E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP18;TP75;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3278-15

论文引用格式:Zhang W, Chen T, Pei G S, Li B C, Wei M T and Yao Y Z. 2026. Semantic enhancement with vision foundation models for multitemporal multimodal remote sensing image matching. Journal of Image and Graphics, 31(9):3278-3292(张旺, 陈涛, 裴根生, 李宝晨, 韦明韬, 姚亚洲. 2026. 大模型语义增强的多时相多模态遥感影像匹配方法. 中国图象图形学报, 31(9):3278-3292)[DOI:10.11834/jig.260202]

大模型语义增强的多时相多模态遥感影像匹配方法

张旺[1],陈涛[1],裴根生[1],李宝晨[2],韦明韬[3],姚亚洲[1*]

1. 南京理工大学计算机科学与工程学院,南京　210094; 2. 陆军工程大学野战工程学院,南京　210007;
3. 中国航天科工集团第二研究院706所数据技术及应用事业部,北京　100854

摘　要:**目的**　针对多时相、多模态遥感影像在显著光谱差异、复杂季节变化及重复纹理条件下匹配稳定性不足的问题,构建统一评测基准并提升跨模态遥感影像匹配精度。**方法**　基于SeCo(seasonal contrast remote sensing)数据集构建了包含15 000对可见光与近红外跨时相样本的遥感影像匹配基准,其中训练集12 000对、测试集3 000对。通过组织多时相多模态影像对并引入随机单应变换,形成统一数据划分与评价协议;基于此,提出一种基于视觉大模型语义增强的跨模态特征匹配方法,以XoFTR(cross-modal feature matching Transformer)为基础匹配主干,在粗匹配阶段引入冻结的DINOv3(self-distillation with NO labels v3)提取高层稠密语义特征,并设计语义提升模块,实现语义先验与几何特征的协同建模,从而增强模型对跨时相、多模态稳定区域的感知与表征能力。**结果**　实验结果表明,所提方法在严格阈值MHA(mean homography accuracy)@3和中等阈值MHA@5下分别达到33.03%和37.03%,优于各对比方法;在较宽松阈值MHA@7下达到39.17%,略低于SP+SG(SuperPoint+SuperGlue)的42.34%,但明显优于原始XoFTR和直接域内微调模型;定性实验结果表明,本文方法能够有效抑制跨区域误匹配,并获得更密集、更稳定的正确对应关系。**结论**　视觉大模型提供的高层语义先验能够有效缓解多时相多模态遥感影像间的表征差异,为大模型赋能航空遥感高精度匹配提供了一种轻量、可复现的语义注入路径。本文公开代码地址:https://github.com/heng-shan/Dino_ft。

关键词:遥感多时相;多模态图像匹配;视觉基础模型;语义增强;XoFTR;DINO

Semantic enhancement with vision foundation models for multitemporal multimodal remote sensing image matching

Zhang Wang[1], Chen Tao[1], Pei Gensheng[1], Li Baochen[2], Wei Mingtao[3], Yao Yazhou[1*]

1. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing　210094, China;
2. Field Engineering College, Army Engineering University of PLA, Nanjing　210007, China; 3. Data Technology and Application Division, Institute 706, the Second Academy of China Aerospace Science & Industry Corp., Beijing　100854, China

Abstract: The robust matching of multitemporal and multimodal remote sensing images is a fundamental prerequisite for image registration, change detection (CD), target localization, image fusion, and three-dimensional reconstruction in aerial and satellite remote sensing applications. However, stable correspondence estimation remains highly challenging when the matching process is simultaneously affected by significant spectral discrepancy, complex seasonal variation, repetitive textures, weakly textured regions, and large appearance inconsistency across acquisition times. In visible-to-

收稿日期:2026-04-16;**修回日期**:2026-06-08;**预印本日期**:2026-06-15

*通信作者:姚亚洲　yazhou.yao@njust.edu.cn

基金项目:国家自然科学基金项目(62506169,62472222)

Supported by: National Natural Science Foundation of China(62506169,62472222)

near-infrared (NIR) remote sensing scenarios, the same geographic area may present markedly different reflectance patterns, local structures, and texture distributions under different seasons and spectral bands, making traditional handcrafted descriptors and many existing deep matching methods unreliable when they depend mainly on local photometric similarity or low-level geometric consistency. To address these issues, this paper studies the problem from the joint perspectives of benchmark construction and semantically enhanced matching design. With the goal of establishing reliable geometric supervision under a unified evaluation protocol, we constructed a standardized benchmark based on the public seasonal contrast (SeCo) dataset. This is achieved by organizing cross-temporal visible and near-IR image pairs and introducing random homography perturbations. The resulting benchmark contains 15 000 image pairs, including 12 000 training pairs and 3 000 testing pairs, and provides a reproducible experimental setting for quantitative comparison under severe temporal and spectral changes. The proposed protocol also evaluates models using homography reprojection error and multithreshold matching homography accuracy (MHA), thus enabling a unified assessment of geometric precision and matching robustness. Building upon this benchmark, a semantically enhanced cross-modal feature matching framework is proposed on the basis of XoFTR —— a hierarchical coarse-to-fine architecture for cross-modal correspondence estimation. The proposed method injects dense semantic priors extracted by a frozen DINOv3 model into the coarse matching stage of XoFTR through a lightweight Semantic Boosting Module, thereby enabling joint modeling of high-level semantic consistency and geometric correspondence. For each modality branch, the DINOv3 features are aligned to the coarse feature resolution and fused with geometric tokens, allowing the model to emphasize semantically stable regions while preserving local geometric sensitivity. This design is important because the coarse stage determines the candidate regions explored by subsequent fine matching; here, later local refinement is unlikely to recover correct matches once the global correspondence field is biased by unstable appearance cues. The overall pipeline follows a progressive coarse-to-fine paradigm in which multiscale geometric features are first extracted from visible and near-infrared inputs, coarse tokens are enhanced by DINOv3-derived semantic priors, global correspondences are established through transformer-based context interaction, and local fine-level refinement is subsequently performed to improve correspondence precision. Furthermore, training is conducted in a single-stage in-domain optimization manner in which the DINOv3 encoder is kept frozen throughout optimization and the XoFTR backbone is jointly learned together with the semantic boosting components. This strategy allows the method to leverage semantic priors from large vision models in a lightweight and reproducible manner without introducing excessive trainable parameters or unstable large-model adaptation behavior. Extensive experiments under the unified SeCo evaluation protocol demonstrate that the proposed method outperforms the XoFTR baselines across all MHA thresholds. On the constructed benchmark, the proposed model achieves 33.03%, 37.03%, and 39.17% in MHA@3, MHA@5, and MHA@7, respectively. These outperform the original pretrained XoFTR model by 12.91%, 7.68%, and 4.62% and surpass the direct in-domain fine-tuning baseline by 14.64%, 15.85%, and 16.61%, respectively. Qualitative comparisons further reveal that the proposed framework suppresses cross-region mismatches and yields denser correct correspondences in scenes with strong temporal appearance shifts and cross-modal ambiguity. These results indicate that simply adapting a geometric matching network to the target domain is insufficient for robust multitemporal multimodal remote sensing matching. In comparison, integrating high-level semantic priors from a vision foundation model can significantly enhance the stability of global correspondence establishment and improve final geometric estimation accuracy. The study reveals that high-level semantic representations learned by large-scale vision models can effectively alleviate the representation discrepancy caused by temporal evolution and spectral inconsistency. Furthermore, these can be incorporated into a hierarchical cross-modal matching framework in a computationally efficient and practically effective manner, thereby providing a lightweight and reproducible semantic injection route for large-model-empowered high-precision remote sensing image matching. The source code is publicly available at https://github.com/heng-shan/Dino_ft.

Key words: multi-temporal remote sensing; multi-modal image matching; vision foundation models; semantic enhancement; cross-modal feature matching Transformer (XoFTR); self-distillation with NO labels (DINO)

0 引 言

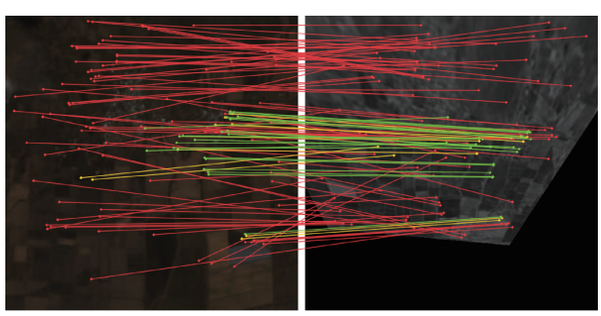
遥感影像匹配是影像配准(Ma等,2021)、变化检测(杜培军 等,2025)、目标定位(程红 等,2015;石争浩 等,2023)、影像融合(唐霖峰 等,2023)及三维重建(王文举 等,2025)等任务的重要基础,其匹配精度与鲁棒性直接影响遥感信息提取与定量分析的可靠性。与自然场景相比,遥感影像通常具有观测范围大、尺度变化复杂、局部重复纹理丰富及弱纹理区域占比较高等特点;当匹配任务进一步扩展到多时相、多模态场景时,问题将更加复杂。以可见光与近红外遥感影像为例,不同时相下地表覆盖、植被生长和阴影分布均可能发生明显变化,不同波段之间还存在显著的辐射差异,使同一地物在纹理、亮度和局部梯度层面表现出较大不一致性。因此,多时相多模态遥感影像匹配不仅面临几何变化带来的挑战,还受到复杂光谱差异与时相变化的共同影响(Jiang等,2021)。

