PI-ADFM: Enhancing Multi-Modal Remote Sensing Image Matching through Phase-Integrated Aggregated Deep Features

Haiqing He, Shixun Yu, Yongjun Zhang, *Member*, *IEEE*, Yufeng Zhu, Ting Chen, and Fuyang Zhou

***Abstract*—Geometric distortions and significant nonlinear radiometric differences in multi-modal remote sensing images (MRSIs) introduce substantial noise in feature extraction. Single-branch convolutional neural networks fail to capture global image features and integrate local and global information effectively, yielding deep descriptors with low discriminability and limited robustness. Moreover, the lack of comprehensive training data further limits the network's performance, which pose a formidable challenge to existing matching methods in securing adequate and evenly distributed corresponding points. This paper proposes a novel method called Phase-Integrated Aggregated Deep Feature Matching (PI-ADFM), designed to address these challenges. Initially, a phase structure feature detector is introduced, which amalgamates the structural attributes and phase information of images to distill keypoints that are highly repeatable and exhibit minimal redundancy. Subsequently, an attention-based multi-level feature interaction and aggregation module (MFIAM) is crafted to encapsulate a comprehensive representation of both local and global features of keypoints. This is followed by the integration of a dense feature fusion module (DFFM) designed to sift through and amalgamate key features, thereby capturing highly discriminative deep semantic features that serve as descriptors for similarity measures. Finally, a multi-level outlier removal strategy is proposed to effectively reduce mismatches. Experimental results substantiate that, in juxtaposition with state-of-the-art methods, the PI-ADFM method has significantly augmented the count of matches for optical-infrared and optical SAR images by a factor of at least 1.7 and 3.7, respectively, while concurrently enhancing the accuracy by a minimum of 10% and 6%, respectively. These enhancements markedly bolster the robustness and reliability of MRSI matching endeavors. The source code of this study is freely available at https://github.com/hyhqing/PI-ADFM.**

***Index Terms*—Multi-modal images, feature matching, feature interaction and aggregation, deep learning, phase structure.**

Manuscript received…; revised…; accepted... Date of publication…; date of current version… This work was supported in part by the National Natural Science Foundation of China under Grants 42261075 and 41861062, in part by the Jiangxi Provincial Natural Science Foundation under Grant 20224ACB212003, in part by the Jiangxi Provincial Training Project of Disciplinary, Academic, and Technical Leader under Grant 20232BCJ22002, and in part by the self-deployed Project of National Key Laboratory of Uranium Resources Exploration-Mining and Nuclear Remote Sensing, East China University of Technology under Grant 2024QZ-TD-24. *(Corresponding author: Haiqing He; Yongjun Zhang.)*

Haiqing He, Shixun Yu, Yufeng Zhu, and Fuyang Zhou are with the National Key Laboratory of Uranium Resources Exploration-Mining and Nuclear Remote Sensing, East China University of Technology, Nanchang 330013, China, and also with School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China (e-mail: hyhqing@163.com; ysxun3624@163.com; yfzhu@ecut.edu.cn; fuy_zhou@163.com).

Yongjun Zhang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (zhangyj@whu.edu.cn).

Ting Chen is with the School of Water Resources and Environmental Engineering, East China University of Technology, Nanchang 330013, China (e-mail: ct_201607@ecut.edu.cn).

Digital Object Identifier…

I. Introduction

MULTI-modal remote sensing images (MRSIs) synthesize data from a diverse array of sensors, including those capturing visible light, near-infrared spectra, and employing synthetic aperture radar (SAR) technology. Each sensor modality contributes distinct information, enriching the dataset. The heterogeneity of images procured by disparate platform sensors endows them with a complementary nature, thereby presenting an expansive repository for the sophisticated extraction and analysis of remote sensing intelligence and big data. The amalgamation of remote sensing images is extensively leveraged across a spectrum of applications, such as change detection [1], image registration [2], image fusion [3], image stitching [4], and 3D reconstruction [5]. Within these applications, image matching emerges as an indispensable preliminary step. However, the variance in imaging modalities, viewpoints, resolutions, and temporal acquisition intervals among MRSIs introduces considerable geometric and nonlinear radiometric discrepancies. These factors render the task of MRSI matching exceptionally intricate.

As depicted in Fig. 1, MRSIs exhibit background alterations, intricate geometric distortions, and pronounced nonlinear radiometric variations. Achieving robust MRSI matching hinges on the accurate extraction of salient features and their corresponding descriptors. To tackle these challenges, a multitude of image matching methods has been introduced, which can be broadly categorized into two paradigms: handcrafted feature-based methods [6], [7] and deep learning-based methods [8], [9]. Handcrafted feature-based methods predominantly depend on manually formulated feature detectors and descriptors to facilitate matching. These can be further bifurcated into region-based and feature-based matching strategies, contingent upon the necessity for a descriptor. Region-based matching techniques employ the grayscale values of predefined templates for gauging similarity. Such approaches are susceptible to variations in template dimensions, textural attributes, and geometric

distortions. Conversely, feature-based matching methods craft descriptors predicated on the point, line, and surface characteristics of the imagery. Nonetheless, these features are confined to the lower echelons of image representation, and in scenarios characterized by complex geometric distortions and substantial nonlinear radiometric disparities in MRSIs, they are impeded from yielding semantic features with high repeatability.

Deep learning-based methods harness learning algorithms to extract high-level features. In recent years, these learning-based matching methods have garnered considerable attention, with many leveraging high-level features in lieu of manually designed counterparts. High-level features are characterized by superior distinctiveness and robustness compared to handcrafted features, and deep learning-based methods have demonstrated enhanced performance in image matching tasks [10]. Convolutional neural networks (CNNs) have been integrated into the domain of image matching as extractors of high-level semantic features. They surpass traditional handcrafted features by learning robust features and descriptors through a series of nonlinear transformations, capturing nuanced local semantic details from images [11]. Nonetheless, recent research indicates that CNNs may have limited capacity to discern the spatial relationships between feature positions. When confronting MRSIs, CNN-based methods encounter challenges due to the weak global modeling capabilities of single-branch networks, which hinder the capture of comprehensive global semantic features and impede the integration of feature information [12]. To bolster global feature representation and emphasize the spatial relationships of features, researchers have incorporated the Transformer attention mechanism into image matching [13]. This mechanism is underpinned by self-attention, which amplifies the focus on more salient features during the learning process. By merging CNNs with the Transformer attention mechanism within a cohesive framework, it becomes feasible to capture a spectrum of local and global semantic features, enabling multi-level feature interaction and aggregation (MFIAM), and to construct robust feature descriptors. This synergistic integration is anticipated to enhance performance in MRSI matching.

Despite delivering commendable results within specific domains, prevailing MRSI matching methods grapple with enduring challenges. Traditional keypoint detection using image intensity or gradient information is highly susceptible to noise and nonlinear radiometric distortions. Deep learning-based keypoint detection may suffer from spatial localization degradation due to insufficient retention of low-level structural features in deeper, more semantically enriched layers. In contrast, phase-structural information demonstrates greater resilience to such nonlinear radiometric variations, making it a robust alternative for keypoint detection. Hence, a phase-structural feature detection algorithm is proposed. Existing matching methods exhibit poor adaptability to complex geometric distortions and significant nonlinear radiometric differences. Research has shown that global features possess strong invariance, while single-branch CNNs struggle to capture global features and facilitate the interaction and fusion of local and global features, hindering the extraction of discriminative and robust features from MRSI. To address this issue, we integrate the advantages of both local and global features by constructing a deep feature aggregation network for deep descriptor extraction. A Transformer branch is incorporated into the network to capture global image features, and an attention-based multi-level feature interaction and aggregation module is designed to enhance the fusion of local and global features. To further integrate the features extracted by the backbone network into deep descriptors, a dense feature fusion module is introduced, ensuring the captured key features are embedded into deep descriptors for feature matching. The lack of publicly available MRSI training datasets limits the feature representation and overall performance of models. Acquiring diverse training data is essential for developing models with strong generalization capabilities. Furthermore, ambiguities and multiple solutions in corresponding points introduce challenges during image matching. To mitigate these issues, a multi-level outlier removal strategy is designed. Accordingly, this paper proposes an MRSI matching method called phase-integrated aggregated deep feature matching (PI-ADFM). We have thoroughly evaluated it across diverse MRSI scenarios to confirm its applicability and effectiveness.

(a) (b) (c)

Fig. 1. Examples of multi-modal remote sensing images depicting various combinations of imaging modalities: (a) Optical-Optical; (b) Optical-Infrared; and (c) Optical-SAR.

The principal contributions of this paper are delineated as follows:

(1) To extract structural features that are highly repeatable and possess minimal redundancy, we have introduced pointwise shape-adaptive texture filtering (PSTF) [36] for the purpose of texture filtering in images. Furthermore, by integrating phase consistency information, we propose the phase structure feature detector (PSFD) algorithm, which yields high-quality features suitable for multi-modal remote sensing image matching.

(2) In response to the nonlinear variations in the background of multi-modal remote sensing imagery to obtain highly discriminative feature descriptors, we introduce an interactively aggregated deep feature extraction module that integrates CNN and Transformer attention mechanisms,

enabling the capture of both local and global features. This module encompasses a multi-level feature interaction component tasked with the aggregation and interaction of deep features across dual branches at various stages. Additionally, a feature fusion component consolidates key features into robust deep descriptors, thereby amplifying their distinctiveness.

(3) In pursuit of the efficient training of the PI-ADFM method to enhance its generalization capabilities, we have curated a representative MRSI dataset, which includes three distinct data types: optical, SAR, and infrared. The PI-ADFM method, trained on this comprehensive dataset, has achieved high-precision matching across various platforms and temporal conditions, thereby showcasing its robust generalization ability on analogous data.

The structure of this paper is as follows: Section 2 reviews existing methods in MRSI matching. Section 3 details the PI-ADFM method. Section 4 outlines the dataset, presents experimental results, and evaluates both quantitative and qualitative findings. Section 5 concludes the research.

II. RELATED WORK

High-precision MRSI matching has long been one of the research hotspots. As previously mentioned, multimodal image matching methods can be broadly divided into two categories: handcrafted feature-based methods and deep learning-based methods. This section briefly reviews previous relevant work on image matching.

A. Handcrafted Feature-based Methods

Traditional image matching techniques predominantly depend on image intensity and gradients. Methods based on intensity include normalized cross-correlation [14] and mutual information [15], which utilize pixel values for measuring image similarity. However, these methods are highly sensitive to variations in lighting and noise, particularly when there are substantial differences in intensity between images. They also demonstrate limited robustness in addressing nonlinear radiometric variations. To counter these limitations, researchers have introduced frequency-domain-based methods. The underlying concept involves performing image matching in the frequency domain subsequent to the application of the Fourier transform. Li et al. [16] proposed a method leveraging directional phase features, employing the Harris detector for point feature identification. They utilized Log-Gabor filters to derive phase feature maps in multiple orientations for these points, constructing descriptors from these maps. Ye et al. [17] expanded upon the histogram of oriented phase congruency (HOPC) by introducing the channel feature of oriented gradients (CFOG). This approach calculates gradients in multiple directions to create a three-dimensional directional feature map. Feature descriptors were then formulated by convolving these maps with a Gaussian-like kernel, thereby enhancing matching efficiency and precision. Fan et al. [18] presented angle-weighted oriented gradients (AWOG) descriptors, which allocate gradient values between the two most salient directions and apply a three-dimensional phase correlation for similarity measurement, substantially boosting matching performance. While these frequency-domain methods offer some resilience to noise and nonlinear radiometric distortions, they still confront challenges, including robustness and computational complexity.

Additionally, feature-based matching methods, encompass algorithms such as SIFT [19], SURF [20], and ORB [21]. SIFT-like algorithms are robust to linear radiance differences and seem to have limited adaptation to nonlinear radiance differences. Building upon these foundational algorithms, numerous researchers have contributed improvements, propelling extensive research in the domain of MRSI matching. Li et al. [22] introduced the radiation-variation insensitive feature transform (RIFT) algorithm, which achieves rotation invariance in feature detection through phase consistency, utilizing the maximum index map (MIM) for descriptor construction. These handcrafted descriptors maintain rotation invariance and have yielded satisfactory matching outcomes in MRSIs. Yao et al. [7] proposed the co-occurrence filtering spatial matching (CoFSM) method, which fortifies MRSI matching by employing an optimized euclidean distance function to mitigate the impact of outliers. Gao et al. [23] advanced a feature-based local main orientation multiscale mistogram (MS-HLMO) registration algorithm, employing harris corner detection for feature identification, followed by the computation of partial main orientation histograms. They integrated a histogram-like feature descriptor with generalized gradient location and orientation to counter scale and rotation distortions stemming from diverse viewpoints. Cristiano et al. developed a multispectral feature descriptor specifically tailored for imagery captured across different frequencies of the electromagnetic spectrum. This descriptor has been evaluated on public datasets, confirming its superiority. Subsequently, they introduced a scale- and rotation-robust feature descriptor to address issues related to image scale and rotation. The results demonstrated that this descriptor possesses strong robustness against geometric transformations [24], [25]. Xiong et al. [26], [27] employed self-similarity features for multimodal remote sensing image matching. Their method enhances matching performance by utilizing an improved feature detector to maximize self-dissimilarity and a directional self-similarity feature descriptor. Ye et al. [28], [29] leveraged robust filtering for multimodal remote sensing image matching. Their approach, which combines first and second-order gradients to construct matching descriptors, employs Fourier transform techniques and integral images to enhance matching efficiency, yielding promising results in both optical and SAR image matching. Hu et al. [30] presented a multiscale structure feature transformation method tailored for MRSI matching, utilizing a multiscale structure feature detector for keypoint extraction and employing logarithmic coordinate descriptors based on multiscale structural orientations. This methodology has shown robustness against rotation distortions and pronounced

nonlinear radiometric differences. Wan et al. [31] proposed an MRSI matching method leveraging weighted structure saliency features (WSSF), which constructs scale space through texture scale filtering. They amalgamated second-order Gaussian directional filters, edge confidence maps (ECMs), and phase features to generate salient structural feature maps, exhibiting enhanced robustness over traditional gradient-based maps. Despite these advancements, handcrafted feature-based matching methods encounter intrinsic limitations, particularly in terms of limited representational capacity and a lack of robustness against geometric distortions and significant nonlinear radiometric variations.

