Multimodal Remote Sensing Image Robust Matching Based on Second-Order Tensor Orientation Feature Transformation

Yongjun Zhang, *Member, IEEE*, Peihao Wu, Yongxiang Yao, Yi Wan, *Member, IEEE*, Wenfei Zhang, Yansheng Li, *Senior Member, IEEE*, and Xiaohu Yan

***Abstract*— Nonrigid deformation (NRD) and image noise in multimodal remote sensing images (MRSI) lead to abrupt changes in feature directions, resulting in sensitivity to rotational variation, sparse correct matches, and high false match rates. In order to address these challenges, this article proposes a second-order tensor orientation feature transformation (SOFT) method to improve the rotational invariance of MRSI matching and increase the number of correct matches (NCMs). The SOFT method has two main contributions: 1) a novel second-order tensor orientation descriptor is constructed by generating a tensor orientation feature map using a designed second-order tensor function, which is then combined with a gradient location and orientation histogram (GLOH)-like descriptor framework to achieve robust rotational invariance in multimodal image matching and 2) an error-removal global-local iterative optimization (EGIO) is introduced, employing a skewness of mixed pixel intensity (SMPI) function to automatically select matching seed points, followed by an iterative partition optimization strategy for refining corresponding points. Experiments on 744 groups of typical MRSIs demonstrate that the SOFT method significantly outperforms nine state-of-the-art methods, achieving an average 97% improvement in the NCMs, an average 25.51% improvement in the rate of correct matches (RCMs), and an average reduction in RMSE of 2.69 pixels. The proposed SOFT method, thus, offers robust MRSI matching with strong rotational invariance and precise identification of corresponding points, proving its effectiveness for complex remote sensing scenarios. Access to experiment-related data and codes will be provided at https://skyearth.org/research/.**

***Index Terms*— Bidirectional matching, gradient location and orientation histogram (GLOH)-like, multimodal remote sensing image (MRSI), rotation invariant, second-order tensor orientation feature, skewness of mixed pixel intensity (SMPI).**

Received 22 September 2024; revised 28 December 2024; accepted 20 January 2025. Date of publication 27 January 2025; date of current version 18 February 2025. This work was supported in part by the Key Program of the National Natural Science Foundation of China under Project 42030102; in part by the National Natural Science Foundation of China under Project 42471470 and Project 42401534; in part by Shenzhen Science and Technology Program under Grant 20220812102547001; in part by the Research Projects of Department of Education of Guangdong Province under Grant 2024KTSCX052; and in part by Shenzhen Polytechnic University Research Fund under Grant 6023310030K, Grant 6022312044K, Grant 6023240118K, and Grant 6024310045K. *(Corresponding authors: Yongxiang Yao; Yi Wan.)*

Yongjun Zhang, Peihao Wu, Yongxiang Yao, Yi Wan, Wenfei Zhang, and Yansheng Li are with the School of Remote Sensing Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zhangyj@whu.edu.cn; wupeihao@whu.edu.cn; yaoyongxiang@whu.edu.cn; yi.wan@whu.edu.cn; zhangwenfei@whu.edu.cn; yansheng.li@whu.edu.cn).

Xiaohu Yan is with the School of Undergraduate Education, Shenzhen Polytechnic University, Shenzhen 518055, China (e-mail: yanxiaohu@szpu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3535154

I. Introduction

MULTIMODAL remote sensing image (MRSI) matching is the process of identifying and aligning two or more images with overlapping regions acquired by different sensors [1]. With the rapid development of sensor technology and artificial intelligence science, the data sources of MRSIs are becoming more and more abundant [2]. MRSIs are widely used in many fields, such as change detection, 3-D reconstruction, simultaneous localization and mapping (SLAM) positioning, and carbon neutrality; however, the prerequisite for the application of MRSI is that they need to be matched and aligned.

A large number of experts and scholars have carried out multimodal research on region-based methods, feature-based methods [3], and deep learning-based methods [4] and achieved certain results; however, on whether it is based on traditional multimodal matching or deep learning-based matching, there are still challenges in achieving robust rotation-invariant matching of MRSIs. There are mainly two problems: 1) due to the contrast difference, intensity difference, and Nonrigid deformation (NRD) between the MRSI, the extreme value of the orientation feature of the MRSI changes suddenly, resulting in the rotation relationship between the images and 2) the high error rate of MRSI matching makes it difficult to identify the correct corresponding points.