围绕上述问题,已有研究从手工特征设计、深度局部特征学习以及上下文关联匹配等角度开展了大量探索。早期方法主要依赖SIFT(scale-invariant feature transform)(Wu 等,2013)、ORB(oriented FAST and rotated BRIEF)(Rublee等,2011)、AKAZE(accelerated-KAZE)(Tareen和Saleem,2018)等人工设计的局部描述子(朱柏和叶沅鑫,2024)。此类方法虽然具有实现简单、可解释性强等优点,但在跨模态辐射差异显著的场景中往往难以保持稳定的特征重复性。随着深度学习的发展,D2-Net(a trainable CNN for joint detection and description of local features)(Dusmanu 等,2019)、DeDoDe(detect, don't describe-describe, don't detect)(Edstedt等,2024)等方法通过学习式描述子提升了局部特征判别性;进一步地,LoFTR(local feature Transformer)(Sun 等,2021)、XoFTR(Tuzcuoğlu 等,2024)和MINIMA(modality invariant image matching)(Ren等,2025)等基于分层匹配范式的方法,能够在更大范围内建模跨图像上下文信息,为弱纹理和大位移场景下的匹配提供更强支撑。然而,针对多时相多模态遥感场景,现有方法存在两个突出问题:1)缺乏统一的大规模标准化实验基准,不同方法之间缺少一致的数据组织方式与评价协议;2)匹配框架大多以几何纹理一致性建模为主,对遥感场景中相对稳定的高层语义结构利用不足,在季节变化显著或跨谱段外观差异较大的区域容易出现匹配退化。

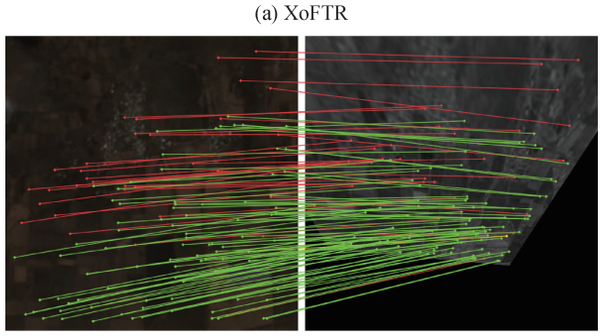
近年来,以DINO(self-distillation with NO labels)(Caron 等,2021)、SAM(segment anything model)(Ravi 等,2024)和CLIP(contrastive language-image pre-training)(Radford等,2021)等为代表的通用模型在大规模预训练支持下,具备较强零样本泛化能力和高层语义表征能力。相较于传统局部纹理特征,这类模型提取的密集语义特征更加关注场景结构、地物类别及区域关系,在面对多时相、多模态遥感影像时通常具有更好的语义稳定性。这为复杂遥感场景中的匹配建模提供了新思路:将视觉大模型的高层语义先验有效注入现有几何匹配框架,从而在保持几何约束能力的同时增强模型对跨时相、跨模态稳定区域的识别能力。相较于计算代价高且训练不稳定的全量微调,如何以轻量而有效的方式利用视觉大模型的语义先验,已成为当前大模型赋能航空遥感研究中的重要问题。

基于上述认识,本文围绕多时相多模态遥感影像匹配中基准缺失与语义利用不足两个关键问题展开研究。一方面,基于公开SeCo(seasonal contrast)(Mañas等,2021)数据集构建了可见光与近红外跨时相匹配任务标准化实验基准,并建立统一数据划分与评价协议;另一方面,提出了一种基于视觉大模型语义增强的跨模态特征匹配方法。该方法以XoFTR为匹配主干,在粗匹配阶段引入冻结的DINOv3(Siméoni等,2025)密集语义特征,并设计语义增强模块(semantic boosting module, SBM),实现高层语义先验与几何特征的协同建模。实验结果表明,所提方法在该基准上相较原始XoFTR模型和域内微调基线模型均取得更优的匹配效果和MHA(mean homography accuracy)指标。图1展示了SeCo基准中一组具有明显季节差异和可见光—近红外光谱差异的测试样本,并给出了XoFTR与本文方法在该样本上的匹配可视化结果,该样本存在明显光谱差异和季节外观变化,图中连线表示估计匹配关系,绿色表示正确匹配,红色表示错误匹配。LoFTR获得129对匹配,其中35对为5像素阈值内正确匹配;所提方法获得2 048对匹配,其中439对为5像素阈值内正确匹配。验证了视觉大模型高层语义先验对复杂遥感匹配任务的有效性。

本文主要贡献如下：1）构建了面向可见光—近红外跨时相匹配的 SeCo 遥感影像匹配基准，包含 15 000 对样本，并建立统一的随机单应评测协议。2）提出了一种基于视觉大模型语义增强的跨模态特征匹配方法，在 XoFTR 粗匹配阶段注入冻结 DINOv3 的密集语义先验，通过语义先验与几何特征融合，增强跨时相多模态场景下的全局对应关系建模能力。3）通过实验验证了大模型语义先验在多时相多模态遥感影像匹配中的有效性。结果表明，相比原始预训练模型和仅进行域内微调的基线模型，本文方法取得了更优的匹配精度与鲁棒性，为大模型赋能航空遥感高精度匹配提供了可行路径。



(a) XoFTR



(b) 语义增强匹配方法

图 1　XoFTR 与所提方法匹配结果对比

Fig. 1　Comparison of matching results between XoFTR and the proposed method ((a) XoFTR; (b) semantically enhanced matching method)

1　相关工作

1.1　跨模态影像的鲁棒特征表示

早期多模态影像匹配研究主要通过设计对辐射变化具有一定鲁棒性的人工特征来缓解模态差异，其核心思想是利用局部几何结构、梯度分布或相位信息构建无需大规模训练数据的稳定描述子。代表性方法包括 SIFT（Wu 等，2013）、ORB（Rublee 等，2011）、AKAZE（Tareen 和 Saleem，2018），以及面向非线性辐射差异设计的 RIFT（radiation change insensitive feature transform）（李焱磊 等，2023）和 HOPC（histogram of orientated phase congruency）（Ye 和 Shen，2016）等。SIFT 依靠尺度空间关键点检测和梯度方向直方图描述，在尺度与旋转变化下具有较好稳定性；ORB 结合 FAST 与 BRIEF（binary robust independent elementary features），具有计算效率高、资源开销低等优点；AKAZE 在非线性尺度空间中提取特征，对局部纹理和边缘结构具有较强刻画能力。相比之下，RIFT 和 HOPC 更关注多模态辐射差异适应能力，分别通过相位一致性及相关方向统计增强跨模态结构响应，在可见光、红外和 SAR（synthetic-aperture radar）等异源影像匹配中表现出一定优势。

手工描述子方法具有无需训练、可解释性强和算法流程稳定等优点，但其主要依赖边缘、角点和纹理等浅层局部先验。面对显著季节变化、地表覆盖变化和跨谱段成像差异时，此类方法难以建立稳定密集的对应关系，因此后续研究逐步转向深度特征学习与全局上下文建模方法。

随着卷积神经网络（convolutional neural network，CNN）的发展，研究者开始利用孪生网络或伪孪生网络直接从多源数据中学习特征映射关系（严镕宇 等，2023），以替代传统手工设计的局部描述子。相较于依赖人工先验的手工特征，深度特征具有更强的非线性表达能力，能够通过端到端训练挖掘模态间潜在的结构一致性，因此在复杂光照、视角变化和跨域匹配任务中表现出更强鲁棒性。

早期深度特征学习方法主要关注局部描述子的判别性增强，例如 MatchNet（unifying feature and metric learning for patch-based matching）（Han 等，2015）、DeepDesc（Mitra 等，2017）、TFeat（Balntas 等，2016）、L2-Net（Tian 等，2017）和 HardNet（Chao 等，2019）等，通常通过构造正负样本对并结合度量学习损失，在局部图像块上学习更稳定的特征嵌入。此后，SOS-Net（second order similarity regularization）（Tian 等，2019）、R2D2（repeatable and reliable detector and descriptor）（Revaud 等，2019）和 SuperPoint（DeTone 等，2018）等方法进一步将关键点检测与描述子学习纳入统一框架，从而提升了局部特征的重复性与判别性。在检测与描述联合优化方面，D2-Net（Dusmanu 等，2019）将特征检测与描述统一到共享

卷积特征图中，DeDoDe（Edstedt等，2024）则进一步联合建模关键点位置与描述特征，在局部匹配精度上表现出较强竞争力。

深度局部特征学习方法提升了遥感影像匹配的特征表达能力，但其仍主要依赖稀疏关键点或局部图像块，难以充分建模大范围场景的全局对应关系。因此，面向覆盖范围广、弱纹理区域多且模态差异显著的遥感影像匹配任务，基于深度上下文关联的分层匹配范式逐渐成为重要发展方向。

1.2 基于深度上下文关联的分层匹配范式

为了克服大尺度遥感影像中局部搜索范围受限、弱纹理区域难以建立稳定对应关系等问题，基于深度上下文关联的分层匹配范式（hierarchical matching paradigm）逐渐成为特征匹配研究的重要方向。与仅依赖局部关键点和局部描述子的匹配方法不同，这类方法通过在低分辨率特征层上建立全局上下文关联，先完成粗层对应关系估计，再在局部窗口内逐步细化匹配结果，从而兼顾大范围搜索能力与局部定位精度。