B. Deep Learning-based Methods

Deep learning-based matching methods have emerged as a prevalent research focus due to their formidable capacity for nonlinear feature representation, effectively overcoming the limitations inherent in traditional matching techniques. These approaches leverage CNNs to serve as high-level feature extractors, capturing contextual image features through a series of complex nonlinear mappings. Learning-based matching methods are broadly divided into two categories: template-based learning methods [32] and feature learning-based methods [33]. Mei et al. [34] introduced a deep neural network-based method for rapid template matching, extracting pixel distribution information via the expanded slice transform (eSLT) matrix. This process standardizes images of varying modalities to a uniform surface image modality, facilitating MRSI matching. Cao et al. [35] utilized a pre-trained VGG network to distill deep image features and proposed an MRSI matching technique. However, these deep learning-based methods are not without their drawbacks, including challenges in template size determination and elevated computational expenses.

To overcome the limitations of template-based learning methods, particularly their vulnerability to nonlinear radiometric distortions, researchers have introduced a variety of feature learning-based methods. Han et al. [36] initially presented MatchNet, leveraging a Siamese network architecture for feature extraction and similarity assessment. Dusmanu et al. [37] subsequently proposed the innovative D2-Net for MRSI matching, employing the network as both a feature detector and descriptor generator to facilitate end-to-end matching. Revaud et al. [38] further refined the D2-Net algorithm, enhancing keypoint repeatability and descriptor reliability. Sarlin et al. [33], building upon the SuperPoint network framework, introduced SuperGlue, a method adept at concurrent feature matching and outlier removal. Aguilera et al. [39] proposed a novel network designed for the learning of local feature descriptors, employing two distinct spectral descriptors to train a quadruple network. Cristiano et al. [40] augmented the TFeat architecture with an additional mapping layer incorporating Log-gabor filters, thereby enhancing the network's capacity to process nonlinear intensity variations within images. Lan et al. [41] crafted CMM-Net, an evolution of D2-Net, tailored for heterogeneous remote sensing image matching. This network utilizes a dynamic adaptive Euclidean distance threshold coupled with the RANSAC algorithm to effectively constrain and eliminate mismatches. Quan et al. [42] bolstered feature representation through the integration of an adaptive learning attention module, capitalizing on the complementary nature of diverse features for MRSI matching. Continuing this progression, Quan et al. [43] advanced the approach by integrating an attention learning module into a deep CNN and devising a feature optimization loss function. This innovation helps to mitigate network overfitting, thereby enhancing the generalization of semantic features and achieving robust registration of MRSIs. Ye et al. [44] proposed a hybrid matching method that enhances structural features with attention mechanisms, integrating traditional methods with deep learning techniques to improve the matching performance of optical and SAR images. Zhang et al. [45] proposed the modality-invariant consistency matching (MICM) method, which amalgamates CNN and Transformer architectures to synthesize multi-modal features, thereby augmenting feature representation and enabling MRSI matching. Despite the considerable advancements in matching performance by current learning-based methods, challenges persist. For instance, the efficacy of MRSI matching is often curtailed by the insufficient representational capacity of local and global features in single-branch networks. Moreover, there is an absence of comprehensive MRSI datasets, which are indispensable for training robust and widely applicable networks.

III. METHODOLOGY

In this section, we introduce a novel MRSI matching method, which incorporates key components such as PSFD, descriptor generalization through the MFIAM, and a multi-level outlier removal strategy, as depicted in Fig. 2. The foundational framework is composed of three key components: (1) the application of PSFD to detect keypoints that are both highly repeatable and exhibit minimal redundancy; (2) the utilization of MFIAM for the generalization of deep descriptors; and (3) the implementation of a multi-level outlier removal strategy to ensure high-precision matching.

A. Phase Structure Feature Detector

MRSIs, due to their diverse imaging modalities, are prone to significant nonlinear radiometric distortions and speckle noise. Traditional feature detection methods, which depend on image intensity and gradient information, are susceptible to these distortions and noise. In contrast, phase consistency information demonstrates greater robustness against such nonlinear radiometric variations. This robustness stems from the fact that, despite capturing different textural information through various sensors, images retain similar structural features. Accordingly, this paper introduces a phase structure feature detector that capitalizes on both the structural and phase information of images to identify keypoints that are

highly repeatable and exhibit minimal redundancy, as illustrated in Fig. 3. The method commences with the application of PSTF [46] to extract the structural features of the image. It then calculates the phase consistency of the filtered image to generate an edge moment map. Ultimately, the Brisk algorithm [47] is applied for keypoint detection. This method harnesses the structural features of the image to delineate edge characteristics, mitigate noise expression, and isolate salient keypoints.

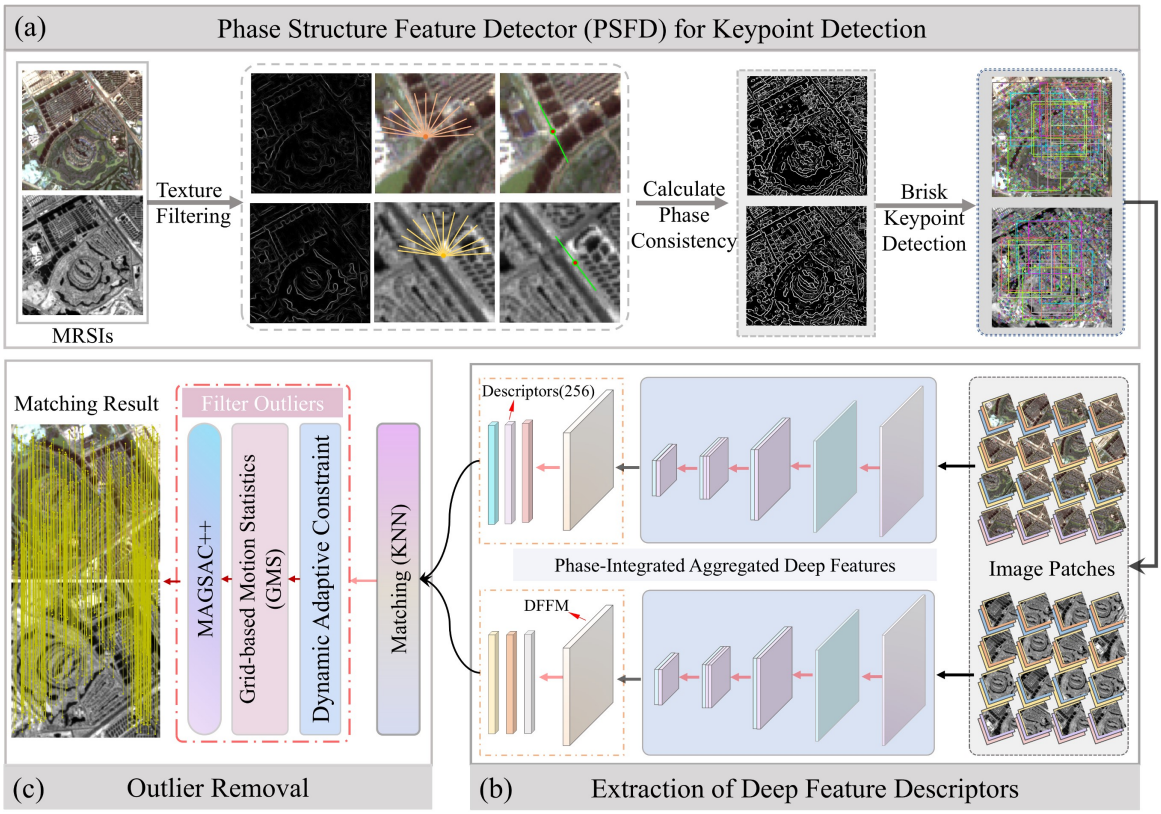


Fig. 2. Schematic representation of the PI-ADFM method's framework, encompassing three principal stages: (a) Phase structure feature detector (PSFD) for keypoint detection; (b) Extraction of deep feature descriptors; and (c) Outlier removal.

(1) PSTF: To effectively mitigate noise interference and accentuate salient structural features, we employ PSTF to extract the image's structural characteristics. The PSTF algorithm is integrated within a joint bilateral filtering framework, which is adept at preserving edges while smoothing homogeneous regions. The formulation of PSTF is delineated as follows:

$$O_p = \frac{1}{K_p}\sum_{q\in N_p} G\left(d(p,q)\right)g\left(\left\|I_q - I_p\right\|\right)I_q, \tag{1}$$

where O_p and I_q respectively represent the pixel values of the input and output, p and q represent the index positions of two pixel values, G and g represent gaussian filtering functions, $d(p,q)$ denotes the distance between p and q, and K_p is the normalization factor for neighboring weights.

In the PSTF algorithm, the orientation of filtering is a critical determinant of the filtering outcome, rendering the acquisition of a guidance image essential. Initially, it is imperative to compute the texture cleanliness for each pixel, centered at pixel p. Line segments of a predefined length k are constructed in n directions, with each segment being assigned a cleanliness measure. The ECM of the input image is derived using a structure-learning-based edge classifier [42]. The clearest line segment is defined as:

$$L_\theta(p) = \arg\min\left\{L_c\left(l_\theta(p)\right) \mid l_\theta(p)\right\}, \tag{2}$$

$$L_c\left(l_\theta(p)\right) = \sigma^2\left(ECV\left(l_\theta(p)\right)\right), \tag{3}$$

where $l_\theta(p)$ represents the line segment centered at pixel p, θ denotes the direction, $L_c\left(l_\theta(p)\right)$ signifies the cleanliness of the line segment, and $ECV\left(l_\theta(p)\right)$ denotes the confidence of pixel values on the line segment. Upon obtaining the cleanliness of pixels, the guidance image can be computed using the following mathematical formula:

$$\begin{cases} D_\theta(p) = \dfrac{1}{k}\displaystyle\sum_{q\in L_\theta(p)} I(q) \\ s_\theta(p) = \exp\left(-\sigma L_c\left(l_\theta(p)\right)\right) \\ G(p) = \dfrac{1}{\displaystyle\sum_{\theta\in(\theta_1\cdots\theta_n)} s_\theta(p)}\displaystyle\sum_{\theta\in(\theta_1\cdots\theta_n)} s_\theta(p)D_\theta(p) \end{cases}, \tag{4}$$

where $D_\theta(p)$ represents the average of $L_\theta(p)$. $s_\theta(p)$ represents the smoothness of the image. $G(p)$ represents the guidance information.

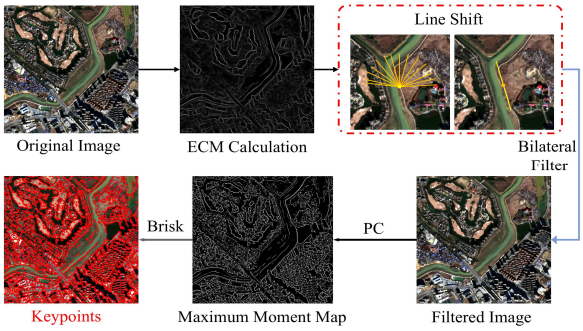


Fig. 3. Workflow of the phase structure feature detector.

(2) Maximum moment map: To further enhance the edge information of the image, phase consistency (PC) calculation is applied to the texture-filtered image. PC can be computed using the following formula:

$$PC(x,y)=\frac{\sum_m\sum_n w_n(x,y)\lfloor A_{mn}(x,y)\Delta\phi_{mn}(x,y)-T\rfloor}{\sum_m\sum_n A_{mn}(x,y)+\varphi}, \quad (5)$$

where $w_n(x,y)$ denotes the weight factor for frequency expansion. $A_{mn}(x,y)$ is the amplitude of the wavelet scale n and orientation at (x,y). T is the noise threshold, φ is a small value used to avoid division by zero. $\lfloor . \rfloor$ denotes a mathematical operation where a value equals itself if it is positive; otherwise, it is zero. $\Delta\phi_{mn}(x,y)$ represents the phase deviation function. Furthermore, the mathematical expression can be represented by:

$$\begin{aligned}A_{mn}(x,y)\Delta\phi_{mn}(x,y)=&\left(E_{mn}(x,y)\phi_E(x,y)+O_{mn}(x,y)\phi_O(x,y)\right)\\&-\left|\left(E_{mn}(x,y)\phi_E(x,y)-O_{mn}(x,y)\phi_O(x,y)\right)\right|\end{aligned}. \quad (6)$$

Then, the maximum moment map can be calculated using:

$$\begin{aligned}M_{\max}&=\frac{1}{2}\left(c+a+\sqrt{b^2+(a-c)^2}\right)\\a&=\sum_\theta\left(PC(\theta)\cos(\theta)\right)^2\\b&=2\sum_\theta\left(PC(\theta)\cos(\theta)\right)\left(PC(\theta)\sin(\theta)\right)\\c&=\sum_\theta\left(PC(\theta)\sin(\theta)\right)^2\end{aligned}, \quad (7)$$

where $M_{\max}$ denotes the maximum moment map. θ is the direction of PC.

This study employs the Brisk algorithm for the rapid extraction of keypoints on the maximum moment map, aiming to identify keypoints that are both robust and repeatable. Based on the comparative results of keypoint detection algorithms such as SIFT, SURF, KAZE, AKAZE, and ORB [49], the Brisk algorithm has been found to offer a good balance between the accuracy and efficiency of keypoint extraction. Consequently, we have opted for the Brisk algorithm to extract keypoints in this study. Fig. 4 presents a comparative evaluation of various keypoint detection methods. Figs. 4(a)-(c) illustrate the impact of nonlinear radiometric distortions and noise on the detected keypoints, which results in the inclusion of noisy keypoints. While these methods may detect a plethora of keypoints, a significant portion of these may be noise points or features that lack robustness. Consequently, this could lead to an escalation in the temporal and computational costs associated with the matching process, potentially compromising the precision of the matching accuracy.