We, therefore, propose a second-order tensor orientation feature transformation (SOFT) MRSI matching method (see Fig. 1). In order to further enhance feature extraction robustness, the proposed SOFT method integrates a GLOH-like descriptor framework. The gradient location and orientation histogram (GLOH) descriptor, an extension of the well-known scale-invariant feature transform (SIFT) descriptor, provides enhanced robustness by offering a more comprehensive representation of image features. By using log-polar spatial bins, GLOH achieves finer feature discrimination and improved rotation invariance, making it particularly suitable for the complex and variable conditions present in MRSI matching. This integration allows the SOFT method to maintain consistent identification of corresponding points even in the presence of nonlinear distortions and varying sensor modalities, effectively addressing some of the inherent challenges in multimodal image matching.

Subsequently, the SOFT method uses a second-order tensor orientation feature to accurately compute the feature

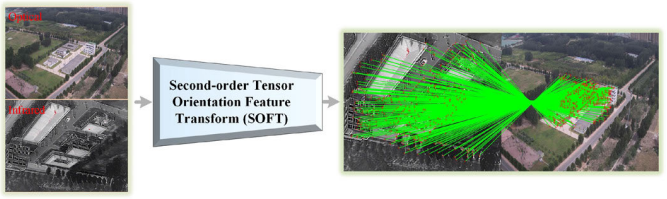

Fig. 1. Matching results of our proposed SOFT method.

descriptor, which is further refined by the error-removal global-local iterative optimization (EGIO) method. This EGIO method is specifically designed to remove outliers and enhance the accuracy of corresponding point identification, thereby ensuring robust matching with strong rotation invariance in MRSI. Within the EGIO method, we introduce the skewness of the mixed pixel intensity (SMPI) feature, inspired by skewness metrics used in the probability distribution analysis of real-valued random variables. Skewness [5] is a statistical measure that quantifies the asymmetry of a distribution around its mean, ideally suited for identifying key features from the complex mixed pixels in remote sensing imagery.

The proposed SOFT method has two main contributions.

1) A novel SOFT method is proposed. The method generates a tensor orientation feature map through the designed second-order tensor orientation function, and the orientation feature map combines a GLOH-like descriptor framework to calculate the descriptor vector, which significantly improves the rotation of the descriptor.
2) An EGIO is proposed. This method uses the designed SMPI to adaptively extract the seed points and then combines the seed points to achieve fine filtering of the corresponding points through the EGIO method, which significantly improves the accuracy of the recognition rate of the corresponding point.

This article is structured as follows. Section I describes the purpose of the study, the limitations of previous studies, and the significance of this paper. Section II reviews the related methods and their challenges. Section III details the processing of the proposed SOFT method. Section IV provides a comprehensive experimental analysis, including the effects of different parameter settings on SOFT performance. Finally, Section V summarizes the contributions of this study and suggests future directions.

II. Related Work

Image matching is divided into traditional methods and deep learning methods [6], [7]. Traditional methods can be further divided into region-based matching and feature-based matching.

A. Region-Based Methods

Region-based matching methods (also known as intensity-based methods) are more commonly used to measure similarity based on strength and mutual information (MI) methods or similarity measures in the transform domain [8]. In the MRSI iterative process, various criteria are designed according to the intensity difference between two images [9]. The commonly used similarity measures include a sum of squared differences (SSDs), a sum of absolute differences, cross correlation, and normalized cross correlation (NCC). SSD and NCC are sensitive to NRDs and are not suitable for MRSI matching. In contrast, MI is more robust to complex NRDs and has been successfully applied to multisource image alignment; however, MI is usually computationally intensive [10], and MI is more sensitive to noise and has poor positional accuracy [11].

In order to ensure the matching accuracy while resisting the problems of nonlinear grayscale differences, NRD, and geometric difference, some experts and scholars have proposed transfer optimization to maximize MI [12] and a combination of oriented gradient distance histogram and gray wolf optimizer [13], to thus improve accuracy and avoid matching results falling into local optimum solutions. Meanwhile, matching based on efficiency divergence as a similarity measure [14], matching based on the histogram of orientation phase consistency (PC) [15], based on channel features of orientated gradients [16], and controlled structure feature matching [17] combining steerable filters of first- and second-order channels, which are better at overcoming problems such as NRD and contrast differences between MRSIs, can achieve robust matching in MRSIs with displacement differences only, but do not have geometric invariance and perform poorly, especially against scale and rotation invariance. The main reason is that they rely on MRSI's own geospatial landmarks (e.g., rational polynomial coefficient parameters for satellite images and position and orientation system data for unmanned aerial vehicle (UAV) images); however, they do not work for some satellite images, UAV images, or ground images without spatial reference information. Feature-based methods that do not rely on spatial reference information are, therefore, of wide research value.