在这一方向中，SuperGlue（Sarlin等，2020）可视为从局部特征匹配迈向上下文关联匹配的代表性工作。该方法在给定局部关键点及描述子的基础上，引入图神经网络和最优传输机制，对两幅图像中的特征进行上下文关联和全局一致性约束，从而显著提升了匹配的鲁棒性。进一步地，LoFTR（Sun等，2021）首次摆脱了对显式关键点检测的依赖，采用无检测器的方式直接在密集特征层面建立粗匹配关系，并通过分层匹配框架在弱纹理和重复纹理区域取得了优于传统局部方法的性能。LoFTR的成功表明，在深度特征层面联合建模全局上下文信息，是解决复杂场景匹配问题的有效路径。

针对多源遥感影像匹配任务，XoFTR（Tuzcuoğlu等，2024）在LoFTR框架基础上进一步面向跨模态场景进行了适配，通过强化不同模态之间的特征交互能力，提升了模型对跨光谱差异的感知能力。与此同时，MINIMA（Ren等，2025）等工作尝试借助更强的预训练表示或跨模态适配机制，进一步增强模型在复杂场景中的泛化性能。尽管上述方法在跨模态匹配方面取得了显著进展，但其核心仍主要依赖于几何纹理特征的一致性建模。当面对多时相遥感影像中的季节变化、植被状态变化以及地表覆盖演化时，仅依赖几何结构的分层匹配框架仍然容易出现稳定性下降的问题。

因此，如何在分层匹配过程中进一步引入对时相变化更稳定的高层语义先验，成为当前提升多时相多模态遥感影像匹配鲁棒性的关键问题。基于这一认识，本文在XoFTR的粗匹配阶段引入视觉大模型提供的高层语义信息，通过语义增强的方式提升跨时相多模态场景下的粗层特征表达能力。

1.3 视觉大模型赋能的语义一致性建模

近年来，视觉基础模型（vision foundation model，VFM）的发展为复杂遥感场景下的影像匹配提供了新的研究思路。代表性模型如DINO（Siméoni等，2025）、CLIP（Radford等，2021）、SAM（Ravi等，2024）以及基于ViT/MAE（masked AutoEncoder）（Cong等，2022）架构的大规模预训练模型，通过大规模数据驱动学习获得了具有较强泛化能力的高层视觉表征。与传统局部纹理和几何特征不同，这类模型提取的密集语义特征更关注场景结构、地物类别及区域语义关系，因此在面对光谱差异、辐射变化和季节演化时，通常表现出更强的语义稳定性。对于多时相多模态遥感影像而言，这种高层语义一致性能够为跨模态匹配提供额外的先验约束，从而缓解仅依赖底层几何纹理所带来的不稳定问题。

然而，现有视觉大模型在遥感领域的应用主要集中于分类、检测、分割和变化分析等高层解译任务，如何将其语义不变性有效转化为匹配任务中的几何一致性约束仍有待深入探索。尤其是在基于由粗到细（coarse-to-fine，C2F）的分层匹配框架中，若能够在粗匹配阶段显式引入视觉大模型提供的高层语义先验，则有望增强模型对跨时相、跨模态稳定区域的感知能力。基于这一认识，本文将视觉大模型的密集语义特征引入XoFTR粗匹配阶段，以实现语义信息与几何特征的协同建模，从而提升多时相多模态遥感影像匹配的鲁棒性与精度。

2 匹配基准构建

针对现有多时相多模态遥感影像匹配研究中缺乏统一数据组织方式与评价协议的问题，本文基于公开遥感序列数据构建了一个面向可见光与近红外影像匹配的标准化实验基准。该基准旨在模拟复杂光谱差异与季节变化条件下的跨时相跨模态匹配场景，并为不同类型算法提供统一、可复现的训练与测

试环境。

2.1 数据来源与样本组织

本文以公开的SeCo（Mañas等，2021）数据集为原始数据源。SeCo基于Sentinel-2多光谱遥感影像构建，覆盖全球多个地理区域和不同季节时相，具有较丰富的跨区域、跨季节观测样本。依托其多光谱波段信息，本文进一步组织可见光与近红外模态的匹配样本：其中，可见光分支由B4、B3和B2波段合成为RGB影像，用于表征地物的视觉纹理与颜色结构；近红外分支采用B8波段构建单通道NIR影像，用于保留植被、水体及地表反射特性在近红外波段下的差异响应。本文采用固定随机种子42对样本对进行划分，训练集和测试集之间不存在完全相同的影像对重叠。训练集包含12 000对样本，测试集包含3 000对样本。当前划分属于样本对级随机划分，并非严格的地理区域无重叠划分；训练集和测试集在部分Sentinel-2场景编号上存在重叠，本文将该设置作为可控评测基准，用于比较不同方法在统一协议下的性能。跨时相样本的时间间隔范围为62～330天，其中训练集时间间隔中位数为160天，测试集时间间隔中位数为155天。

2.2 几何变换协议与样本生成

原始Sentinel-2影像经过正射校正与地理配准后，跨时相样本之间通常已具备较好的全局几何一致性。为了模拟实际遥感匹配任务中由于平台姿态扰动、成像距离变化及局部视角差异所引起的几何失配，本文在测试协议中对影像对施加已知二维单应变换（homography transformation）。旋转角度从$[-30^\circ, 30^\circ]$均匀采样，水平和垂直缩放因子从$[0.85, 1.06]$均匀采样，平移扰动服从标准差为24像素的高斯分布，剪切扰动服从标准差为0.12的高斯分布，透视扰动服从标准差为0.001 2的高斯分布。设参考影像与待匹配影像分别为$\boldsymbol{I}_0$和$\boldsymbol{I}_1$，本文对$\boldsymbol{I}_0$和$\boldsymbol{I}_1$分别施加随机单应矩阵$\boldsymbol{H} \in \mathbf{R}^{3\times3}$，得到变换后的影像$\boldsymbol{I}_0'$和$\boldsymbol{I}_1'$，并以$(\boldsymbol{I}_0', \boldsymbol{I}_1')$作为最终评测输入。由于在正射视角和局部平面假设下，遥感影像间由尺度、平移、旋转以及轻微透视变化引起的失配可以由单应模型较好近似，而仿射变换又是单应变换的特例，因此采用随机单应扰动能够较为合理地覆盖遥感匹配中的主要几何变化形式。本文所采用的随机变换包括旋转、平移、缩放、剪切及轻微透视扰动，从而为每一对样本构建具有严格几何真值的评价环境。

2.3 评价指标体系

本文采用单应矩阵精度（mean homography accuracy，MHA）作为核心评价指标。对于测试集中第n对样本，设真实单应矩阵为$\boldsymbol{H}_n$，由匹配结果估计得到的单应矩阵为$\hat{\boldsymbol{H}}_n$。给定一组参考采样点集合$\mathcal{P}$，其平均重投影误差定义为

$$e_n = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\left\|\hat{\boldsymbol{H}}_n p - \boldsymbol{H}_n p\right\|_2 \tag{1}$$

在此基础上，单应矩阵精度定义为

$$MHA@\tau = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}(e_n < \tau) \tag{2}$$

式中，N表示测试样本数，τ表示允许的像素误差阈值，$\mathbf{1}(\cdot)$为示性函数。本文报告MHA@3、MHA@5和MHA@7三项结果，以分别衡量严格、中等和相对宽松容差条件下的匹配成功率，如图2所示。

需要说明的是，本文未采用平均匹配数等指标作为主要评价依据。其原因在于，不同匹配方法在稀疏与稠密输出形式、置信度筛选策略以及后处理机制上存在显著差异，匹配点数量本身并不能直接反映几何估计的准确性，也难以在不同范式之间形成公平比较。相比之下，MHA直接度量预测匹配是否足以支持正确的全局几何恢复，更符合遥感影像配准任务的实际目标，也与XoFTR等分层匹配方法常用的单应估计评价方式保持一致。

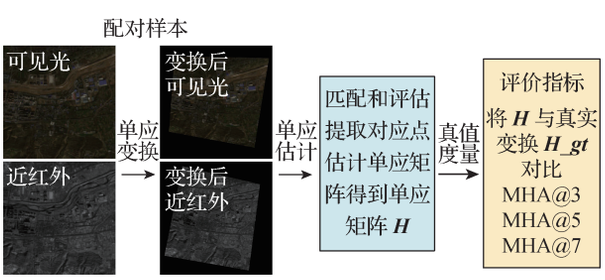


图2 评测协议

Fig. 2 Evaluation protocol

2.4 基准特点分析

与自然场景匹配基准相比，本文构建的多时相多模态遥感影像匹配基准具有以下特点。首先，不同波段之间存在显著的光谱响应差异，如图3所示，尤其是可见光（visible light，Vis）与近红外（near-infrared，NIR）影像在植被、水体和裸地等地物上的成像特性差异明显，对依赖底层纹理一致性的匹配方法提出了更高要求。其次，跨季节观测引入了明

显的时相动态性，地表覆盖状态、阴影分布和辐射特性均可能发生变化，从而使得传统几何纹理特征的稳定性受到挑战。最后，本文采用固定样本划分、统一几何扰动协议和统一评价指标，为不同类型的匹配模型提供了标准化的实验平台，从而保证了方法比较的可复现性与科学性。