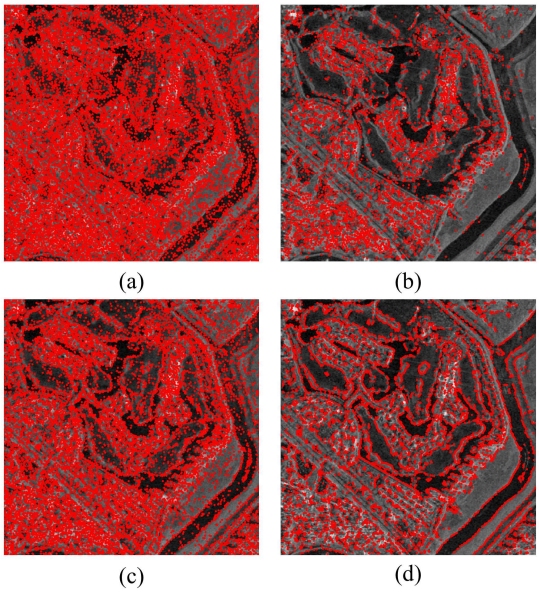


Fig. 4. Comparison of keypoint detection performance across various methods using a SAR image. Subfigures (a), (b), (c), and (d) respectively depict the keypoints detected using the FAST[43], BRISK, PC+FAST, and PSDF algorithms.

In contrast, Fig. 4(d) demonstrates the efficacy of the PSFD algorithm, which significantly mitigates the impact of noise to extract high-quality keypoints, markedly reduces data redundancy, and ensures a uniform distribution of these keypoints across the SAR image.

B. Multi-Level Feature Interaction and Aggregation Module

Existing learning-based methods commonly employ single-branch CNNs for the extraction of deep semantic image features. Primarily adept at capturing local features, CNNs exhibit inherent limitations in global feature extraction, which impedes the comprehensive representation of multi-level spatial semantics. To mitigate this, some researchers have integrated attention modules within CNN architectures to bolster the generalization of global semantic features.

However, the reliance on a single-branch network structure remains restrictive for encapsulating both local and global feature generalization, thus complicating the robust matching of MRSIs. To counter this limitation, this study introduces an

attention-based MFIAM, as illustrated in Fig. 5.

The PI-ADFM is structured around two branch networks: a CNN for local feature extraction and a transformer for global feature extraction. Our PI-ADFM, which is underpinned by an attention mechanism, facilitates the interaction and refinement of local and global features across the stages of the dual-branch network. Subsequently, the channel attention-based dense feature fusion module (DFFM) consolidates deep features to form generalized deep descriptors.

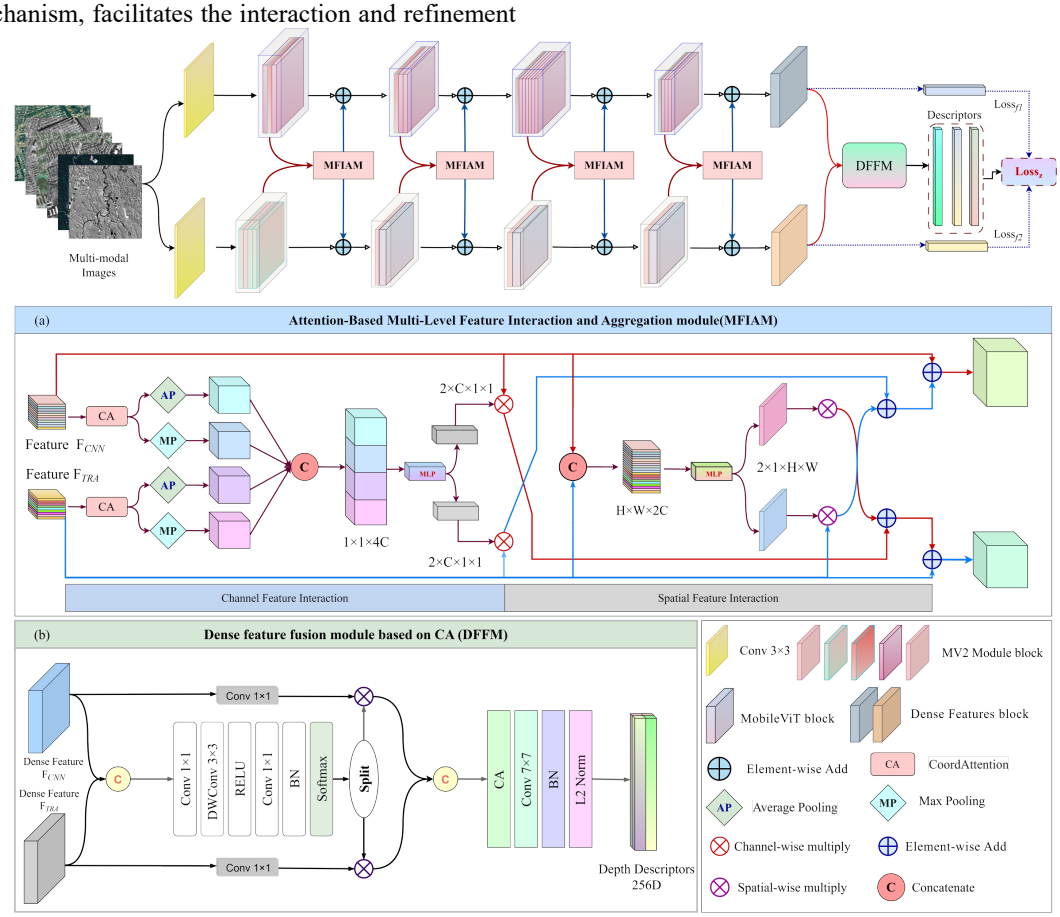


Fig. 5. Network architecture diagram highlighting two principal components: (a) the Attention-based Multi-level Feature Interaction and Aggregation Module (MFIAM), and (b) the Dense Feature Fusion Module (DFFM)

This PI-ADFM network architecture holistically addresses the extraction of both local and global semantic features, as well as the spatial and channel correlations inherent in feature representation, facilitating feature interaction and aggregation across various stages. Specifically, the attention-based MFIAM incorporates a parallel dual-branch structure: the upper branch employs a CNN to capture local features, while the lower branch leverages the Transformer attention mechanism for global feature capture and for conducting feature interaction and aggregation.

Furthermore, we have developed the DFFM based on coordinate attention (CA) to synthesize the deep features extracted by the dual branches. The CA within the DFFM is designed to enhance the salience of key features. Ultimately, the deep semantic features extracted by the PI-ADFM method are integrated to construct comprehensive deep feature descriptors.

(1) Attention-based MFIAM: The CNN branch is tasked with extracting local detail information from MRSIs, whereas the Transformer branch is responsible for capturing global features. To facilitate effective interaction and fusion of these complementary features, this study introduces the attention-based MFIAM, as depicted in Fig. 5(a). Drawing inspiration from a module presented in [50], this PI-ADFM network architecture promotes the interaction and refinement of spatial and channel features across distinct stages, enabling the extraction of features that are both highly repeatable and discernible. The attention-based MFIAM primarily encompasses two interaction stages: channel feature interaction and spatial feature interaction, which are pivotal for the aggregation of robust and descriptive features.

(a) Channel feature interaction: To effectively capture global features from the dual-branch channels and enhance the semantic representation of features, we initially employ CA to augment the positional information of feature maps. Specifically, feature maps F_{CNN}, derived from the CNN and F_{TRA}, derived from the Transformer, are enhanced by CA. This enhancement provides a robust foundation for the precise localization of feature correspondences. Following this, we perform global max and average pooling operations along the channel dimension to generate four distinctive feature vectors

$F_{CNN_{\mathrm{AP}}}$, $F_{CNN_{\mathrm{MP}}}$, $F_{TRA_{\mathrm{AP}}}$, and $F_{TRA_{\mathrm{MP}}}$, which are concatenated to form $F_C \in \mathbb{R}^{1\times1\times4C}$. Finally, a multi-layer perceptron (MLP) is utilized for deep interaction, where weights of the vectors are computed using the sigmoid function, and subsequently divided into two equally sized weight vectors $W_{CNN_C} \in \mathbb{R}^c$ and $W_{TRA_C} \in \mathbb{R}^c$. After achieving deep feature interaction, the corrected feature vectors $F^C_{CNN_{\mathrm{rct}}}$ and $F^C_{TRA_{\mathrm{rct}}}$ can be obtained using Equation 10.

$$F_C = f_{Con}\begin{pmatrix}\left(f_{\mathrm{AP}}\left(f_{\mathrm{CA}}\left(F_{CNN}\right)\right), f_{\mathrm{MP}}\left(f_{\mathrm{CA}}\left(F_{CNN}\right)\right)\right)\\ \left(f_{\mathrm{AP}}\left(f_{\mathrm{CA}}\left(F_{TRA}\right)\right), f_{\mathrm{MP}}\left(f_{\mathrm{CA}}\left(F_{TRA}\right)\right)\right)\end{pmatrix}, \quad (8)$$

$$W_{CNN_C}, W_{TRA_C} = f_{\mathrm{spilt}}\left(w_C\left(\mathrm{MLP}\left(F_C\right)\right)\right), \quad (9)$$

$$\begin{cases} F^C_{CNN_{\mathrm{rct}}} = W_{CNN_C} \otimes F_{CNN} \\ F^C_{TRA_{rct}} = W_{TRA_C} \otimes F_{CNN} \end{cases}, \quad (10)$$

where w_C is the Sigmoid weight function. $\otimes$ represents the channel multiplication operation.

(b) Spatial feature interaction: We further introduce a spatial feature perception module to enable interaction and correction of features at different stages of the feature maps. First, feature map $F_{Conv} = \mathbb{R}^{\mathrm{H\times W\times 2C}}$ is obtained by applying two 1×1 convolutional layers and the ReLU function for feature interaction along the channel dimension, concatenating the input feature maps F_{CNN} and F_{TRA}. Then, compute the weights of the vectors using the Sigmoid function, dividing them into two equally sized weight vectors $W_{CNN_S} \in \mathbb{R}^{\mathrm{H\times W}}$ and $W_{TRA_S} \in \mathbb{R}^{\mathrm{H\times W}}$. After feature correction, feature maps $F^S_{CNN_{\mathrm{rct}}}$ and $F^S_{CNN_{\mathrm{rct}}}$ can be obtained.

$$F_{\mathrm{Conv}} = Conv_{1\times1}\left(\mathrm{Relu}\left(Conv_{1\times1}\left(f_{con}\left(F_{CNN}, F_{TRA}\right)\right)\right)\right), \quad (11)$$

$$W_{CNN_S}, W_{TRA_S} = f_{\mathrm{spilt}}\left(w_S \times F_{\mathrm{Conv}}\right), \quad (12)$$

$$\begin{cases} F^S_{CNN_{\mathrm{rct}}} = W_{CNN_S} \otimes F_{CNN} \\ F^S_{TRA_{rct}} = W_{TRA_S} \otimes F_{TRA} \end{cases}, \quad (13)$$

Then, after the interaction and correction of channel and spatial information from different branches, the corrected features can be calculated by:

$$\begin{cases} F_{CNN_{out}} = F_{CNN} + F^C_{CNN_{rct}} + F^S_{CNN_{rct}} \\ F_{TRA_{out}} = F_{TRA} + F^C_{TRA_{rct}} + F^S_{TRA_{rct}} \end{cases}, \quad (14)$$

(2) DFFM: The dual-branch network is adept at extracting both local and global dense features from the imagery. However, these raw features are not directly amenable to feature matching. To surmount this limitation, we have architected the DFFM, which is predicated on CA. As depicted in Fig. 5(b), this module synthesizes the deep-level features extracted by the dual-branch network into comprehensive deep descriptors. The DFFM is composed of two interlinked submodules: a deep feature aggregation module (FAM) that consolidates feature information, and a feature fusion module (FFM) that integrates these features to forge robust descriptors. In FAM, the outputs $F_{CNN_{out}}$ and $F_{TRA_{out}}$ from the dual-branch network are first concatenated, and two 1×1 convolutional layers along with depthwise separable convolution (DWConv) are used for deep feature aggregation. Next, the softmax function is used to compute the weight maps W_{CNN_D} and W_{TRA_D} for the two different features, and these weight maps are utilized to obtain the aggregated features. The mathematical operations involved in FAM can be represented by:

$$F^{CNN}_{Conv}, F^{TRA}_{Conv} = Conv_{1\times1}\left(F_{CNN_{out}}\right), Conv_{1\times1}\left(F_{TRA_{out}}\right), \quad (15)$$

$$W_{CNN_D}, W_{TRA_D} = soft\max\left(f\left(Conv, BN, \mathrm{Re}lu\right)\times f_{con}\left(F_{CNN_{out}}, F_{TRA_{out}}\right)\right), \quad (16)$$

$$\begin{cases} F^D_{CNN} = W_{CNN_D} \otimes F^{CNN}_{Conv} \\ F^D_{TRA} = W_{TRA_D} \otimes F^{TRA}_{Conv} \end{cases}, \quad (17)$$

where $f\left(Conv, BN, \mathrm{Re}lu\right)$ represents a composite function composed of convolution, BN, and Relu. In FFM, it is necessary to fuse the aggregated dense features to obtain deep feature descriptors for matching. First, we use CA to enhance the representation of key features and suppress irrelevant features. Next, we use convolutional layers and BN layers to obtain a 256-dimensional feature vector α. To facilitate feature matching, we normalize α using L_2. The mathematical operations involved in FFM can be expressed as:

$$\alpha = f\left(Conv, BN, \mathrm{Re}lu\right)\times f_{CA}\left(f_{con}(F^D_{CNN}, F^D_{TRA})\right), \quad (18)$$

$$\alpha = L_2\left(\alpha\right), \quad (19)$$

C. Loss Function

To facilitate the training of the PI-ADFM network, which possesses a composite structure, we implement a composite loss function that encompasses both a primary loss component and an auxiliary loss component. The primary loss is based on the hybrid similarity measure and triplet loss (HSMTL) [51], which serves to supervise the main network. This loss function is designed to minimize the distance between highly similar feature descriptors while maximizing the distance between dissimilar ones, thereby enhancing the discriminative power of the features.