B. Feature-Based Methods

Feature-based methods, starting with the SIFT matching proposed by Lowe [18] have seen rapid development of many SIFT-like methods [19]. These methods have been explored from various perspectives, such as scale robustness, rotational invariance, binary description optimization, descriptor optimization, and multifeature extraction. The inability of gradient features to accommodate modal differences in multimodal images, however, makes such methods inappropriate for multimodal image matching. Chen et al. [20] proposed the partial intensity invariant feature descriptor (PIIFD) algorithm, which achieves rotational invariance by computing the grayscale features of the image and works for small NRDs and contrast differences between images. Ma et al. [21] proposed the position scale orientation-SIFT (PSO-SIFT) algorithm, which works well for both nonlinear brightness differences and rotation variations by building new image gradient features, whereas contrast differences and signal-to-noise differences are more sensitive. Sedaghat and Mohammadi [22] proposed the histogram of the oriented self-similarity algorithm HOSS, which can guarantee rotation invariance better, and it performs

well in large contrast and nonlinear radiometric aberrations in multimodal images. In contrast, the oriented self-similarity (OSS) method proposed by Xiong et al. [23] has overcome the differences in MRSI, but there is a loss of rotational invariance to the images. This author's improved ASS [24] algorithm on OSS is much better adapted to rotational invariance but still not well adapted to MRSI with large NRD. Other experts and scholars have addressed MRSI matching from a phase-coherence model. For example, Li et al. [25] proposed the radiation invariant transformation feature matching (RIFT) method, which includes a maximum index map that can overcome the NRD discrepancy of MRSI better; however, it requires a strategy of ring feature calculation to overcome the rotation discrepancy, which is less technically efficient. Li et al. [26], therefore, proposed a new rotation strategy to optimize the efficiency. Yao et al. [27] proposed a histogram of the absolute phase orientation matching method, in which an absolute phase orientation feature is designed to adapt to differences between MRSIs and resist scale, displacement, and rotation differences between images, but the limitation of this method is that it can only be applied to matching tasks with small rotation differences. Yang et al. [28] proposed a robust matching algorithm that designs a local phase sharpness orientation feature to accommodate MRSI matching and improve the applicability of MRSI rotation differences. Yao et al. [29] proposed a multiorientation feature-based diffusion tensor descriptor (MoTIF), which can be better used in MRSI matching with large noise differences, but this method has limited support for rotation transformations. Recently, the cooccurrence filter space matching (CoFSM) method [1], Max-index-based local self-similarity descriptor method, rotation-invariant self-similarity descriptor matching method [30], adjacent self-similarity matching method, multiscale adaptive binning phase congruency feature, and matching algorithm [31], [32] have been proposed to reduce MRSI differences by improving the image scale space for matching. The matching algorithm [33] is for enhancing multimodal image similarity by establishing local normalized filtering. In addition, a recent study [34] improved the matching accuracy by reducing nonlinear geometric and radial distortions through detailed texture removal and radiation invariant similarity functions. The AMES method [35] optimized the filtering parameters through adaptive prediction, enhanced cross-modal feature extraction, and used a coarse-to-fine strategy for matching, which provided a higher success rate (SR) and matching accuracy. All of these methods have improved the MRSI matching problem, and all have good rotational invariance, but these methods suffer from high computational complexity and low computational efficiency and do not support scale differences to different degrees.