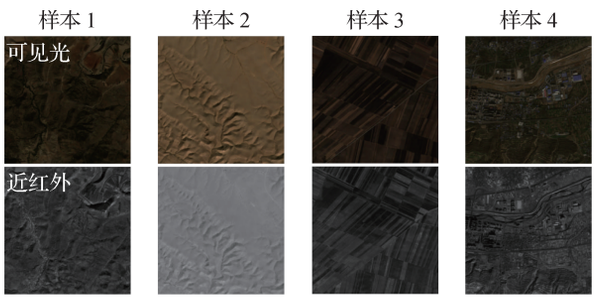


图3　多时相可见光/近红外基准样本

Fig. 3　Multi-temporal visible/near-infrared benchmark samples

3　大模型语义增强的跨模态匹配

针对多时相多模态遥感影像在光谱差异显著、季节变化复杂条件下匹配鲁棒性不足的问题，本文提出一种融合视觉大模型语义先验的遥感影像匹配方法。该方法利用视觉大模型所提供的高层语义表征，增强跨时相、跨模态条件下稳定地物结构的一致性建模能力，以弥补传统深度匹配网络主要依赖局部纹理与几何线索、因而在剧烈辐射变化场景中表征能力不足的局限。

3.1　总体框架概述

本文方法采用C2F的分层匹配框架，在粗匹配阶段完成跨模态影像之间的全局对应关系建模，并在细匹配阶段实现局部精细对齐。不同于传统深度匹配方法主要依赖局部几何纹理一致性，本文在粗匹配阶段引入视觉大模型提供的高层密集语义先验，以增强模型对多时相、多模态遥感场景中稳定场景结构与语义一致性的建模能力，从而提升复杂条件下的匹配鲁棒性与几何精度。所提方法的整体框架如图4所示。

给定一对待匹配的多时相多模态遥感影像 $\boldsymbol{I}_0$ 和 $\boldsymbol{I}_1$，利用XoFTR主干网络提取两幅影像的粗粒度几何特征 $\boldsymbol{F}_0$ 和 $\boldsymbol{F}_1$，用于建立初始的跨模态结构表示。与此同时，采用冻结的DINOv3预训练模型提取对应的高层密集语义特征 $\boldsymbol{S}_0$ 和 $\boldsymbol{S}_1$。在此基础上，本文设计语义增强模块，通过交叉注意力与门控融合机制，将高层语义先验注入粗匹配特征，生成语义增强后的特征表示 $\tilde{\boldsymbol{F}}_0$ 和 $\tilde{\boldsymbol{F}}_1$。随后，增强后的粗层特征进入粗匹配模块以建立全局对应关系，并在此基础上结合中层与细层局部特征完成逐级细化，最终获得亚像素级匹配结果。整体而言，本文方法实现了高层语义信息与底层几何线索的协同建模，为多时相多模态遥感影像匹配提供了一种更加稳健的技

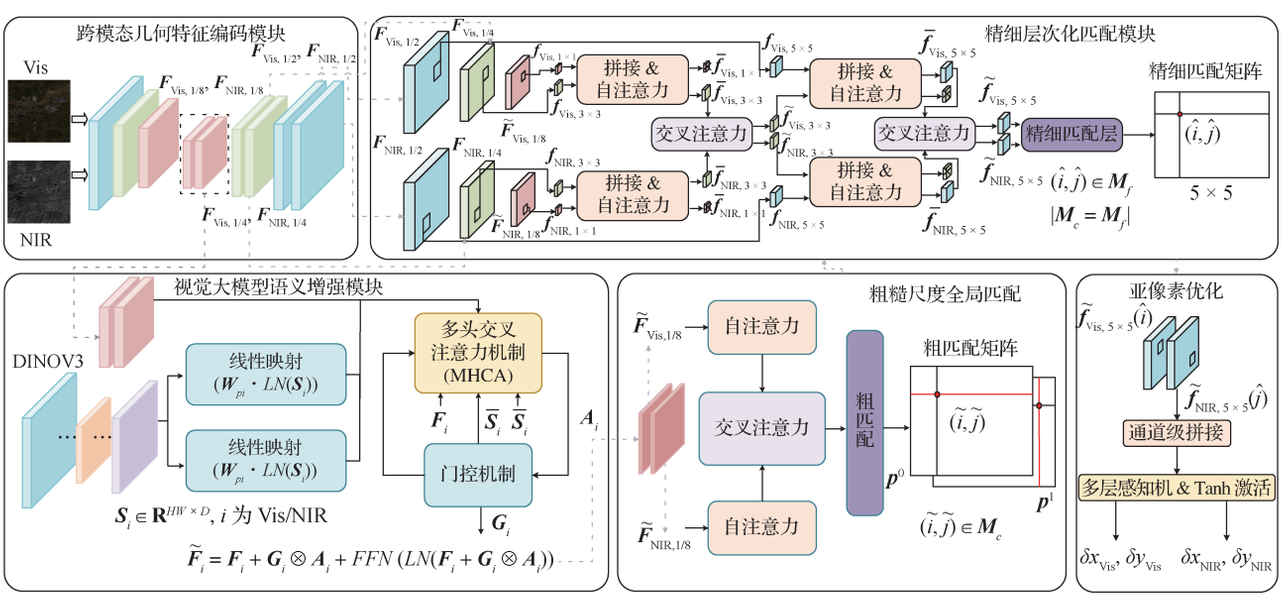


图4　所提基于视觉大模型语义增强的多时相多模态遥感影像匹配总体框架

Fig. 4　Overall framework of the proposed vision foundation model semantic-enhanced method for multi-temporal multi-modal remote sensing image matching

术路径。

3.2 跨模态几何特征编码

多时相多模态遥感影像匹配的关键在于构建对场景几何结构稳定且具有判别性的跨模态特征表示。针对遥感影像中尺度变化显著、局部重复纹理丰富、弱纹理区域较多及跨模态辐射响应差异明显等问题，本文采用XoFTR作为跨模态几何特征编码主干。不同于传统检测后再匹配的稀疏范式，XoFTR通过共享参数的卷积特征提取网络构建密集特征表示，并为后续基于Transformer的上下文关联建模提供统一基础，因此更适合复杂遥感场景下的跨模态匹配。

几何主干网络 $\mathcal{E}_g$ 首先对两幅影像进行多尺度特征编码，可表示为

$$\left(\boldsymbol{F}_i^c, \boldsymbol{F}_i^m, \boldsymbol{F}_i^f\right) = \mathcal{E}_g(\boldsymbol{I}_i), \qquad i \in \{0,1\} \tag{3}$$

式中，$\boldsymbol{F}_i^c \in \mathbf{R}^{H_c \times W_c \times C_c}$、$\boldsymbol{F}_i^m \in \mathbf{R}^{H_m \times W_m \times C_m}$ 和 $\boldsymbol{F}_i^f \in \mathbf{R}^{H_f \times W_f \times C_f}$ 分别表示粗层、中层和细层特征图。为建立后续的全局匹配关系，本文选取粗层特征作为主要匹配载体，并在加入二维位置编码（positional encoding，PE）后将其展平（*flatten*）为序列形式，具体为

$$\boldsymbol{X}_i = Flatten\left(PE\left(\boldsymbol{F}_i^c\right)\right) \in \mathbf{R}^{N_c \times C_c} \tag{4}$$

上述粗层特征在保留边缘轮廓、角点分布、局部纹理及空间布局等几何结构信息的同时，也为后续全局上下文建模与跨模态特征交互提供了统一表示基础。与此同时，中层与细层特征 $\boldsymbol{F}_i^m$ 和 $\boldsymbol{F}_i^f$ 则为后续局部窗口细化与亚像素级匹配提供多尺度局部信息支撑。

3.3 视觉大模型语义增强模块

为缓解多时相多模态遥感影像中几何结构相对稳定但外观差异显著所导致的粗匹配不稳定问题，本文在粗匹配阶段引入大模型高层语义先验。采用冻结的DINOv3作为语义编码器 $\mathcal{E}_s$，对输入影像分别提取密集语义特征，并将其空间分辨率对齐到粗层特征尺度，表示为

$$\boldsymbol{S}_i = \mathcal{E}_s(\boldsymbol{I}_i) \in \mathbf{R}^{N_c \times D_s}, \qquad i \in \{0,1\} \tag{5}$$

与几何主干提取的粗层特征 $\boldsymbol{X}_i$ 相比，$\boldsymbol{S}_i$ 更侧重于编码场景中的高层语义属性及稳定区域结构关系，因而能够在跨时相、跨模态条件下提供相对一致的语义参考。由于几何特征与语义特征在分布形式和通道维度上存在差异，本文首先对语义特征进行归一化LN操作 $LN(\cdot)$ 与线性映射 $\boldsymbol{W}_p$，将其投影到与粗层几何特征一致的维度空间，具体为

$$\bar{\boldsymbol{S}}_i = \boldsymbol{W}_p LN(\boldsymbol{S}_i) \in \mathbf{R}^{N_c \times C_c} \tag{6}$$

在完成维度对齐后，通过多头交叉注意力 $MHA(\cdot)$ 实现粗层几何特征与同模态语义特征之间的显式交互。具体地，以几何特征作为查询项，以投影后的语义特征作为键和值，可得

$$\boldsymbol{A}_i = MHA\left(LN(\boldsymbol{X}_i), \bar{\boldsymbol{S}}_i, \bar{\boldsymbol{S}}_i\right), \qquad i \in \{0,1\} \tag{7}$$

式中，$\boldsymbol{A}_i \in \mathbf{R}^{N_c \times C_c}$ 为语义引导下生成的增强响应。

该交互过程在每个模态分支内部独立进行，目的在于利用视觉大模型所提供的高层语义信息增强本分支粗层特征的表达能力，而非直接在两个模态之间进行语义对齐。

考虑到高层语义信息若过强注入，会削弱原始几何特征对局部结构细节的敏感性，本文进一步引入门控控制机制，对语义增强分量进行自适应调节。具体地，将归一化后的几何特征与交叉注意力输出在通道维进行拼接，并通过门控网络生成逐通道权重，具体为

$$\boldsymbol{G}_i = \sigma\left(MLP\left([\,LN(\boldsymbol{X}_i), \boldsymbol{A}_i\,]\right)\right) \tag{8}$$