Concurrently, the auxiliary loss incorporates second-order similarity regularization (SOSR) [52] to supervise the branch networks, aiming to mitigate the risks of gradient vanishing and overfitting, thus ensuring a more robust training process.

The HSMTL includes two components: L_2 normalization and triplet loss. The normalization operation helps mitigate issues such as parallel gradients and enhances the network's robustness to image intensity. The formula for L_2 normalization is as follows:

$$R_{L_2} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|x_i\right\| - \left\|x_i^+\right\|\right)^2, \tag{20}$$

where x_i and x_i^+ are a pair of positive descriptors before L_2 normalization. To better generate gradients during network training, a hybrid similarity measure is used to enhance the triplet loss, which can be mathematically calculated using:

$$L_{\text{triplet}} = \frac{1}{N}\sum_{i=1}^{N}\max\left(0, m + H_s\left(\theta_i^{\text{pso}}\right) - H_s\left(\theta_i^{\text{neg}}\right)\right), \tag{21}$$

$$H_s(\theta) = \frac{1}{Z}\left(k\left(1 - H(\theta)\right) + D(\theta)\right), \tag{22}$$

where $k \in (1 \cdots + \infty)$ is a scalar used to adjust the proportion. Z represents the maximum magnitude gradient normalization factor. θ denotes the angle between descriptors. $H(\theta) = \cos(\theta)$, $D(\theta) = \sqrt{2\left(1 - \cos(\theta)\right)}$. The HSMTL can be defined as:

$$L_{\text{HSMTL}} = L_{\text{triplet}} + \delta R_{L_2}, \tag{23}$$

where δ is the regularization parameter.

Additionally, to effectively supervise the learning process of the branch network and to mitigate the risks of overfitting and gradient vanishing, we employ the SOSR loss function. This function leverages both first order similarity (FOS) and second-order similarity for the regularization component of the loss. The FOS calculation is based on a predefined formula that measures the initial-order proximity between feature descriptors, typically expressed as:

$$L_{\text{FOS}} = \frac{1}{N}\sum_{j=1}^{N}\max\left(0, t + d_j^{\text{pos}} - d_j^{\text{neg}}\right)^2, \tag{24}$$

where d_i^{pos} and d_i^{neg} represent the distances between positive samples and negative samples, respectively. The formulation for calculating the second-order similarity among descriptors is:

$$L_{\text{SOS}} = \frac{1}{N}\sum_{j=1}^{N} d^{(2)}\left(x_j, x_j^{'}\right), \tag{25}$$

where $d^{(2)}\left(x_j, x_j^{'}\right)$ is the similarity measure between x_j and x_j^+. x_j and $x_j^{'}$ are a pair of matching positive sample descriptors. The SOSR loss can be mathematically calculated using:

$$L_{\text{SOSR}} = L_{\text{FOS}} + L_{\text{SOS}}. \tag{26}$$

The joint loss function, which amalgamates the HSMTL with the SOSR loss, is a critical component of our training regimen. This composite function is designed to refine the learning process by incorporating both first- and second-order similarities. It can be mathematically expressed as follows:

$$L_z = \phi L_{HSMTL} + \varphi L_{SOSR}^{CNN} + \gamma L_{SOSR}^{TRA}, \tag{27}$$

where ϕ, φ, γ represent three weighting factor. L_{SOSR}^{CNN} and L_{SOSR}^{TRA} denote the loss values in the CNN and Transformer branches, respectively. By incorporating the joint loss function, the training process of the network can be more effectively optimized, leading to better generalization and robustness.

D. Outlier Removal

MRSI often exhibits complex geometric distortions and nonlinear radiometric differences, which inevitably leads to the presence of mismatches. To address this issue, we propose a multi-level outlier removal strategy designed to maximize the preservation of correctly matched points. This strategy comprises two stages: 1) An adaptive threshold constraint; and 2) Coarse filtering based on grid-based motion statistics (GMS) [53], followed by fine matching using the MAGSAC++ algorithm [54].

Traditional methods often rely on ratio thresholds to identify high-quality candidate matches, yet selecting the optimal threshold value can be challenging. To overcome this challenge, our study employs an adaptive threshold constraint approach, which offers a more flexible and effective strategy for extracting candidate matches. Initially, the nearest neighbor algorithm is employed to obtain the closest distance D_{f} and the second closest distance D_{s} of the candidate matches. The average difference D_{Avg} between D_{f} and D_{s} serves as the criterion, formulated as:

$$D_{\text{Avg}} = \frac{1}{n}\sum_{i=1}\left(D_{\text{s}} - D_{\text{f}}\right), \tag{28}$$

$$D_{\text{s}} - D_{\text{f}} > D_{\text{Avg}}, \tag{29}$$

when the candidate matches satisfy equation (29), they are considered to have good quality.

Although adaptive threshold constraints can reduce many mismatches, some may persist. To further refine the matching process, we employ GMS for motion smoothing, which coarsely filters the corresponding points. This approach effectively mitigates the interference from multiple feasible solutions. Additionally, we apply the MAGSAC++ algorithm, leveraging iteratively reweighted least squares for robust estimation, to achieve fine matching. This enhances both the quantity and accuracy of correct matches.

E. Implementation Details

In an effort to expedite the training process, we have incorporated pre-trained weights from the ImageNet into the dual-branch backbone network of the PI-ADFM. The network's backbone, designed for attention-based MFIAM, consists of an improved MobilenetV2 [55] and MobileViTv2 [56]. It is structured into four stages, with feature interaction and correction mechanisms at each stage, culminating in the extraction of 256-dimensional descriptors via the DFFM. To train a robust and generalizable network, we mixed three types of data from the multi-modal remote sensing image dataset as input for training, allowing the deep feature descriptors extracted by the PI-ADFM method to exhibit stronger discriminative power and better robustness.

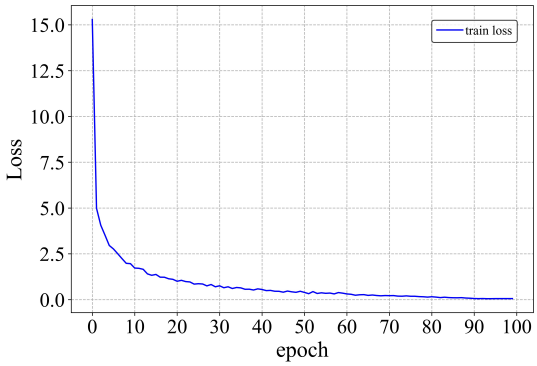


Fig. 6. Loss value during the training process.

Experiments were conducted utilizing an NVIDIA GeForce RTX 4070Ti GPU with 8GB VRAM, employing the PyTorch framework. The training regimen involved the Adam optimizer, initialized with a learning rate of 1e-3, a batch size of 24, and spanning 100 epochs. The regularization parameter for HSMTL was set to the default value of 1.2. After extensive trials, the weight ratios for the joint loss components—ϕ, φ, and γ —were determined to be 0.5, 0.25, and 0.25, respectively. The learning rate was managed using a polynomial decay strategy, and the network was optimized using the aforementioned joint loss function. The loss values during the training process are shown in Fig. 6. As the network training progresses, the loss values gradually decrease and tend to converge. In the final few epochs of training, the loss stabilizes, indicating that the network training is complete.

IV. Experimental Results

In this section, we first provide an overview of the training dataset and experiments, followed by a description of the evaluation metrics. Finally, we analyze and compare the experimental results.

A. Training Dataset

To train the PI-ADFM network, we created an MRSI matching dataset. This dataset includes three types: optical-optical, optical-near-infrared, and optical-SAR, covering remote sensing images with multiple temporal phases, various sensors, and different spatial resolutions. It also includes data on different land cover types, such as buildings, roads, coastlines, forests, lakes, farmland, urban, and rural areas. The dataset comprises 60,000 pairs of image patches, with 20,000 pairs for each type. Some examples are shown in Fig. 7. It should be noted that this study employed transfer learning to share some of the weight parameters from publicly available networks, thus the number of samples was sufficient to meet the training requirements. Consequently, no data augmentation was performed during the training process in this study.

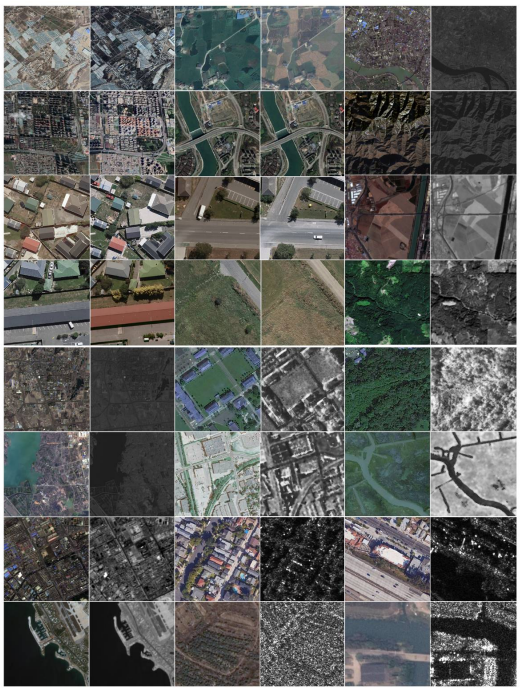


Fig. 7. Training samples.

(1) Optical-Optical Dataset: The optical dataset comprises historical satellite images from Google Earth and the WUH dataset [57]. We sourced 40 multi-temporal pairs of 5000×5000 pixel images from Google Earth, captured between 2020 and 2024, featuring a variety of landscapes such as cities, rural areas, farmlands, forests, rivers, and coastlines. These images have resolutions varying from 5 to 10 m. After performing geographic registration and geometric correction, we cropped the large images into 14,000 pairs of 256×256 pixel patches.

The WUH dataset, utilized for change detection, includes two sets of pre-registered aerial photos from 2012 and 2016 over the same area in New Zealand, with a high resolution of 0.075 m. We manually selected regions with minimal land feature changes and cropped them to acquire 6,000 patch pairs. In total, the optical dataset encompasses around 20,000 pairs of 256×256 pixel multi-temporal patches, with 18,000 reserved for training and 2,000 for testing.

(2) Optical-Infrared Dataset: The dataset primarily includes optical and infrared images from Landsat-9 and Sentinel-2 satellites. The Sentinel-2 data is mainly sourced from the SEN12MS-CR dataset [58]. We obtained Landsat-9 images from the Geospatial Data Cloud, capturing various terrains such as plains, mountains, and hills, with a resolution of 30 m. The optical images are true-color composites created from Landsat-9 bands 2, 3, and 4, captured in 2022. In contrast, the infrared images correspond to Landsat-9 band 5, captured in 2023. After preprocessing, these images were cropped into 10,000 pairs of 256×256 pixel patches.

The SEN12MS-CR dataset, known for its large-scale cloud-free imagery with even distribution across major continents, provided us with cloud-free summer data from Sentinel-2, encompassing all available bands. We created optical images

by combining bands 2, 3, and 4 and used band 8 for the infrared images. From this dataset, we manually selected 10,000 pairs of high-quality images, each 256×256 pixels in size. In total, the optical-infrared dataset comprises approximately 20,000 image pairs, allocated for training (18,000 pairs) and testing (2,000 pairs).

(3) Optical-SAR Dataset: The optical-SAR dataset is sourced from the QXS-SAROPT dataset [59], created by Huang et al. in 2021, and the SAR2Opt dataset [60]. The QXS-SAROPT dataset includes 20,000 pairs of optical and SAR remote sensing image patches, derived from Gaofen-3 SAR satellite images and Google Earth images, covering three port cities: Santiago, Shanghai, and Qingdao. From this dataset, we manually selected 16,000 high-quality image pairs, each 256×256 pixels in size.

The SAR2Opt dataset primarily consists of TerraSAR-X satellite images and Google Earth images, capturing SAR images with a spatial resolution of 1 meter from 2007 to 2013 across 10 cities in Asia, North America, Oceania, and Europe. By manually selecting ground control points, we registered Google Earth images with the corresponding SAR images, resulting in 600×600 pixel images. We utilized only the training portion of this dataset, cropping it into 6,000 pairs of 256×256 pixel images. In total, the optical-SAR dataset comprises approximately 20,000 pairs of image patches, with 18,000 pairs allocated for training and 2,000 pairs reserved for testing.

B. Experimental Data

To validate the effectiveness of the proposed method, we conducted matching experiments using both a self-constructed image dataset and publicly available datasets. Table I details the data sources, sensor types, pixel sizes, and spatial resolutions of the images in our dataset. This dataset encompasses three types of image pairs—optical-optical, optical-infrared, and optical-SAR—comprising a total of 30 matching test pairs, with an equal distribution of 10 pairs per category. Captured by various platforms and sensors at different times and locations, these images exhibit variations in temporal, spatial, scale, and textural attributes. The experimental dataset covers various land cover types, such as urban areas, rural areas, rivers, farmland, lakes, and roads. To address the issue of matching remote sensing images with different spatial resolutions, we initially performed a geographic registration on the image pairs. Subsequently, we extract sub-image blocks that correspond to identical ground areas as candidates for matching, without the necessity of resampling the images of different resolutions, thus preserving the original image detail information to the fullest extent possible. The experimental images originate from a range of satellite sources, including Google Earth, the Landsat series, Sentinel-1 and Sentinel-2, ZY-3, GF-2, GF-3, and TerraSAR-X. Such a diverse set is well-suited for assessing the performance of our method in matching different MRSI types. Fig. 8 displays a selection of these experimental images, showcasing the variety of scenes they represent.