C. Deep-Based Methods

With the rapid development of deep learning techniques, these methods have been widely applied to MRSI matching. Early methods, such as convolutional neural network (CNN)-based feature matching [36] and Superglue matching with graph neural networks (GNNs) [37], were used to improve image matching accuracy. These methods, however, still have limitations in matching efficiency and handling grayscale differences and modality variations. Deep learning-based MRSI matching methods have been extensively studied to address these issues. In the domain of CNNs, the D2-Net network [38] has been used for multisource image feature extraction and description, significantly improving matching performance. Additionally, XFeat [39] proposed extracting more robust keypoints using efficient CNNs, further optimizing feature learning and matching accuracy. In terms of Transformer architectures, a Transformer-based MRSI patch matching method [40] successfully applies a Transformer encoder architecture to improve MRSI patch matching accuracy. LoFTR [41], based on the Transformer framework, uses cross-self-attention and cross-attention mechanisms for feature extraction and similarity learning, significantly improving matching performance on weak-textured images. SE2-LoFTR [42] adds rotational invariance to LoFTR, further improving matching accuracy. Matchformer [43] proposed a framework for simultaneous feature extraction and similarity learning, optimizing the matching process. In the field of GNNs, CoAM [44] uses common attention modules and saliency scores to improve matching accuracy. LightGlue [45] builds on Superglue, enhancing matching speed through the use of self-attention and cross-attention mechanisms in GNNs. Efficient image matching methods based on GNNs [46] have improved matching efficiency, particularly in large-scale datasets. In dense matching methods, DKM [47] and RoMa [48] are two representative dense matching methods. While they are capable of matching a large number of keypoints, they require longer matching times and perform less effectively when applied to MRSI compared to sparse methods. Overall, these deep learning methods demonstrate strong feature learning capabilities and have shown great potential in MRSI matching. Large differences in ground features between multimodal images, along with the difficulty in obtaining training samples, however, limit the generalization ability and applicability of these methods.

III. Method

The proposed SOFT method is shown in Fig. 2, and the MRSI is preprocessed. Then, the maximum moment map is generated with the help of the PC model, and the Block-Harris detector [49] is used to complete the feature point extraction. The construction of the descriptor is the key part of this article, which can be subdivided into two steps: 1) a novel second-order tensor orientation function is designed to generate a tensor orientation feature map, which is used to describe the main direction information of the feature points and 2) the orientation feature map is combined with the improved GLOH-like descriptor framework to calculate the descriptor vector so that the descriptor has good rotation invariance. Next, the initial matching is done through a two-way matching strategy.

Then, EGIO was proposed. This method includes four steps: 1) automatically completing the global seed point calculation by designing the SMPI function; 2) partitioning and screening the candidate points for the seed points in 1); 3) iterating the

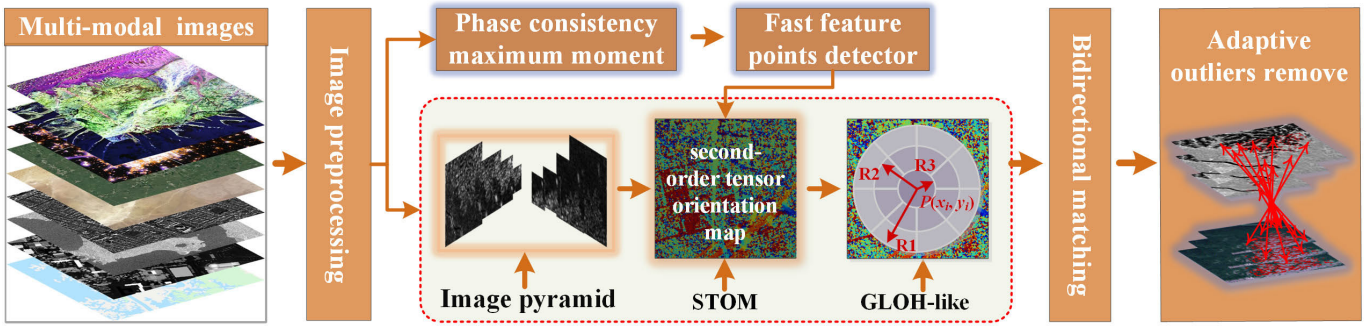

Fig. 2. Technical roadmap of the proposed method.

candidate points Optimization; and 4) global conversion model calculation.

A. Feature Point Extraction

First, a simple preprocessing is performed on the MRSI, which requires feature point detection. Considering the nonlinear distortion of MRSIs, differences in contrast, and differences in texture details, etc., it further increases the difficulty of identifying feature points. The PC model has a good ability to extract image structure features and edge features, and the maximum moment in the PC model can be convenient for extracting edge and corner features of the image. In this article, therefore, Block-Harris feature point extraction is performed on the maximum moment map, and nonmaximum value suppression is performed to retain significant feature points.

B. Second-Order Tensor Orientation Descriptor

The construction of a second-order tensor orientation descriptor is the key to achieve rotation-invariant multimodal feature matching. First, Gaussian image pyramid scale shadow diffusion is performed on the preprocessed image to obtain multiscale features. Then, the second-order tensor orientation feature of the image is constructed. Finally, a 204-D descriptor vector is obtained by using the directional feature map combined with the GLOH-like feature calculation framework, and the rotation invariant feature matching of MRSIs can be realized through this descriptor (see Fig. 2).