式中，$[\cdot,\cdot]$ 表示特征拼接，$MLP(\cdot)$ 表示两层前馈感知机，$\sigma(\cdot)$ 表示sigmoid激活函数，$\boldsymbol{G}_i \in \mathbf{R}^{N_c \times C_c}$ 为门控权重。随后，本文采用残差式融合策略生成语义增强后的粗层表示，具体为

$$\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i + \boldsymbol{G}_i \otimes \boldsymbol{A}_i \tag{9}$$

式中，$\otimes$ 表示逐元素乘法。为进一步提升特征表达能力并稳定训练过程，在残差融合结果上进一步施加前馈变换（FFN），得到最终输出。具体为

$$\tilde{\boldsymbol{X}}_i = \tilde{\boldsymbol{X}}_i + FFN\left(LN(\tilde{\boldsymbol{X}}_i)\right) \tag{10}$$

经语义提升模块增强后的粗层特征 $\tilde{\boldsymbol{X}}_0$ 和 $\tilde{\boldsymbol{X}}_1$ 将被送入后续粗层特征交互模块 $\mathcal{T}(\cdot)$，以建立跨模态全局对应关系，即

$$\boldsymbol{Y}_0, \boldsymbol{Y}_1 = \mathcal{T}(\tilde{\boldsymbol{X}}_0, \tilde{\boldsymbol{X}}_1) \tag{11}$$

与在匹配结果层面进行后处理不同，本文在粗匹配构建阶段显式引入语义先验，使高层语义信息能够在全局对应关系建立之前参与特征表达，从而增强模型对多时相、多模态稳定区域的识别能力与匹配鲁棒性。如图5所示，在具有明显季节变化和跨模态外观差异的样例中，基线模型的响应往往较为分散，或者峰值落在错误区域，而引入DINO语义特征后，响应更集中且更接近真实对应点，说明语义增强能够显著提升粗匹配定位的稳定性与准确性。

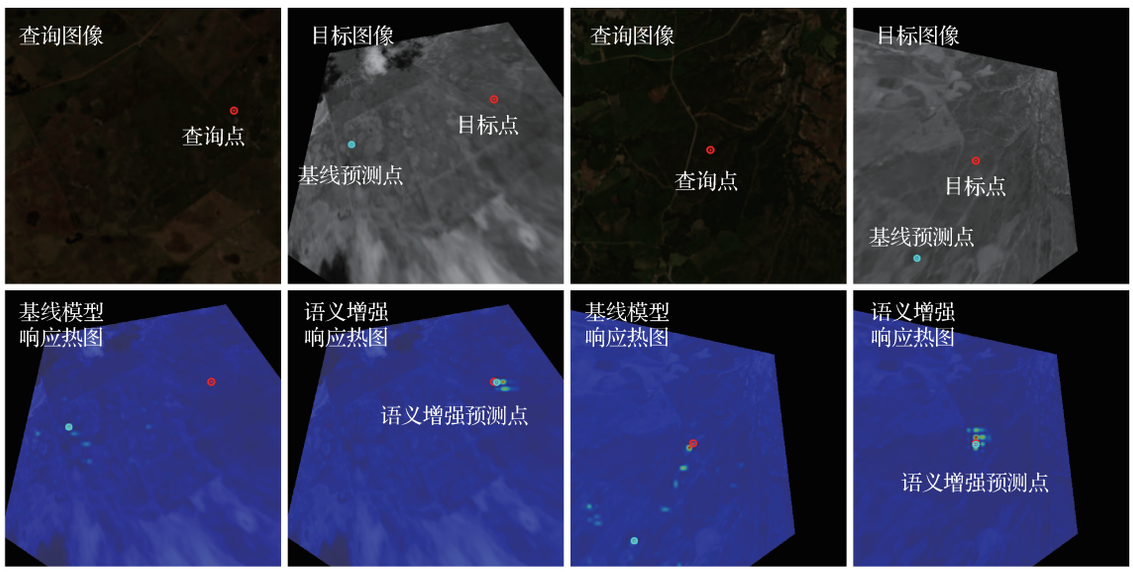


图5　XoFTR 与引入 DINO 语义提升后的 XoFTR 在粗匹配响应上的对比

Fig. 5　Comparison of coarse matching responses between the XoFTR and the DINO-enhanced XoFTR

3.4　粗尺度全局匹配

在获得语义增强后的粗层特征 $\tilde{\boldsymbol{X}}_0$ 和 $\tilde{\boldsymbol{X}}_1$ 后，本文采用基于自注意力和交叉注意力机制的粗层特征交互模块对两幅影像进行全局上下文化建模，以建立跨模态初始对应关系。由于粗层特征分辨率较低，该阶段能够在可接受的计算代价下实现较大范围的全局搜索，因此适合作为多时相多模态匹配中的初始对应关系建立阶段。

语义增强后的特征表示为 $\boldsymbol{Y}_0$、$\boldsymbol{Y}_1 \in \mathbf{R}^{N_c \times C_c}$，为降低密集特征交互带来的计算复杂度，该阶段采用线性注意力机制近似标准多头注意力，从而使特征交互复杂度相较于标准全连接注意力显著降低。这种设计使模型能够在粗层密集特征图上执行全局上下文关联，同时保持较高的推理效率，满足遥感影像匹配对全局感受野和计算可扩展性的双重需求。

在获得上下文化粗层特征后，本文进一步通过粗层匹配模块计算两幅影像间的特征相似性矩阵 $\boldsymbol{M}$，具体为

$$\boldsymbol{M} = \frac{1}{\tau_c}\left(\frac{\boldsymbol{W}_c\boldsymbol{Y}_0}{\sqrt{\boldsymbol{C}_c}}\right)\left(\frac{\boldsymbol{W}_c\boldsymbol{Y}_1}{\sqrt{\boldsymbol{C}_c}}\right)^{\mathrm{T}} \tag{12}$$

式中，$\boldsymbol{W}_c$ 表示粗层匹配投影矩阵，τ_c 表示温度系数。沿行列方向执行 softmax 归一化，得到双向匹配置信分布 $\boldsymbol{P}$。随后，通过置信度阈值约束与双向最近邻筛选生成粗层候选对应关系。阈值为 γ_c，粗层匹配集合可记为

$$\boldsymbol{M}_c = \{(u,v) \,|\, \boldsymbol{P}(u,v) > \gamma_c\} \tag{13}$$

进一步结合局部最大响应约束和边界剔除策略，获得最终的粗层匹配结果。

3.5　精细层次化匹配与亚像素优化

粗层全局匹配输出的对应关系 $\boldsymbol{M}_c$ 为后续高精度定位提供了候选匹配骨架，但由于粗层特征分辨率较低，其匹配结果仍只能反映近似的区域级对应关系，难以直接满足遥感影像精细配准对像素级定位精度的要求。为此，本文继承 XoFTR 的层次化细匹配思想，在粗层匹配结果约束下，进一步结合中层与细层局部特征进行逐级细化，从而实现从粗层结构一致性到细层局部几何一致性的渐进式匹配优化。该过程既利用了粗层匹配提供的全局搜索能力，又通过局部窗口内的上下文交互提升了细粒度对应关系的判别性。

粗层匹配集合为 $\boldsymbol{M}_c = \{(\boldsymbol{p}^c_{0,m}, \boldsymbol{p}^c_{1,m})\}^{M}_{m=1}$，其中 $\boldsymbol{p}^c_{0,m}$ 和 $\boldsymbol{p}^c_{1,m}$ 分别表示第 m 个粗匹配在两幅影像粗层特征图上的位置。为了为细匹配提供更加稳定的上下文先验，本文首先将粗层交互前后的特征进行融合，构造细化阶段的粗层上下文表示，具体为

$$\boldsymbol{C}_i = \psi\left([\,Reshape(\boldsymbol{Y}_i), \boldsymbol{F}^c_i\,]\right) \tag{14}$$

式中，$Reshape(\cdot)$ 表示将粗层序列特征恢复为二维特征图，$\psi(\cdot)$ 表示由卷积映射构成的融合变换。随后，以每个粗匹配位置为中心，分别从中层特征图和细层特征图中裁剪（Crop）局部窗口，记为

$$\boldsymbol{\omega}^m_{i,m} = Crop(\boldsymbol{F}^m_i, \boldsymbol{p}^c_{i,m}, \boldsymbol{W}_m) \in \mathbf{R}^{W^2_m \times C_m} \tag{15}$$

$$\boldsymbol{\omega}_{i,m}^{f} = Crop(\boldsymbol{F}_{i}^{f}, \boldsymbol{p}_{i,m}^{c}, \boldsymbol{W}_{f}) \in \mathbf{R}^{W_f^2 \times C_f} \tag{16}$$

式中，$\boldsymbol{W}_m$和$\boldsymbol{W}_f$分别表示中层和细层局部窗口尺寸。按照本文实现，$\boldsymbol{W}_m = 3$，$\boldsymbol{W}_f = 5$，因此分别对应3 × 3和5 × 5的局部特征窗口。