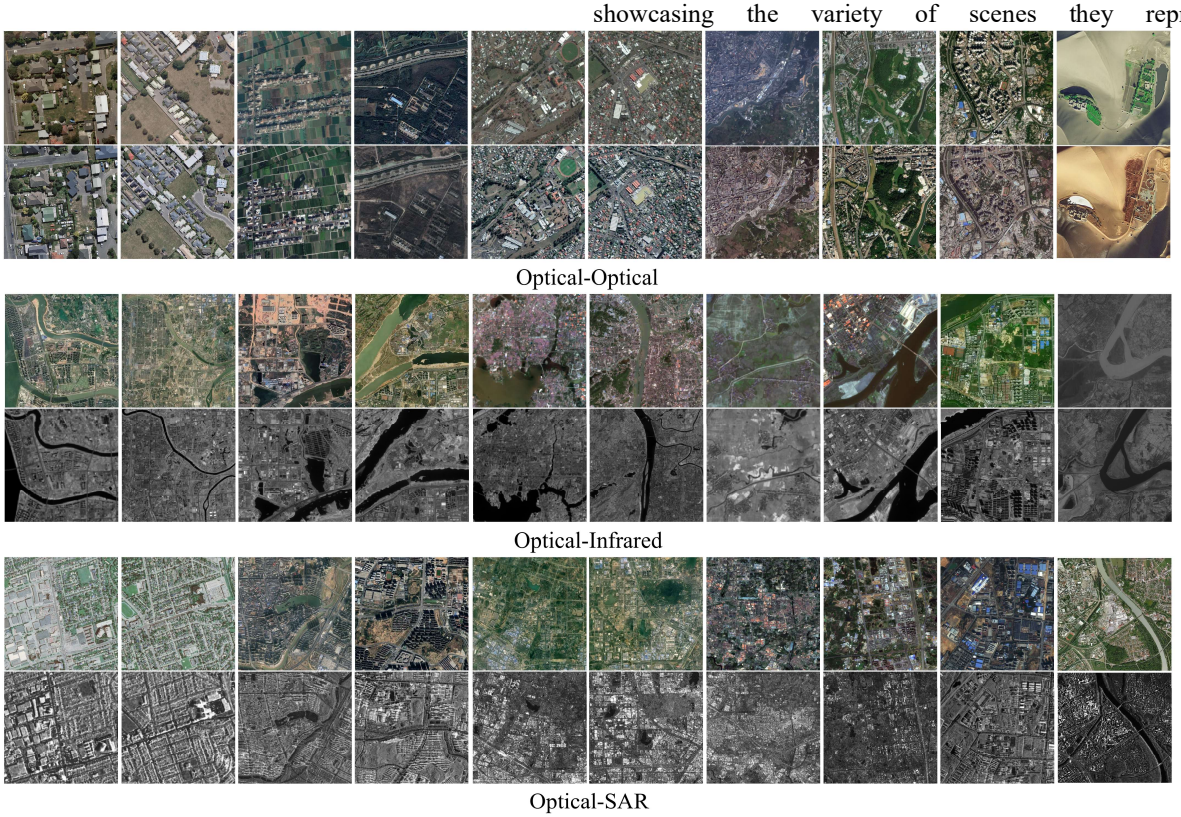


Fig. 8. Examples of multi-modal remote sensing images (MRSIs), illustrating the combination of optical, infrared, and SAR images.

TABLE I
DATA DESCRIPTION

Image type	No.	Image source	Pixel	Resolution (Unit: m)
Optical-Optical	1	Aircraft / Aircraft	512×512 / 512×512	0.3 / 0.3
	2	Aircraft / Aircraft	512×512 / 512×512	0.3 / 0.3
	3	Google Earth / Google Earth	512×512 / 512×512	6 / 8
	4	Google Earth / Google Earth	512×512 / 512×512	6 / 5
	5	Google Earth / IKONOS	512×512 / 512×512	4 / 4
	6	Google Earth / IKONOS	512×512 / 512×512	4 / 4
	7	Google Earth / ZY-3	512×512 / 512×512	8 / 6
	8	Google Earth / ZY-3	512×512 / 512×512	6 / 6
	9	Arcgis / Jilin 1	512×512 / 512×512	5 / 5
	10	MRSIDatasets / MRSIDatasets[61]	500×422 / 500×422	- / -
Optical-Infrared	11	Google Earth / Landsat-8	512×512 / 512×512	25 / 30
	12	Google Earth / Landsat-9	512×512 / 512×512	30 / 30
	13	Google Earth / Sentinel-2	512×512 / 512×512	10 / 10
	14	Google Earth / Sentinel-2	512×512 / 512×512	10 / 10
	15	Landsat-8 / Landsat-9	512×512 / 512×512	30 / 30
	16	Landsat-9 / Landsat-8	512×512 / 512×512	30 / 30
	17	Sentinel-2 / Sentinel-2	512×512 / 512×512	10 / 10
	18	Sentinel-2 / Sentinel-2	512×512 / 512×512	10 / 10
	19	Jilin 1 / GF2	512×512 / 512×512	5 / 3
	20	MRSIDatasets / MRSIDatasets[61]	512×512 / 512×512	- / -
Optical-SAR	21	Google Earth / TerraSAR-X	600×600 / 600×600	1 / 1
	22	Google Earth / TerraSAR-X	600×600 / 600×600	1 / 1
	23	Arcgis / GF3	600×600 / 600×600	3 / 1.5
	24	Arcgis / GF3	600×600 / 600×600	1.5 / 1.5
	25	Sentinel-2 / Sentinel-1	600×600 / 600×600	10 / 10
	26	Sentinel-2 / Sentinel-1	600×600 / 600×600	10 / 10
	27	Google Earth / Sentinel-1	600×600 / 600×600	10 / 10
	28	Google Earth / Sentinel-1	600×600 / 600×600	10 / 10
	29	GF2 / GF3	600×600 / 600×600	3 / 1.5
	30`	MRSIDatasets / MRSIDatasets[61]	500×600 / 500×600	- / -

C. Evaluation Metrics

To comprehensively evaluate the performance of the PI-ADFM method, we use the number of correct matches (NCM) and root mean square error (RMSE) as quantitative evaluation metrics. In this study, we establish a criterion that if the correspondence between a keypoint on the reference image and a keypoint on the matched image is within a given distance ε, the matching point is considered correct. The mathematical expression for this criterion is:

$$\mathrm{Cor}(X):\left\|X_i-\widehat{X}_i\right\|\le\varepsilon, \tag{30}$$

where X_i represents the keypoints on the reference image, and $\widehat{X}_i$ represents the corresponding points on the matched image. RMSE reflects the fluctuation in the accuracy of matches. The transformation matrix is computed based on the correct matches, and using the transformation parameters, the coordinates (x_i,y_i) on the reference image and the transformed coordinates (x_i',y_i') are calculated. Then, RMSE is computed using:

$$\mathrm{RMSE}=\sqrt{\frac{1}{\mathrm{NCM}}\sum_{i=1}^{\mathrm{NCM}}\left[(x_i'-x_i)^2+(y_i'-y_i)^2\right]}. \tag{31}$$

Additionally, another metric known as the F-measure is utilized to evaluate the matching performance, which is defined as the harmonic mean of precision and recall. The calculation formula is as follows:

$$F-measure=\frac{2\times MP\times \mathrm{Re}\,call}{MP+\mathrm{Re}\,call}, \tag{32}$$

where MP represents the accuracy of matches, calculated as the total NCM divided by the total number of matches; Recall is the ratio of NCM to the total number of key points, used to assess the accuracy of correctly matched points.

D. Experimental Results and Analysis

To comprehensively evaluate the performance of the PI-ADFM method, six representative MRSI matching methods were selected for comparison. The descriptions of these methods are as follows: (1) RIFT [26], which is based on handcrafted features utilizing phase consistency information for feature detection and employing maximum and minimum torque graphs for feature description; (2) MS-HLMO [27], which employs Harris for feature detection and uses generalized gradients and directional gradient histograms for feature description; (3) WSSF [31], which utilizes structural and phase information for both feature detection and description; (4) R2D2 [38], which is an end-to-end matching network integrating reliability indicators; (5) CMM-Net [41], which is specifically designed for matching heterogeneous remote sensing images; and (6) LightGlue [62], a deep neural network dedicated to matching sparse local features within image pairs, employs an adaptive mechanism to adjust computational complexity across image pairs of varying difficulty, thereby achieving rapid and accurate matching. In the experiments, the PI-ADFM method set an accuracy threshold of 3 pixels, denoted as $\varepsilon = 3$ [31], [41]. To ensure fairness in experimental comparisons, all methods were evaluated using parameters recommended by the authors who proposed these methods.

(1) Qualitative Evaluation: Figs. 9-11 present comparative results across three types of experimental data, demonstrating that the PI-ADFM method successfully identifies a large number of correct matches in all image pairs.

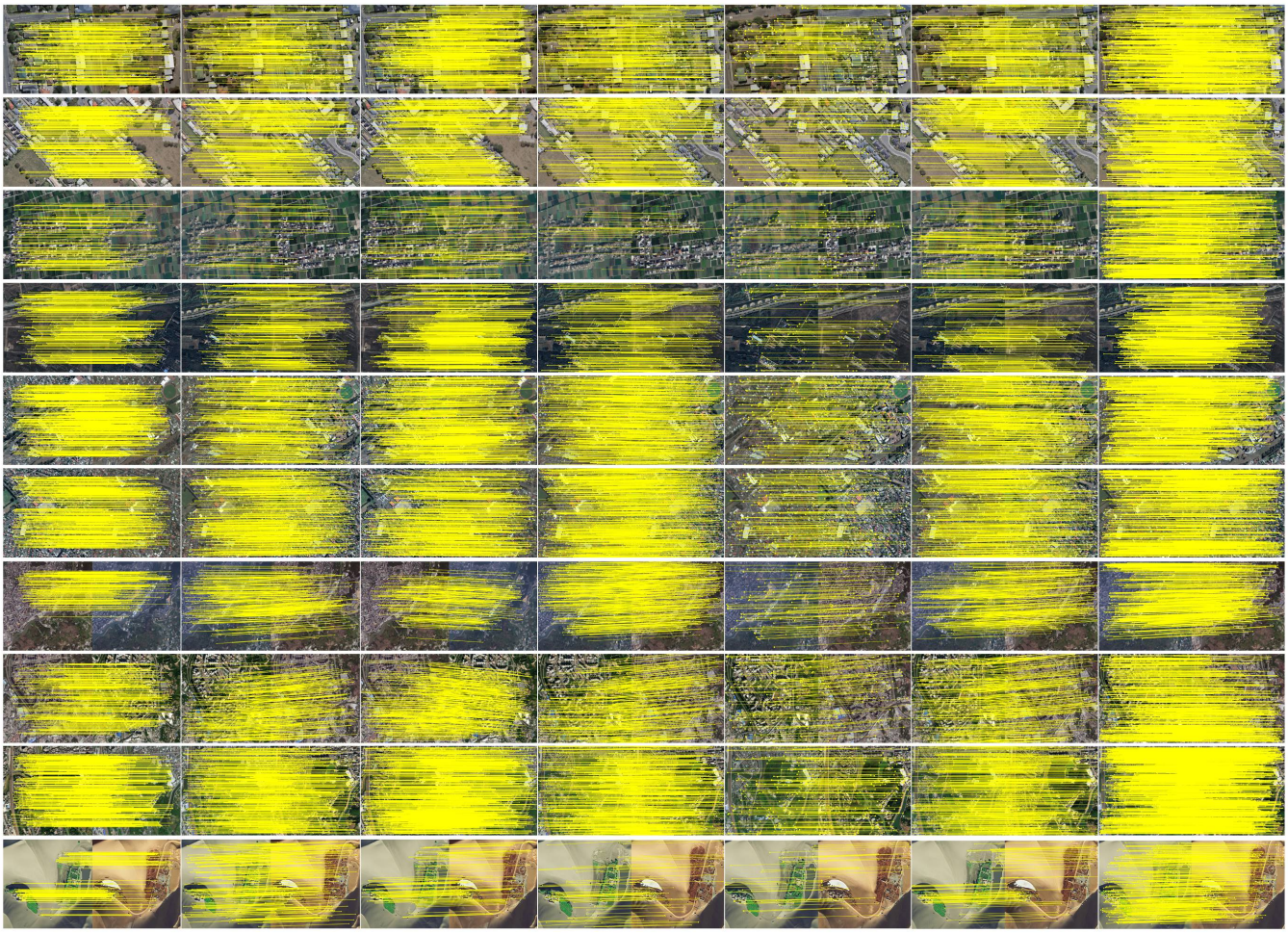


Fig. 9. Optical-Optical image matching results: (from left to right: results obtained by RIFT, MS-HLMO, WSSF, R2D2, CMM-Net, LightGlue, and our method), (from top to bottom: pair1-pair10).

As depicted in Fig. 9, various methods generally perform well in identifying correct matches within optical image pairs. However, the performance of the three handcrafted feature-based methods—RIFT, MS-HLMO, and WSSF—declines notably on image pairs characterized by significant temporal and background variations, such as those labeled as pair3, pair7, and pair10. Among the these, WSSF demonstrates superior performance in optical image matching, particularly in terms of the NCM. It excels by integrating structural features and phase information to extract keypoints and descriptors, showcasing its robustness against complex backgrounds. In comparison, CMM-Net records fewer correct matches in optical images. Despite its simultaneous acquisition of keypoints and descriptors, the method's limited network depth hampers the extraction of sufficient semantic information, thereby reducing the number of correct matches.

Conversely, R2D2 offers distinct advantages for optical image matching by emphasizing both the repeatability of keypoints and the reliability of their detection. LightGlue achieves good matching performance on optical image pairs, primarily due to its training on such pairs, which gives it an advantage in handling optical imagery. In summary, the PI-ADFM method outperforms other state-of-the-art methods by integrating local and global semantic features into robust deep descriptors, which significantly enhances image matching performance. It not only achieves a higher NCM in optical images but also ensures a more uniform distribution of these matches across the image pairs.