1) Second-Order Tensor Orientation Feature: The key to the second-order tensor orientation descriptor method is to build a second-order tensor orientation map, which calculates the main direction and statistical descriptor vector of the feature points through the feature map. The image gradients cannot be directly relied on in image registration owing to their high sensitivity to image distortions. In this section, the second-order gradient was first calculated, followed by computing the second-order gradient amplitude in the horizontal and vertical directions by using the improved Sobel template. Their equations are given in the following equation:

$$\nabla S_x = \begin{bmatrix} -1 & 0 & 1 \\ -\sqrt{5} & 0 & \sqrt{5} \\ -1 & 0 & 1 \end{bmatrix}, \quad \nabla S_y = \begin{bmatrix} -1 & -\sqrt{5} & -1 \\ 0 & 0 & 0 \\ 1 & \sqrt{5} & 1 \end{bmatrix}. \tag{1}$$

The first-order and second-order gradients of the image are computed by the two templates of (1), its mathematical expression is (1) and as follows:

$$\boldsymbol{G}(x,y)^1_\sigma = \sqrt{(\boldsymbol{L}(x,y)\cdot\sigma\cdot\nabla S_x)^2 + \left(\boldsymbol{L}(xy)\cdot\sigma\cdot\nabla S_y\right)^2} \tag{2}$$

$$\begin{cases} \boldsymbol{G}(x,y)^2_x = (\boldsymbol{G}(x,y)^1_\sigma \cdot \sigma \cdot \nabla S_x)^2 \\ \boldsymbol{G}(x,y)^2_y = (\boldsymbol{G}(x,y)^1_\sigma \cdot \sigma \cdot \nabla S_y)^2 \end{cases} \tag{3}$$

where $\boldsymbol{G}(x,y)^1_\sigma$ denotes the first-order gradient; $\boldsymbol{G}(x,y)^2_\sigma$ denotes the second-order gradient; $\boldsymbol{L}(x,y)$ represents the grayscale of the image and is the standard deviation of Gaussian distribution. ∇S_x and ∇S_y denote the Sobel template in the horizontal and vertical directions, respectively. The more detailed calculation is presented in the following equation:

$$W_\sigma = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{4}$$

The tensor provides the edge information in terms of its shape and direction. Despite its shape changes with contrast and illumination, the edge direction always remains unaltered. Consequently, the tensor model is frequently used in extracting the structural features of the image [50]. The definitive second-order tensor feature expression is given as follows:

$$\begin{bmatrix} \boldsymbol{T}_{xx} & \boldsymbol{T}_{xy} \\ \boldsymbol{T}_{yx} & \boldsymbol{T}_{yy} \end{bmatrix} = \begin{bmatrix} G_\sigma * P(x,y)_{xx} * W_\sigma & G_\sigma * P(x,y)_{xy} * W_\sigma \\ G_\sigma * P(x,y)_{yx} * W_\sigma & G_\sigma * P(x,y)_{yy} * W_\sigma \end{bmatrix} \tag{5}$$

where $P(x,y)_{xx}$ and $P(x,y)_{yy}$ represent the sum of squares of second-order gradients in the x- and y-directions, respectively. $P(x,y)_{xy}$ denotes trace of the second-order gradient and $*$ denote the dot product operation $P(x,y)_{yx} = P(x,y)_{xy}$, where G_σ is the Gaussian kernel function.

Finally, the complete second-order gradient tensor feature is calculated according to (5), as shown in the following equation:

$$G_{\mathrm{STOD}} = \frac{1}{2}\cdot\left[\arctan\left(T_{xy}+T_{yx}, T_{xx}-T_{yy}\right)+\pi\right] \tag{6}$$

where, G_{STOD} represents the final second-order tensor orientation feature map.

In order to further show the advantages of second-order tensor orientation descriptor, we use a set of synthetic aperture