在局部细化阶段，本文首先利用粗层上下文特征对中层局部窗口进行增强。具体地，将中层局部窗口$\boldsymbol{\omega}_{i,m}^{m}$与对应位置的粗层上下文特征联合输入局部自注意力模块，从而使中层特征能够吸收粗层匹配结果所携带的全局上下文信息。随后，对两幅影像的增强中层窗口执行双向交叉注意力，建立局部跨图对应关系，得到跨模态交互后的中层表示，具体为

$$\begin{cases} (\tilde{\boldsymbol{\omega}}_{0,m}^{m}, \tilde{\boldsymbol{\omega}}_{1,m}^{m}) = \boldsymbol{A}_{m}^{\text{cross}} \boldsymbol{U} \\ \boldsymbol{U} = (\boldsymbol{A}_{m}^{\text{self}}(\boldsymbol{\omega}_{0,m}^{m}, \boldsymbol{C}_{0,m}), \boldsymbol{A}_{m}^{\text{self}}(\boldsymbol{\omega}_{1,m}^{m}, \boldsymbol{C}_{1,m})) \end{cases} \tag{17}$$

式中，$\boldsymbol{A}_{m}^{\text{self}}(\cdot)$和$\boldsymbol{A}_{m}^{\text{cross}}(\cdot)$分别表示中层局部自注意力与双向交叉注意力操作。增强后的中层特征进一步通过线性映射投影到细层特征维度，并作为细层局部窗口的上下文引导信息。随后，细层局部窗口在同样的自注意力结合交叉注意力机制下继续进行局部上下文化建模，得到最终用于精细匹配的窗口特征$(\tilde{\boldsymbol{\omega}}_{0,m}^{f}, \tilde{\boldsymbol{\omega}}_{1,m}^{f})$。通过上述设计，细层特征不仅保留了局部纹理和边缘细节，还融合了来自粗层和中层的结构上下文信息，从而显著增强了局部窗口内特征的判别能力。

在获得细层增强窗口后，本文对局部窗口计算细粒度相似性矩阵，具体为

$$\boldsymbol{M}_{m}^{f} = \frac{1}{\tau_f}\left(\boldsymbol{W}_f \tilde{\boldsymbol{\omega}}_{0,m}^{f}\right)\left(\boldsymbol{W}_f \tilde{\boldsymbol{\omega}}_{1,m}^{f}\right)^{\mathrm{T}} \tag{18}$$

式中，τ_f表示温度系数，$\boldsymbol{M}_{m}^{f} \in \mathbf{R}^{W_f^2 \times W_f^2}$为第$m$个局部窗口对的细匹配相似性矩阵。在此基础上，采用双向softmax构造细匹配置信矩阵$\boldsymbol{P}_{m}^{f}$并结合阈值筛选获得局部窗口内的离散匹配位置。由于本文默认采用一对一的细匹配策略，因此每个粗层匹配最终对应一个细层精化结果，即细层匹配数量与粗层匹配数量保持一致。

为了进一步突破离散窗口网格带来的定位误差，本文在细匹配结果基础上引入亚像素回归模块。具体地，将匹配位置对应的两侧细层特征进行拼接，并通过多层感知机预测局部偏移量，具体为

$$(\Delta \boldsymbol{p}_{0,m}, \Delta \boldsymbol{p}_{1,m}) = \rho\left([\tilde{\boldsymbol{w}}_{0,m}, \tilde{\boldsymbol{w}}_{1,m}]\right) \tag{19}$$

式中，$\rho(\cdot)$表示亚像素回归网络，$\tilde{\boldsymbol{w}}_{0,m}$和$\tilde{\boldsymbol{w}}_{1,m}$分别表示窗口内选中的匹配特征。最终，细匹配结果由粗层中心位置、窗口内离散偏移以及亚像素修正共同构成，从而得到最终的高精度对应点。通过这一层次化局部匹配与亚像素细化机制，本文方法能够在保留粗层全局搜索能力的同时，实现对跨模态、多时相遥感影像局部对应关系的精确刻画。

3.6 训练目标与优化策略

本文采用单阶段域内训练策略。在训练过程中，以预训练XoFTR权重作为模型初始化，并在SeCo多时相多模态遥感匹配数据集上直接进行端到端优化。对于本文引入的视觉大模型语义增强分支，DINOv3仅作为固定的高层语义先验提取器，其参数在训练过程中保持冻结；模型实际参与更新的参数主要包括几何匹配主干以及语义提升模块中的可学习参数。该设计能够避免对大模型进行联合微调所带来的训练不稳定性与额外计算开销，使优化过程更加聚焦于跨模态匹配表示与语义融合机制本身。

在监督目标方面，本文延续XoFTR的粗到细训练范式，对粗层匹配、细层匹配以及亚像素细化3个阶段进行联合约束。设总体损失函数记为

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_f \mathcal{L}_f + \lambda_s \mathcal{L}_s \tag{20}$$

式中，$\mathcal{L}_c$表示粗层匹配监督项，用于约束粗层对应关系的建立；$\mathcal{L}_f$表示细层局部匹配监督项，用于提升局部窗口内匹配位置的判别能力；$\mathcal{L}_s$表示亚像素细化监督项，用于进一步优化最终匹配点的定位精度；λ_c、λ_f和λ_s为相应的权重系数。本文采用$\lambda_c = 0.5$、$\lambda_f = 0.3$和$\lambda_s = 100$。该设置参考XoFTR原始训练配置并结合本文单应训练任务进行保持。由于亚像素回归项与粗层、细层匹配分类项具有不同的数值尺度，其原始损失值通常较小，因此设置较大的λ_s用于平衡不同监督项对总损失的贡献。通过上述多阶段联合优化，模型能够同时学习全局结构一致性、局部细粒度对应关系以及高精度几何定位能力。

在实现层面，输入影像首先经过统一尺寸预处理，随后分别送入几何主干与语义编码器提取对应特征。语义特征在空间尺度上对齐到粗层特征分辨率后，经语义提升模块注入粗匹配特征，并进入后续粗层Transformer完成全局上下文交互；细匹配阶段则在粗层匹配结果约束下，结合中层与细层局部窗口特征逐级完成匹配细化。训练过程中采用统一的数据划分、输入分辨率与优化设置；测试阶段对不同

对比模型使用相同的图像尺寸、匹配阈值和评价协议，以保证实验结果的可比性与公平性。

4 实验与结果分析

4.1 实验设置

本文实验基于SeCo多时相多模态遥感匹配数据集开展。为保证实验对比的公平性，所有可训练模型均采用相同的数据划分与评价协议，其中训练集包含12 000对影像，测试集包含3 000对影像，划分随机种子设为42。实验平台采用NVIDIA Tesla V100-SXM2-32GB GPU×1，训练与推理框架均基于PyTorch实现。

本文采用DINOv3 ViT模型作为语义编码器，其patch size为16，隐藏维度为1 024，包含24个Transformer层和16个注意力头，并使用最终层归一化后的patch tokens作为密集语义特征。DINOv3输入尺寸设为224 × 224像素。对于单通道NIR影像，本文将其复制为三通道后输入DINOv3，并采用预训练模型对应的均值[0.430, 0.411, 0.296]和标准差[0.213, 0.156, 0.143]进行归一化。DINOv3输出的patch特征首先恢复为空间特征图，然后通过双线性插值对齐到XoFTR粗层特征分辨率，最后展平为token序列输入语义提升模块。训练过程中DINOv3参数保持冻结。

在训练阶段，域内微调模型seco_ft与本文提出的语义增强模型dino_ft均以相同的XoFTR预训练权重作为初始化，在SeCo数据集上进行单阶段端到端优化。具体地，输入图像尺寸设为640 × 640像素，批大小为2，训练步数为30 000，优化器采用AdamW，初始学习率为2×10^{-4}，权重衰减系数为1×10^{-4}，匹配模式设为rgb_nir。其中，seco_ft表示在SeCo数据上直接进行域内微调的XoFTR模型；dino_ft在相同训练设置下进一步引入冻结的DINOv3 base模型作为高层语义先验提取器，并联合优化几何匹配主干与语义增强模块。作为对照，baseline直接采用原始XoFTR预训练权重进行测试，不进行SeCo域内训练。

在测试阶段，本文采用统一的SeCo评测协议对所有方法进行比较。具体而言，从固定测试集影像对中读取可见光与近红外图像，并对其中一幅影像施加随机单应变换以构造具有已知几何真值的测试样本。各方法测试输入分辨率统一设为640像素；对于XoFTR系列模型，粗匹配阈值设为0.3，细匹配阈值设为0.1，最大保留匹配点数设为2 048。由于部分测试对在随机单应扰动后不满足有效区域约束，即变换后影像的有效非零区域小于整幅图像面积的30%，故最终有效评测对数为2 903。在评价指标方面，本文以单应矩阵精度（MHA）作为核心指标，报告MHA@3、MHA@5和MHA@7，以综合评估各方法在几何估计精度与匹配鲁棒性方面的表现。

在对比方法设置上，本文重点分析baseline、seco_ft和dino_ft 3种XoFTR相关模型的性能差异，并进一步与传统方法及深度学习代表性匹配方法在统一协议下进行比较，以验证所提语义增强策略的有效性。

4.2 实验对比

为验证所提方法在多时相多模态遥感影像匹配任务中的有效性，本文在统一评测协议下，将所提方法与传统手工特征方法、深度局部特征方法以及代表性深度匹配方法进行比较。对比方法包括原始预训练模型baseline、域内微调模型seco_ft、本文提出的语义增强模型dino_ft，以及MINIMA、D2-Net、SuperGlue、SIFT、ORB和AKAZE等方法。表1给出了各方法在SeCo基准上的MHA指标结果。其中，

表1 不同类型匹配方法在SeCo多时相多模态遥感影像测试集上的匹配准确率对比

Table 1 Comparison of matching accuracy for different types of matching methods on the SeCo multi-temporal multi-modal remote sensing image test set