Fig. 10 illustrates the matching results for seven methods applied to optical-infrared images. The significant nonlinear radiometric differences between these image types pose a challenge for matching, leading to a decrease in the NCM for all methods when compared to optical-only image pairs. Notably, RIFT outperforms MS-HLMO and WSSF in terms of NCM on optical-infrared pairs. This advantage is attributed to RIFT's use of phase consistency information for descriptor generalization, which confers robustness against nonlinear radiometric variations. In contrast, the performance of the three deep learning-based methods—R2D2, CMM-Net, and LightGlue—is inferior to that of the handcrafted feature-based methods in terms of matching accuracy. The PI-ADFM method, which employs a dual-branch structure, showcases its strength by effectively handling the nonlinear radiometric differences between optical and infrared images, resulting in a higher NCM.

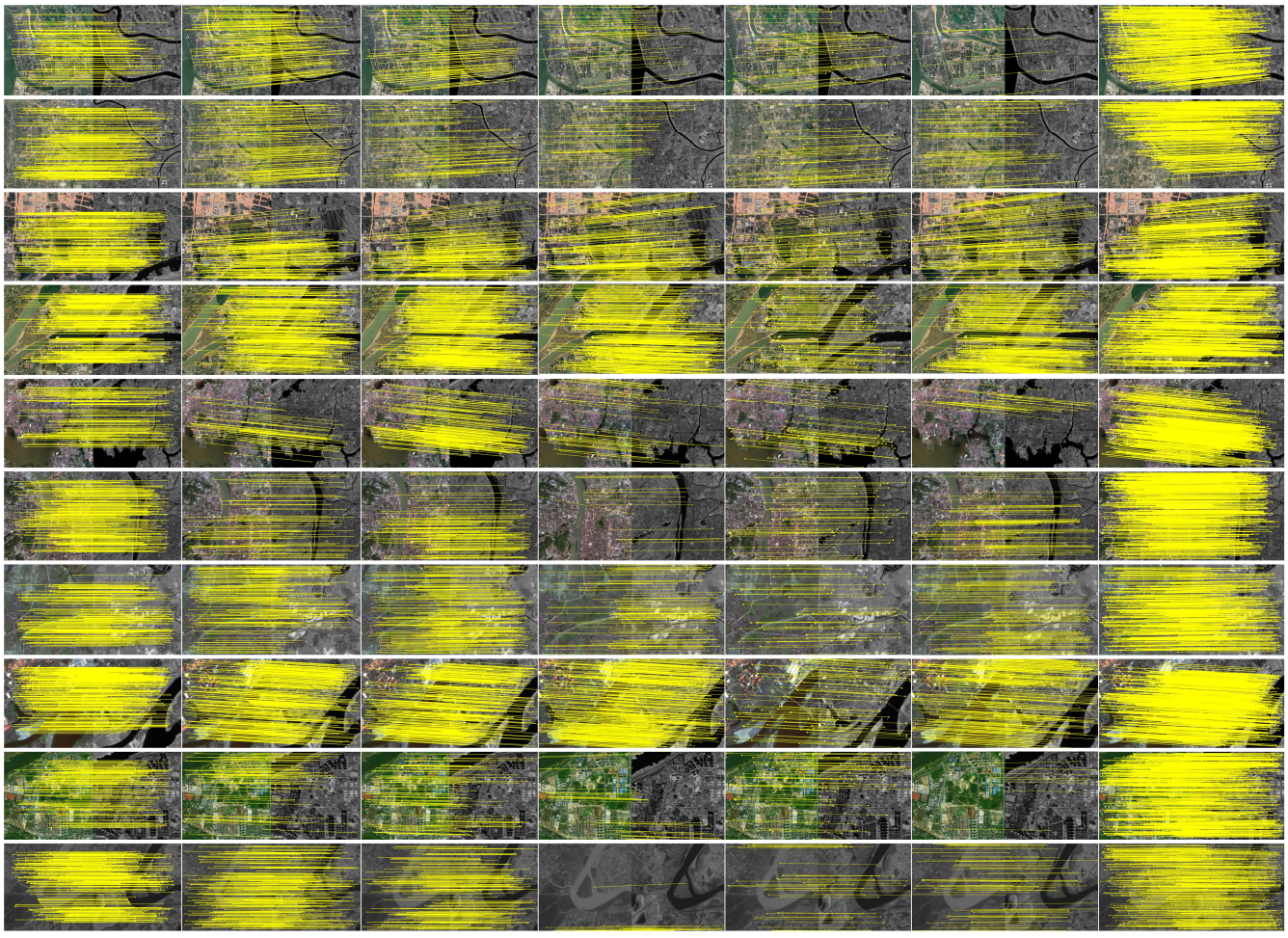


Fig. 10. Optical-Infrared image matching results: (from left to right: results obtained by RIFT, MS-HLMO, WSSF, R2D2, CMM-Net, LightGlue, and our method), (from top to bottom: pair11-pair20).

Fig. 11 displays the matching results of seven methods on optical-SAR images. The complex geometric distortions and significant nonlinear radiometric differences between optical and SAR images make the matching task particularly challenging. As a result, the performance of the six existing methods is less satisfactory compared to the PI-ADFM method. R2D2, CMM-Net, and LightGlue achieve fewer NCM than the handcrafted feature-based methods. Specifically, the performance of R2D2 and LightGlue is hindered by a lack of training data for this type of imagery, which affects their ability to match multiple image pairs. Furthermore, state-of-the-art deep learning-based methods, including R2D2, CMM-Net, and LightGlue, struggle with image pairs that have weak texture information, such as pairs 5 to 8 depicted in Fig. 10. In contrast, the PI-ADFM method demonstrates a stronger generalization ability, identifying a

significantly higher number of NCMs on optical-SAR images. Initially, it employs a structural feature detector to mitigate speckle noise in SAR images, thereby enhancing keypoint repeatability and reducing redundancy. Following this, the attention-based MFIAM interacts with and aggregates local and global semantic information. Meanwhile, the DFFM refines feature representation and consolidates these features into robust deep feature descriptors.

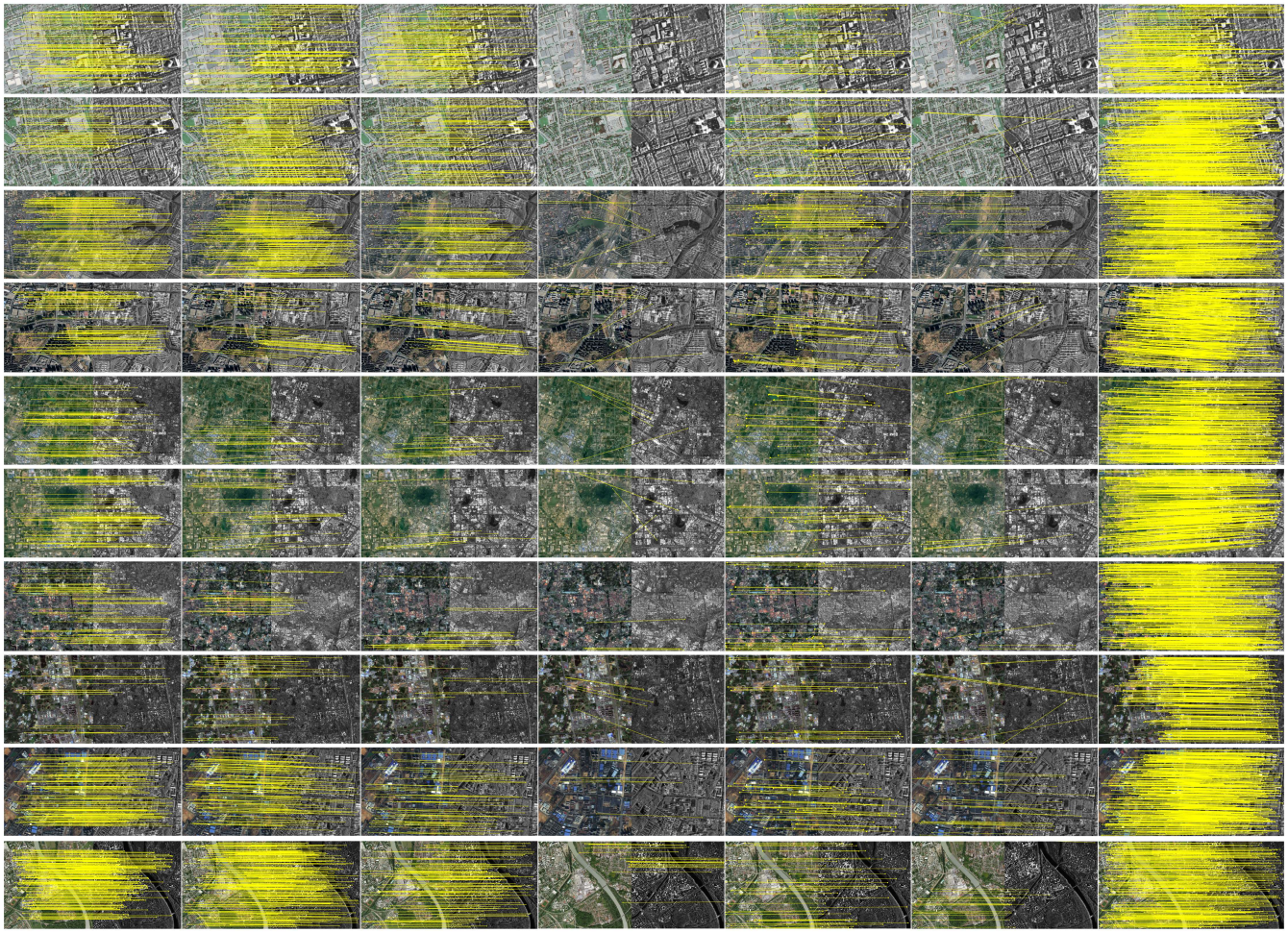


Fig. 11. Optical- SAR image matching results: (from left to right: results obtained by RIFT, MS-HLMO, WSSF, R2D2, CMM-Net, LightGlue, and our method), (from top to bottom: pair21-pair30).

In summary, the PI-ADFM method achieves a higher NCM across all image pairs and exhibits a more uniform distribution of these matches. It demonstrates superior matching performance and better generalization ability compared to other methods. Handcrafted feature-based methods struggle to extract high-level semantic information from multi-modal sources, resulting in inadequate descriptor representation and discriminability. While deep learning-based methods show robustness with optical images, they are sensitive to variations in infrared and SAR images. The suboptimal matching performance can be primarily attributed to two factors: insufficient comprehensive training data for the networks, which affects robustness and generalization; and the constraints of the single-branch network, which limits the interaction and aggregation of deep features. Additionally, all comparison methods exhibit sensitivity to SAR images due to their complex geometric distortions, significant nonlinear radiometric differences when compared to optical images, and the inherent speckle noise in SAR images. These challenges collectively complicate the matching process between optical and SAR images.

(2) Quantitative Evaluation: To provide a quantitative assessment of the matching performance of the seven methods, Fig. 12 compares their performance metrics, including the NCM, RMSE, and F-measure.

The results depicted in Fig. 12 indicate that the PI-ADFM method achieves the best overall performance on MRSI, with higher NCMs and lower RMSE. Specifically, the PI-ADFM method obtains significantly higher NCMs on optical-infrared and optical-SAR images, with RMSE within 2 pixels. The NCMs obtained by the PI-ADFM method are more than those of the handcrafted feature-based methods across all image pairs, with an average NCM of 528 and an average RMSE of 1.689. Among the comparison methods, WSSF performs better than RIFT and MS-HLMO in optical-optical image matching, while RIFT outperforms WSSF and MS-HLMO in optical-infrared and optical-SAR image matching. R2D2 achieves NCMs similar to the PI-ADFM method in a few

optical-optical and optical-infrared image pairs, but its matching accuracy is lower. Additionally, R2D2 exhibits matching errors in optical-SAR images, indicating its lack of robustness to significant nonlinear radiometric differences. The CMM-Net achieves correct matches across all image pairs, with better performance in optical image pairs compared to optical-infrared and optical-SAR images. However, a drawback of this method is its lower matching accuracy. LightGlue demonstrates superior matching performance on optical image pairs relative to optical-infrared and optical-SAR image pairs; however, its performance is markedly inferior when matching optical and SAR images, indicating a limited generalization capability of the network. Moreover, the accuracy of matches achieved by LightGlue is lower compared to that of the PI-ADFM method. In terms of the F-measure, it indicates that the PI-ADFM method exhibits superior matching performance in multi-modal image matching, such as optical to SAR imagery.

