.com

沈羽翔,男,硕士研究生,主要研究方向为多模态遥感影像目标检测。E-mail: shenyuxiang@csu.edu.cn