/%

类型	算法	MHA@3	MHA@5	MHA@7
手工描述子	SIFT	16.29	18.81	20.22
	ORB	9.37	12.75	14.88
	AKAZE	18.19	22.98	26.21
深度特征	D2-Net	0.69	6.23	13.74
	DeDoDe	<u>22.90</u>	27.64	30.47
	SP + MNN	16.57	27.49	34.58
分层匹配	seco_ft	18.39	21.18	22.56
	MINIMA	13.85	20.15	23.11
	SP + SG	22.53	<u>35.38</u>	**42.34**
	dino_ft(本文)	**33.03**	**37.03**	<u>39.17</u>

注：加粗、下划线字体分别表示各列最优、次优结果。

seco_ft表示域内微调模型，SG表示SuperGlue，SP表示SuperPoint，MNN表示互近邻匹配方法，直观匹配对比结果如图6所示。

由表1可见，本文提出的dino_ft在多数精度阈值下取得了最优或次优结果。其中，在严格阈值和中等阈值条件下，dino_ft的MHA@3和MHA@5分别达到33.03%和37.03%，均优于seco_ft以及其他对比方法，表明在粗匹配阶段引入视觉大模型高层语义先验后，模型能够更稳定地建立跨时相多模态影像之间的全局对应关系。在较宽松阈值条件下，dino_ft的MHA@7达到39.17%，仅次于SuperGlue的42.34%，依然表现出较强的整体竞争力。

从同系列模型对比结果来看，dino_ft相较于原始预训练模型baseline在3项MHA指标上均有明显提升，而仅进行SeCo域内微调的seco_ft反而低于baseline。这说明对于多时相多模态遥感场景，仅依赖几何特征进行域内适配难以有效克服光谱差异和时相变化带来的匹配退化问题，而高层语义先验的引入能够显著增强粗层特征的跨模态一致性表达能力。

从不同类型方法的比较结果来看，传统手工特征方法SIFT、ORB和AKAZE的整体表现均弱于dino_ft，说明仅依赖局部梯度或二值描述子的方式难以适应复杂的多时相多模态遥感场景。深度局部特征方法D2-Net和跨模态预训练方法MINIMA的结果同样低于本文方法，表明单纯依赖局部描述子或原始跨模态特征表示难以充分应对该任务中的大尺度时相变化与显著光谱差异。相比之下，本文方法

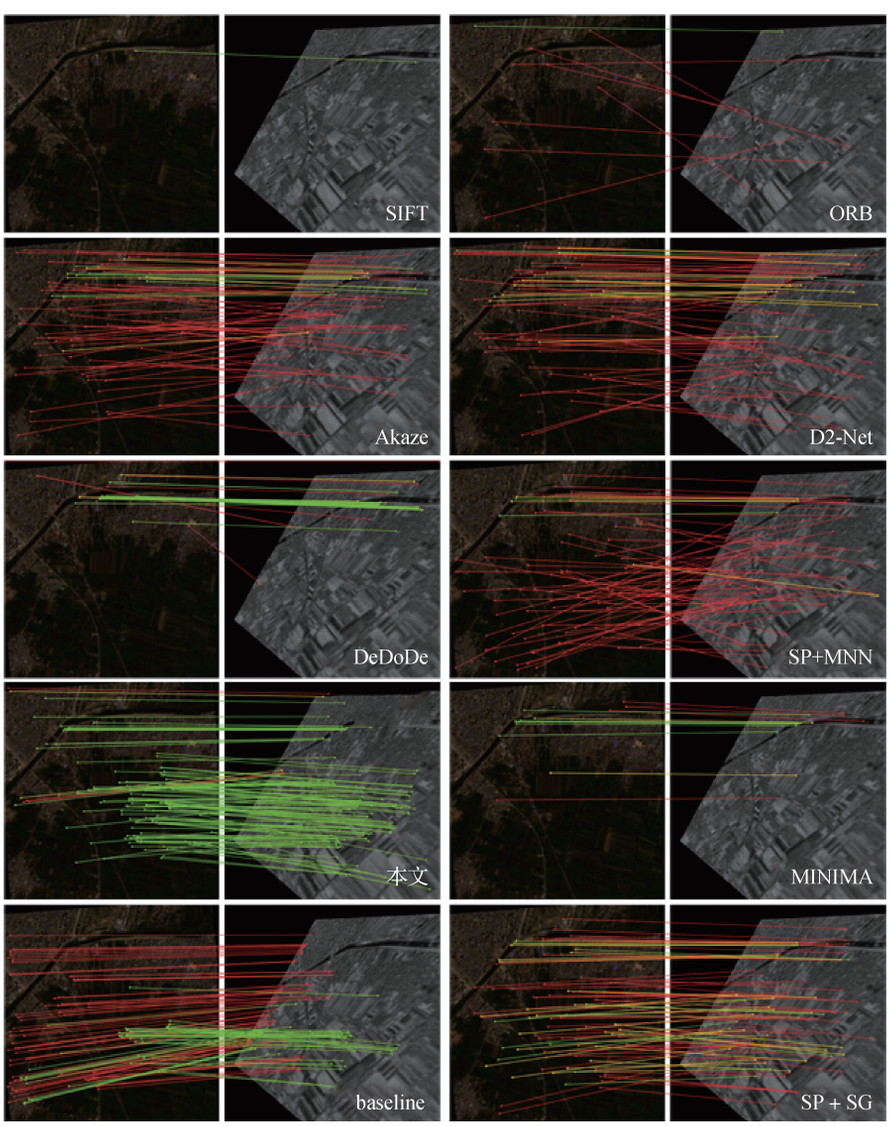


图6　不同方法在多时相多模态遥感影像匹配任务中的定性对比

Fig. 6　Qualitative comparison of different methods on the multi-temporal multi-modal remote sensing image matching task

通过将视觉大模型的高层语义信息显式注入XoFTR的粗匹配阶段，能够在保持分层匹配框架优势的同时进一步提升模型的匹配鲁棒性，从而验证了大模型语义增强策略在多时相多模态遥感影像匹配任务中的有效性。

4.3 语义增强模块消融分析

为验证本文所引入视觉大模型语义增强模块的有效性，进一步在XoFTR框架内部开展消融实验。具体而言，分别比较以下3种模型：1）baseline，即原始预训练XoFTR模型，未在SeCo数据上进行域内训练；2）seco_ft，即在SeCo数据集上直接进行域内微调的XoFTR模型；3）dino_ft，即本文提出的引入冻结DINOv3语义先验并进行联合训练的模型。通过该组实验，可以进一步分析单纯域内微调与语义增强策略对模型性能的实际影响。

表2给出了3种设置下的消融结果。可以看出，seco_ft在3项MHA指标上均低于baseline，说明仅依赖几何特征进行域内微调，难以有效适应多时相多模态遥感场景中的显著光谱差异和季节变化，甚至可能破坏原始预训练模型已有的几何匹配能力。相比之下，dino_ft在MHA@3、MHA@5和MHA@7上分别达到33.03%、37.03%和39.17%，均显著优于baseline和seco_ft，表明在粗匹配阶段引入视觉大模型高层语义先验后，模型能够更有效地建模跨时相多模态场景中的稳定区域对应关系。

表2 消融实验

Table 2 Ablation study

/%

算法	MHA@3	MHA@5	MHA@7
baseline	20.12	29.35	34.55
seco_ft	18.39	21.18	22.56
dino_ft	**33.03**	**37.03**	**39.17**

注：加粗字体表示各列最优结果。

上述结果表明，本文方法的性能提升并非单纯来源于额外训练数据或更长训练过程，而主要得益于语义增强机制本身。视觉大模型提供的高层密集语义特征能够为粗层匹配提供更稳定的语义参考，从而弥补原有几何特征在复杂跨时相跨模态场景下表达能力不足的问题。这也进一步验证了本文方法设计的合理性，即在保持XoFTR分层匹配框架不变的前提下，通过语义先验增强粗匹配特征，能够有效提升整体匹配性能。

4 结 论

针对多时相多模态遥感影像匹配中存在的光谱差异显著、时相变化复杂以及几何特征稳定性不足等问题，本文围绕数据基准构建与匹配方法设计两个方面开展研究。首先，基于SeCo数据集构建了面向可见光与近红外跨时相匹配任务的标准化实验基准，并建立了统一的数据组织方式与评价协议；其次，提出了一种基于视觉大模型语义增强的跨模态特征匹配方法。该方法以XoFTR分层匹配框架为基础，在粗匹配阶段引入冻结的DINOv3密集语义特征，并通过语义提升模块实现高层语义先验与几何特征的有效融合，从而增强模型在复杂跨时相、多模态条件下的全局对应关系建模能力。

实验结果表明，本文方法相较原始预训练模型以及仅进行域内微调的基线模型均取得了更优的匹配性能，验证了视觉大模型高层语义先验对多时相多模态遥感影像匹配任务的积极作用。研究结果说明，在保持分层几何匹配框架优势的基础上，通过显式引入语义一致性信息，能够有效提升模型对复杂场景结构的稳定感知能力，从而改善跨模态匹配的鲁棒性与精度。

未来工作将从以下几个方面进一步展开：1）扩展更加贴近真实应用场景的遥感配准数据与评测协议，增强对复杂几何畸变和未见场景泛化能力的验证；2）探索视觉大模型语义先验与几何匹配过程的更深层次耦合机制；3）面向更多类型的多源遥感数据和匹配主干网络，进一步验证大模型赋能几何匹配方法的通用性与可扩展性。