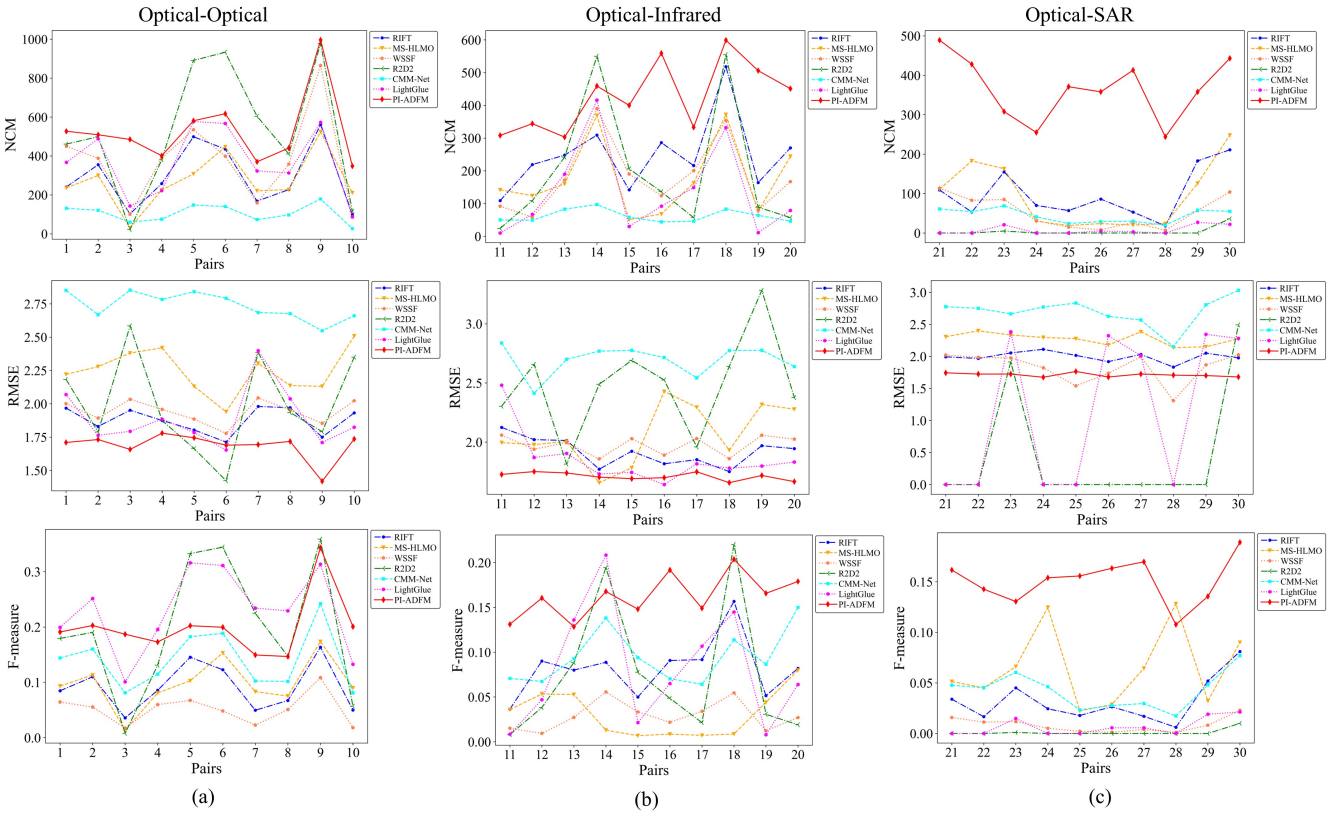


Fig. 12. Quantitative comparison. (a) Optical-Optical. (b) Optical-Infrared. (c) Optical-SAR.

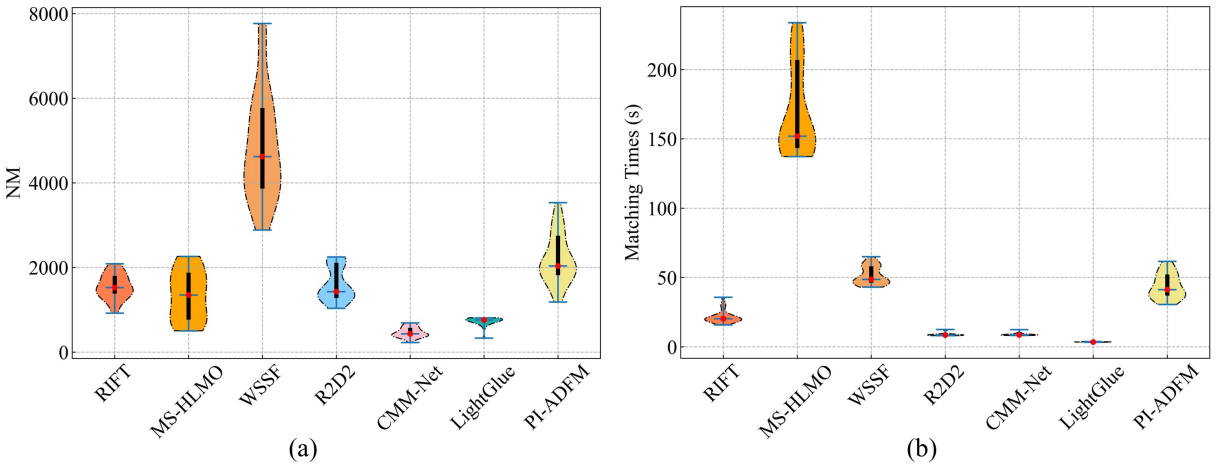


Fig. 13. Comparative performance metrics of seven matching methods, displaying (a) the NMs and (b) the corresponding matching time for each method.

Quantitative results indicate that the PI-ADFM method demonstrates strong robustness against nonlinear radiometric distortions, validating the effectiveness of CNN features and Transformer attention mechanisms in MRSI matching. The PI-

ADFM method combines structural features and phase information to obtain stable keypoints, reducing the impact of noise on keypoint detection. Additionally, the designed MFIAM deeply fuses local and global features, using attention mechanisms in the fusion module to integrate key features, thus generalizing deep descriptors invariant to nonlinear radiometric differences in MRSIs. Finally, a multi-level outlier removal strategy effectively eliminates mismatches, maximizing the retention of correct matches. Therefore, the PI-ADFM method is highly suitable for MRSI matching.

Fig. 13 illustrates a comparison of the number of matches (NMs) and the matching time (MT) between the PI-ADFM method and the other six methods. Fig. 13(a) delineates the NMs comparison, revealing that the WSSF method attains the highest NMs; however, it exhibits lower NCMs compared to the PI-ADFM method, suggesting a higher incidence of mismatches within its set of matched keypoints. The median NM of the PI-ADFM method ranks second, following WSSF, establishing a foundation for securing an adequate NCMs. This is facilitated by the implementation of a multi-level outlier removal strategy, which enhances the inlier rate of matching. The RIFT method demonstrates the most consistent NMs, averaging around 1500, whereas the CMM-Net exhibits the lowest NMs, consequently yielding the lowest average NCMs. Fig. 13(b) illustrates a comparison of matching times between the PI-ADFM method and other comparative methods. Relative to these methods, although the PI-ADFM method does not possess a significant advantage in terms of time efficiency, it markedly outperforms the others in the number of matches and localization accuracy across three types of remote sensing data. Consequently, the additional time expenditure required to achieve higher NCM, F-measure, and RMSE is deemed acceptable.

(3) Evaluation of Image Registration: MRSIs registration is one of the significant applications of image matching, with the accuracy of matching directly affecting the quality of image registration. As depicted in Figure 13, visually, there are no apparent seams in the overlapping regions of all experimental image pairs, indicating satisfactory registration outcomes. This also suggests that the PI-ADFM method possesses excellent matching performance and localization accuracy. The superior registration results across the three types of data demonstrate the robustness of the PI-ADFM method.

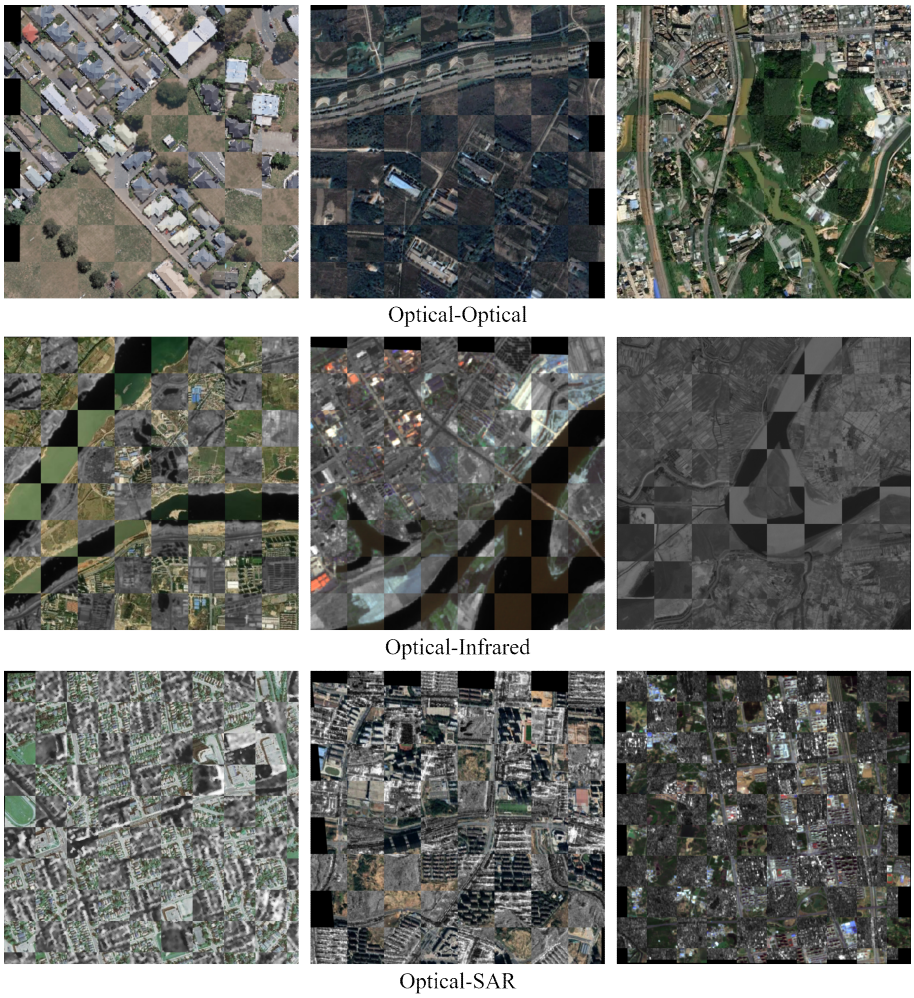


Fig. 14. Registration results for three types of data.

V. Discussion

To substantiate the viability of the PI-ADFM method, ablation studies were meticulously executed to evaluate the influence of the MFIAM and DFFM components on the matching outcomes. The experimental design encompassed a comprehensive selection of 120 image pairs for the matching trials, meticulously allocated into 40 pairs per category. This categorization included pairs of optical-optical and optical-SAR data from the WUH and SAR2Opt datasets, respectively, as well as optical-infrared data extracted from quartet Landsat-9 satellite imagery captured between the years 2022 and 2024. It warrants emphasis that this dataset was distinct from those utilized in the training phase. The experimental findings are delineated in Fig. 15 and Table II, with Table II offering a compilation of the mean metric values for each data category.

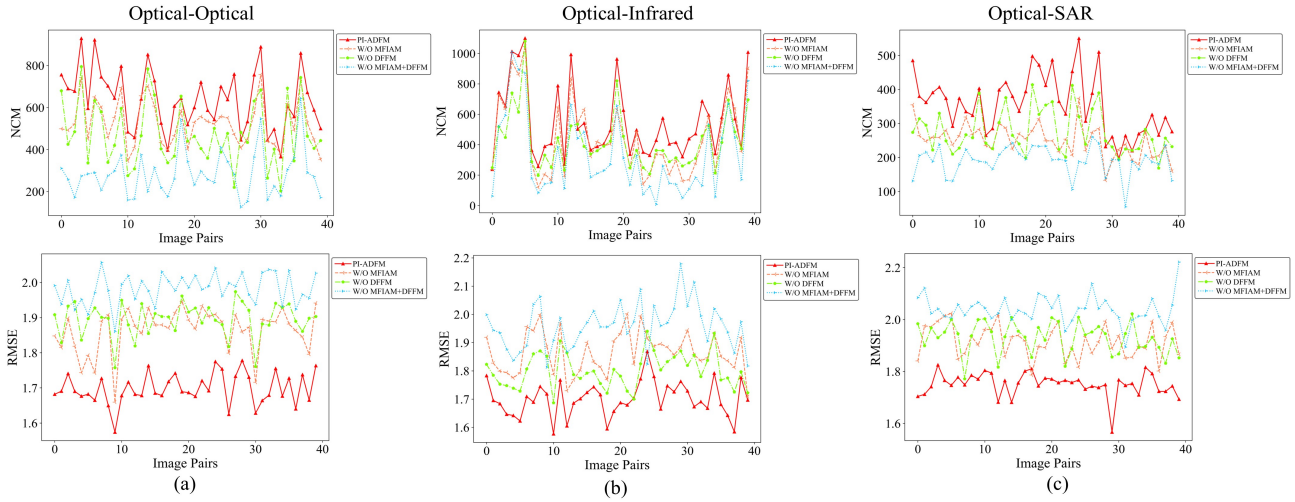


Fig. 15. Ablation Results of the attention-based MFIAM and DFFM. (a) Optical-Optical. (b) Optical-Infrared. (c) Optical-SAR.

TABLE II
Quantitative Matching Results of the PI-ADFM Method

Types	MFIAM	DFFM	NM	NCM	RMSE
Optical-Optical			1556	284	1.983
		✓	1811	524	1.864
	✓		1832	474	1.895
	✓	✓	2132	639	1.698
Optical-Infrared			1583	323	1.959
		✓	1698	418	1.865
	✓		1785	449	1.802
	✓	✓	2065	552	1.701
Optical-SAR			2710	200	2.040
		✓	2927	269	1.916
	✓		2996	275	1.929
	✓	✓	3106	357	1.750

The experimental outcomes underscore the pivotal role of both the MFIAM and DFFM in augmenting the matching performance. Notably, the network's efficacy is markedly compromised upon the sequential removal of these modules. As delineated in Table II, the progressive elimination of the MFIAM and DFFM leads to a decrement in the NMs and NCMs, concomitant with an escalation in the RMSE. The MFIAM is instrumental in facilitating the interplay and amalgamation of local and global features, endowing the resultant features with resilience against geometric distortions and nonlinear radiometric disparities, which in turn bolsters the matching performance. Furthermore, the DFFM module fosters the generalization of salient features, thereby enhancing the network's capacity for feature representation and rendering the ensuing feature descriptors more distinctive. This enhancement, in essence, contributes to the refinement of the matching performance.

VI. Conclusion

To counteract the complexities arising from geometric distortions and nonlinear radiometric variations in MRSI matching, which are often induced by disparate imaging modalities, we propose an innovative MRSI matching method known as PI-ADFM. This novel method commences with the deployment of a phase-structure feature detection algorithm. This algorithm synergizes image structural attributes with phase information, thereby significantly attenuating the adverse effects of image noise and nonlinear radiometric aberrations on the detection of keypoints. Subsequently, the MFIAM is engineered to amalgamate and integrate local and global features, thereby augmenting the capacity for deep feature representation and fortifying the robustness of the ensuing feature descriptors. In the final stage, a tiered strategy for outlier removal is implemented to bolster the inlier rate of NCMs and to refine the precision of the matching process. Compared to the state-of-the-art methods, the PI-ADFM method has significantly augmented the count of matches for optical-infrared and optical-SAR images by a factor of at least 1.7 and 3.7, respectively, while concurrently enhancing the

accuracy by a minimum of 10% and 6%, respectively.

Additionally, we have curated an extensive MRSI dataset, encompassing diverse sensor inputs and sceneries, to facilitate the training of the PI-ADFM network. We have conducted a thorough evaluation of the PI-ADFM method across three distinct data types, juxtaposing its performance against that of predominantly utilized, advanced handcrafted features, and contemporary deep learning-based methods. The outcomes of our evaluation indicate that our method surpasses existing state-of-the-art methods in terms of matching efficacy, yielding a substantial and uniformly dispersed array of matches. However, the PI-ADFM method has not yet addressed the challenges associated with large-scale and rotational variations, which result in suboptimal performance in matching remote sensing images that exhibit such characteristics. Additionally, there is significant scope for improving the matching efficiency. Future work will concentrate on refining the method to improve multi-modal image matching performance in these areas.

REFERENCES

[1] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, "Local Restricted Convolutional Neural Network for Change Detection in Polarimetric SAR Images," *IEEE Trans. Neural Networks Learn. Syst.,* vol. 30, no. 3, pp. 818-833, 2019.

[2] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A Multiscale Framework With Unsupervised Learning for Remote Sensing Image Registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-15, 2022.

[3] N. Liu, W. Li, X. Sun, R. Tao, and J. Chanussot, "Remote Sensing Image Fusion With Task-Inspired Multiscale Nonlocal-Attention Network," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1-5, 2023.

[4] Y. Mo, X. Kang, P. Duan, and S. Li, "A Robust UAV Hyperspectral Image Stitching Method Based on Deep Feature Matching," *IEEE Trans. Geosci. Remote Sens.,* vol. 60, pp. 1-14, 2022.

[5] H. Zhang, Y. Lin, F. Teng, and W. Hong, "A Probabilistic Approach for Stereo 3D Point Cloud Reconstruction from Airborne Single-Channel Multi-Aspect SAR Image Sequences," *Remote Sens.*, vol. 14, no. 22, p. 5715, 2022.

[6] G. Xu, Q. Wu, Y. Cheng, F. Yan, Z. Li, and Q. Yu, "A robust deformed image matching method for multi-source image matching," *Infrared. Phys. Techn.*, vol. 115, p. 103691, 2021.

[7] Y. Yao, Y. Zhang, Y. Wan, X. Liu, X. Yan, and J. Li, "Multi-Modal Remote Sensing Image Matching Considering Co-Occurrence Filter," *IEEE Trans. Image Process.*, vol. 31, pp. 2584-2597, 2022.

[8] S. Cui, M. Xu, A. Ma, and Y. Zhong, "Modality-Free Feature Detector and Descriptor for Multimodal Remote Sensing Image Registration," *Remote Sens.*, vol. 12, no. 18, p. 2937, 2020.

[9] Y. Zhang, G. Ma, and J. Wu, "Air-Ground Multi-Source Image Matching Based on High-Precision Reference Image," *Remote Sens.*, vol. 14, no. 3, p. 588, 2022.

[10] Y. Liao *et al.*, "Feature Matching and Position Matching Between Optical and SAR With Local Deep Feature Descriptor," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.,* vol. 15, pp. 448-462, 2022.

[11] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images," *Remote Sens.*, vol. 14, no. 9, p. 1956, 2022.

[12] Y. Liu, M. Lin, Y. Mo, and Q. Wang, "SAR–Optical Image Matching Using Self-Supervised Detection and a Transformer–CNN-Based Network," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1-5, 2024.

[13] W. Zhong and J. Jiang, "LGFCTR: Local and Global Feature Convolutional Transformer for Image Matching," *Expert Syst. Appl.*, vol. 270, p. 126393, 2025.

[14] J.-C. Yoo and T. H. Han, "Fast Normalized Cross-Correlation," *Circ. Syst. Signal. Pr.*, vol. 28, no. 6, pp. 819-843, 2009.

[15] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Trans. Med. Imaging*, vol. 22, no. 8, pp. 986-1004, 2003.

[16] X. Li, Y. Yang, B. Yang, and F. Yin, "A Multi-source Remote Sensing Image Matching Method Using Directional Phase Feature," *Geomatics and Information Science of Wuhan University*, vol. 45, no. 04, pp. 488-494, 2020.

[17] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and Robust Matching for Multimodal Remote Sensing Image Registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059-9070, 2019.

[18] Z. Fan, L. Zhang, Y. Liu, Q. Wang, and S. Zlatanova, "Exploiting High Geopositioning Accuracy of SAR Data to Obtain Accurate Geometric Orientation of Optical Satellite Images," *Remote Sens.*, vol. 13, no. 17, p. 3535, 2021.

[19] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision.*, vol. 60, no. 2, pp. 91-110, 2004.

[20] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image. Und.*, vol. 110, no. 3, pp. 346-359, 2008.

[21] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.. (ICCV)*, 2011, pp. 2564-2571.

[22] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296-3310, 2020.

[23] C. Gao, W. Li, R. Tao, and Q. Du, "MS-HLMO: Multiscale Histogram of Local Main Orientation for Remote Sensing Image Registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-14, 2022.

[24] C. F. G. Nunes and F. L. C. Pádua, "A Local Feature Descriptor Based on Log-Gabor Filters for Keypoint Matching in Multispectral Images," *IEEE Geosci. Remote Sens. Lett*., vol. 14, no. 10, pp. 1850-1854, 2017.

[25] C. F. G. Nunes and F. L. C. Pádua, "An Orientation-Robust Local Feature Descriptor Based on Texture and Phase Congruency for Visible–Infrared Image Matching," *IEEE Geosci. Remote Sens. Lett*., vol. 21, pp. 1-5, 2024.

[26] X. Xiong, G. Jin, Q. Xu, and H. Zhang, "Self-Similarity Features for Multimodal Remote Sensing Image Matching," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 12440-12454, 2021.

[27] X. Xiong, G. Jin, J. Wang, H. Ye, and J. Li, "Robust Multimodal Remote Sensing Image Matching Based on Enhanced Oriented Self-Similarity Descriptor," *IEEE Geosci. Remote. Sens. Lett.*, vol. 21, pp. 1-5, 2024.

[28] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, and G. Zhang, "A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 331-350, 2022/06/01/ 2022.

[29] Y. Ye, J. Zhang, L. Zhou, J. Li, X. Ren, and J. Fan, "Optical and SAR Image Fusion Based on Complementary Feature Decomposition and Visual Saliency Features," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-15, 2024.

[30] M. Hu, B. Sun, X. Kang, and S. Li, "Multiscale structural feature transform for multi-modal image matching," Inform. Fusion., vol. 95, pp. 341-354, 2023.

[31] G. Wan et al., "Multimodal Remote Sensing Image Matching Based on Weighted Structure Saliency Feature," I *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-16, 2024.

[32] L. Li, L. Han, M. Ding, H. Cao, and H. Hu, "A deep learning semantic template matching framework for remote sensing image registration," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 205-217, 2021.

[33] P. E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching With Graph Neural Networks," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 4937-4946.

[34] L. Mei, C. Wang, H. Wang, Y. Zhao, J. Zhang, and X. Zhao, "Fast template matching in multi-modal image under pixel distribution mapping," *Infrared. Phys. Techn.*, vol. 127, p. 104454, 2022.

[35] F. Cao, T. Shi, K. Han, P. Wang, and W. An, "RDFM: Robust Deep Feature Matching for Multimodal Remote-Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1-5, 2023.

[36] H. Xufeng, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3279-3286.

[37] M. Dusmanu et al., "D2-Net: A Trainable CNN for Joint Description and Detection of Local Features," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8084-8093.

[38] J. Revaud et al., "R2D2: Repeatable and Reliable Detector and Descriptor," 2019, arXiv:1906.06195.

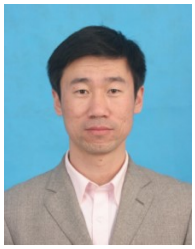
[39] C. A. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-Spectral Local Descriptors via Quadruplet Network," *Sensors.*, vol. 17, no. 4, 2017.
[40] C. Nunes and F. Pádua, "A Convolutional Neural Network for Learning Local Feature Descriptors on Multispectral Images," *IEEE Lat. Am. Trans.*, vol. 20, no. 2, pp. 215-222, 2022.
[41] C. Lan, W. Lu, J. Yu, and Q. Xu, "Deep learning algorithm for feature matching of cross modality remote sensing images," *Acta. Geogr. Sin.*, vol. 50, no. 02, pp. 189-202, 2021.
[42] D. Quan et al., "Multi-Relation Attention Network for Image Patch Matching," *IEEE Trans. Image Process.*, vol. 30, pp. 7127-7142, 2021.
[43] D. Quan et al., "Deep Feature Correlation Learning for Multi-Modal Remote Sensing Image Registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-16, 2022.
[44] Y. Ye, C. Yang, G. Gong, P. Yang, D. Quan, and J. Li, "Robust Optical and SAR Image Matching Using Attention-Enhanced Structural Features," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-12, 2024.
[45] Y. Zhang, C. Lan, H. Zhang, G. Ma, and H. Li, "Multimodal Remote Sensing Image Matching via Learning Features and Attention Mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-20, 2024.
[46] B. Liu and X. Lu, "Pointwise Shape-Adaptive Texture Filtering," *in Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, 2018, pp. 1-6.
[47] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2548-2555.
[48] E. Rosten and T. Drummond, "Machine Learning for High-Speed Corner Detection," *in Proc. ECCV*, 2006, pp. 430-44.
[49] S. A. K. Tareen and Z. Saleem, "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK," in *Proc.* iCoMET, 2018, pp. 1-10.
[50] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679-14694, 2023.
[51] Y. Tian, A. B. Laguna, T. Ng, V. Balntas, K. J. a. C. V. Mikolajczyk, and P. Recognition, "HyNet: Learning Local Descriptor with Hybrid Similarity Measure and Triplet Loss," 2020, arXiv:2006.10202.
[52] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "SOSNet: Second Order Similarity Regularization for Local Descriptor Learning," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11008-11017.
[53] J. Bian, W. Y. Lin, Y. Matsushita, S. K. Yeung, T. D. Nguyen, and M. M. Cheng, "GMS: Grid-Based Motion Statistics for Fast, Ultra-Robust Feature Correspondence," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2828-2837.
[54] D. Baráth, J. Noskova, M. Ivashechkin, and J. Matas, "MAGSAC++, a Fast, Reliable and Accurate Robust Estimator," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1301-1309.
[55] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510-4520.
[56] S. Ji, S. Wei, and M. Lu, "Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574-586, 2019.
[57] S. Mehta et al., " Separable Self-attention for Mobile Vision Transformers," 2022, arXiv:2206.02680.
[58] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A Remote-Sensing Data Set for Multimodal Multitemporal Cloud Removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-14, 2022.
[59] M. Huang et al., "The QXS-SAROPT Dataset for Deep Learning in SAR-Optical Data Fusion," 2021, arXiv:2103.08259.
[60] H. Toriya, A. Dewan, and I. Kitahara, "SAR2OPT: Image Alignment Between Multi-Modal Images Using Generative Adversarial Networks," *in Proc. IGARSS*, 2019, pp. 923-926.
[61] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22-71, 2021/09/01/ 2021.
[62] P. Lindenberger, P. E. Sarlin, and M. Pollefeys, "LightGlue: Local Feature Matching at Light Speed," *in Proc. IEEE Int. Conf. Comput. Vis.(ICVV)*, 2023, pp. 17581-17592.

Haiqing He received the Ph.D. degree in geodesy and survey engineering from Wuhan University, Wuhan, China, in 2013.

From 2017 to 2019, he was with the Lyles School of Civil Engineering, Purdue University, IN, USA, as a Post-Doctoral Fellow. He is currently a Full Professor with the School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang, China, and also with Jiangxi Key Laboratory of Watershed Ecological Process and Information, East China University of Technology, Nanchang, China. His research interests include low-attitude photogrammetry, image matching, artificial intelligence driven remote-sensing image interpretation, and 3D reconstruction.

Shixun Yu received the B.S. degree in surveying and mapping engineering in 2022 from East China Jiaotong University, Nanchang, China, where he is currently working toward the M.S. degree in Surveying and Mapping.

His current research interests include photogrammetry and remote sensing, image processing, and deep learning.



Yongjun Zhang received the BS degree in geodesy, the MS degree in geodesy and surveying engineering, and the PhD degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively. He is currently the dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 180 research articles and three books. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, object information extraction and modeling with artificial intelligence, integration of LiDAR point clouds and images, and 3D city model reconstruction. He is the co-editor-in-chief of the *Photogrammetric Record*.

Yufeng Zhu received the Ph.D. degree in geodesy and surveying engineering, School of Geosciences and Infor-physics, Central South University, Changsha, China, in 2013.

His research interests include theory and application of InSAR technology.

Ting Chen received the B.S. and M.S. degrees in water conservancy engineering from Wuhan University, Wuhan, China, in 2013 and 2016, respectively.

She is currently a Lecturer with the School of Water Resources and Environmental Engineering, East China University of Technology, Nanchang, China. Her research interests include photogrammetry and remote sensing.

Fuyang Zhou received the B.S. and M.S. degrees in surveying and mapping engineering from East China University of Technology, Nanchang, China, in 2021 and 2024, respectively. He is currently pursuing a Ph.D. degree in surveying and mapping engineering. His current research interests include photogrammetry and remote sensing, image processing, and machine learning.