参考文献（References）

Balntas V, Riba E, Ponsa D and Mikolajczyk K. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks//Proceedings of 2016 British Machine Vision Conference 2016. York, UK: BMVA Press

Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. 2021. Emerging properties in self-supervised vision transformers// Proceedings of 2021 IEEE/CVF International Conference on Com-

puter Vision. Montreal, Canada: IEEE: 9630-9640 [DOI: 10.1109/ICCV48922.2021.00951]

Chao P, Kao C Y, Ruan Y S, Huang C H and Lin Y L. 2019. HarDNet: a low memory traffic network//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3551-3560 [DOI: 10.1109/ICCV.2019.00365]

Cheng H, Qiu R C and Sun W B. 2015. Location algorithm of the target in remote sensing image. Infrared Technology, 37(10): 831-835 (程红, 仇荣超, 孙文邦. 2015. 遥感图像目标的定位算法. 红外技术, 37(10): 831-835)

Cong Y Z, Khanna S, Meng C L, Liu P, Rozi E, He Y T, et al. 2022. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery//Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS: 197-211 [DOI: 10.52202/068431-0015]

DeTone D, Malisiewicz T and Rabinovich A. 2018. SuperPoint: self-supervised interest point detection and description//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE: 224-236 [DOI: 10.1109/CVPRW.2018.00060]

Du P J, Fang H, Guo S C, Yang C H and Tang P F. 2025. Research progress of remote sensing change detection based on deep learning: pixel-level, object-level, and scene-level. Remote Sensing Technology and Application, 40(4): 783-794 (杜培军, 方宏, 郭山川, 杨承翰, 唐鹏飞. 2025. 深度学习遥感变化检测研究进展: 像素—对象—场景. 遥感技术与应用, 40(4): 783-794)

Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, et al. 2019. D2-Net: a trainable CNN for joint description and detection of local features//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8084-8093 [DOI: 10.1109/CVPR.2019.00828]

Edstedt J, Bökman G, Wadenbäck M and Felsberg M. 2024. Dedode: detect, don't describe—describe, don't detect for local feature matching//Proceedings of 2024 International Conference on 3D Vision. Davos, Switzerland: IEEE: 148-157 [DOI: 10.1109/3DV62453.2024.00035]

Han X F, Leung T, Jia Y Q, Sukthankar R and Berg A C. 2015. MatchNet: unifying feature and metric learning for patch-based matching//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3279-3286 [DOI: 10.1109/CVPR.2015.7298948]

Jiang X Y, Ma J Y, Xiao G B, Shao Z F and Guo X J. 2021. A review of multimodal image matching: methods and applications. Information Fusion, 73: 22-71 [DOI: 10.1016/j.inffus.2021.02.012]

Li Y L, Liu J B, Liu W C, Liu Y L, Guo Y H, Wang M M, et al. 2023. An improved RIFT algorithm for multi-view radar image registration. Journal of Signal Processing, 39(9): 1633-1650 (李焱磊, 刘静博, 刘文成, 刘云龙, 郭宇豪, 王明明, 等. 2023. 一种用于多视角雷达图像配准的改进RIFT算法. 信号处理, 39(9): 1633-1650) [DOI: 10.16798/j.issn.1003-0530.2023.09.009]

Ma J Y, Jiang X Y, Fan A X, Jiang J J and Yan J C. 2021. Image matching from handcrafted to deep features: a survey. International Journal of Computer Vision, 129(1): 23-79 [DOI: 10.1007/S11263-020-01359-2]

Mañas O, Lacoste A, Giró-i-Nieto X, Vazquez D and Rodríguez P. 2021. Seasonal contrast: unsupervised pre-training from uncurated remote sensing data//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9394-9403 [DOI: 10.1109/ICCV48922.2021.00928]

Mitra R, Zhang J K, Narayan S, Ahmed S, Chandran S and Jain A. 2017. Improved descriptors for patch matching and reconstruction//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 1023-1031 [DOI: 10.1109/ICCVW.2017.125]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Ravi N, Gabeur V, Hu Y T, Hu R H, Ryali C, Ma T Y, et al. 2024. SAM 2: segment anything in images and videos [EB/OL]. [2026-04-12]. https://arxiv.org/pdf/2408.00714.pdf

Ren J W, Jiang X Y, Li Z Z, Liang D K, Zhou X and Bai X. 2025. MINIMA: modality invariant image matching//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 23059-23068 [DOI: 10.1109/CVPR52734.2025.02147]

Revaud J, Weinzaepfel P, De Souza C, Pion N, Csurka G, Cabon Y, et al. 2019. R2D2: repeatable and reliable detector and descriptor [EB/OL]. [2026-04-12]. https://arxiv.org/pdf/1906.06195.pdf

Rublee E, Rabaud V, Konolige K and Bradski G. 2011. ORB: an efficient alternative to SIFT or SURF//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 2564-2571 [DOI: 10.1109/ICCV.2011.6126544]

Sarlin P E, DeTone D, Malisiewicz T and Rabinovich A. 2020. Superglue: learning feature matching with graph neural networks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4937-4946 [DOI: 10.1109/CVPR42600.2020.00499]

Shi Z H, Wu C W, Li C J, You Z Z, Wang Q and Ma C C. 2023. Object detection techniques based on deep learning for aerial remote sensing images: a survey. Journal of Image and Graphics, 28(9): 2616-2643 (石争浩, 仵晨伟, 李成建, 尤珍臻, 王泉, 马城城. 2023. 航空遥感图像深度学习目标检测技术研究进展. 中国图象图形学报, 28(9): 2616-2643) [DOI: 10.11834/jig.221085]

Siméoni O, Vo H V, Seitzer M, Baldassarre F, Oquab M, Jose C, et al. 2025. DINOv3 [EB/OL]. [2026-04-12]. https://arxiv.org/pdf/2508.10104.pdf

Sun J M, Shen Z H, Wang Y A, Bao H J and Zhou X W. 2021. LoFTR:

detector-free local feature matching with transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8918-8927 [DOI: 10.1109/cvpr46437.2021.00881]

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 (唐霖峰, 张浩, 徐涵, 马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报, 28(1): 3-36) [DOI: 10.11834/jig.220422]

Tareen S A K and Saleem Z. 2018. A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK//Proceedings of 2018 International Conference on Computing, Mathematics and Engineering Technologies. Sukkur, Pakistan: IEEE: 1-10 [DOI: 10.1109/ICOMET.2018.8346440]

Tian Y R, Fan B and Wu F C. 2017. L2-Net: deep learning of discriminative patch descriptor in Euclidean space//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6128-6136 [DOI: 10.1109/CVPR.2017.649]

Tian Y R, Yu X, Fan B, Wu F C, Heijnen H and Balntas V. 2019. SOSNet: second order similarity regularization for local descriptor learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1111008-11017 [DOI: 10.1109/CVPR.2019.01127]

Tuzcuoğlu Ö, Köksal A, Sofu B, Kalkan S and Alatan A A. 2024. XoFTR: cross-modal feature matching transformer//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4275-4286 [DOI: 10.1109/CVPRW63382.2024.00431]

Wang W J, Tang B, Gu Z Y and Wang S. 2025. Overview of multi-view 3D reconstruction techniques in deep learning. Computer Engineering and Applications, 61(6): 22-35 (王文举, 唐邦, 顾泽骅, 王森. 2025. 深度学习的多视角三维重建技术综述. 计算机工程与应用, 61(6): 22-35) [DOI: 10.3778/j.issn.1002-8331.2405-0328]

Wu J, Cui Z M, Sheng V S, Zhao P P, Su D L and Gong S R. 2013. A comparative study of SIFT and its variants. Measurement Science Review, 13(3): 122-131 [DOI: 10.2478/msr-2013-0021]

Yan R Y, Li W, Chen Y M, Huang H, Wang W J and Song Y P. 2023. A Siamese network-based image matching algorithm. Journal of Nanjing University (Natural Sciences), 59(5): 770-776 (严镕宇, 李伟, 陈玉明, 黄宏, 王文杰, 宋宇萍. 2023. 一种基于孪生网络的图像匹配算法. 南京大学学报(自然科学), 59(5): 770-776) [DOI: 10.13232/j.cnki.jnju.2023.05.005]

Ye Y X and Shen L. 2016. HOPC: a novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3: 9-16 [DOI: 10.5194/isprs-annals-III-1-9-2016]

Zhu B and Ye Y X. 2024. Multimodal remote sensing image registration: a survey. Journal of Image and Graphics, 29(8): 2137-2161 (朱柏, 叶沅鑫. 2024. 多模态遥感图像配准方法研究综述. 中国图象图形学报, 29(8): 2137-2161) [DOI: 10.11834/jig.230737]

作者简介

张旺, 男, 博士研究生, 主要研究方向为多模态图像匹配。
E-mail: zhangwang@njust.edu.cn
姚亚洲, 通信作者, 男, 教授, 主要研究方向为计算机视觉、模式识别和多媒体技术。E-mail: yazhou.yao@njust.edu.cn
陈涛, 男, 副教授, 主要研究方向为计算机视觉、语义分割和弱监督学习。E-mail: taochen@njust.edu.cn
裴根生, 男, 博士后, 主要研究方向为计算机视觉、图像分割和遥感图像匹配。E-mail: peigsh@njust.edu.cn
李宝晨, 男, 讲师, 主要研究方向为计算机视觉。
E-mail: 15510319144@163.com
韦明韬, 男, 工程师, 主要研究方向为计算机视觉。
E-mail: wmtao@sohu.com