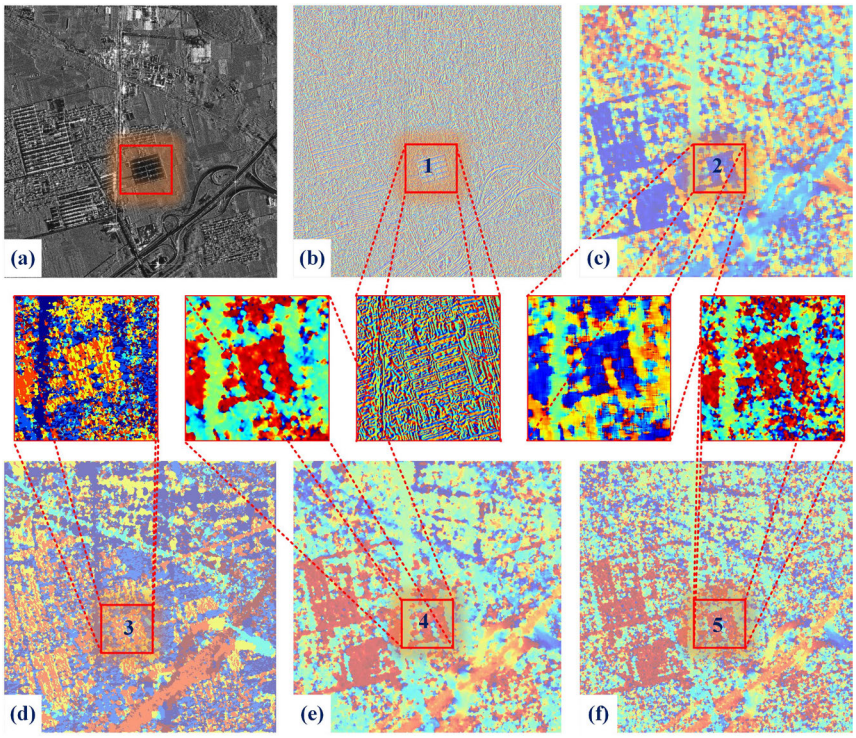

Fig. 3. Orientation maps of several methods. (a) Original image. (b) Orientation map of PSO-SIFT. (c) Orientation map of PIIFD. (d) Orientation map of RIFT. (e) Orientation map of MS-HMLO. (f) Orientation map of our SOFT.

radar (SAR) images and different methods to calculate its orientation feature map, as shown in Fig. 3. Fig. 3(b) is based on the second-order gradient of the image orientation feature map; Fig. 3(c) is the orientation feature map based on image grayscale; Fig. 3(d) is the maximum index orientation map based on image PC model; Fig. 3(e) is the orientation map based on the image average squared gradient; Fig. 3(f) is the second-order tensor orientation map proposed in this article.

Fig. 3 shows that the orientation map of the PSO-SIFT method contains too much detailed information and is susceptible to interference from noise. The characteristics of the orientation maps of the PIIFD and MS-HMLO methods are too coarse, and the detail information is filtered, which easily leads to inaccurate direction calculation; the orientation features of the RIFT method are indexed features calculated by multidimensional features, which have a limited degree of correct description of the main direction of the feature points. The orientation results of the proposed SOFT method, on the other hand, could better demonstrate the directional changes of the features, overcome the directional inversion, and provide a rotation invariant description.

2) Statistics of GLOH-Like Descriptor: After the SOFT calculation is completed, the feature vector of the descriptor needs to be calculated. Among them, one of the classical frameworks is the GLOH descriptor framework, and it has been successfully used in MRSI matching [51]. The division of the circular area has a great impact on accurate matching. Referring to the existing research and the GLOH-like descriptor [31], we propose an improved GLOH-like descriptor to count the feature vector of the image, as shown in Fig. 4. In order to further enhance the stability of the

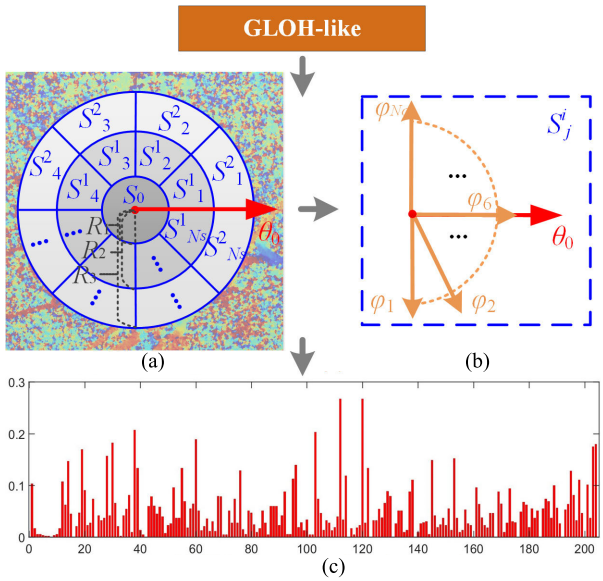

Fig. 4. GLOH-like descriptor template flowchart. (a) 17 subregions. (b) 12-D orientation. (c) 204-D feature vector.

GLOH description, the subregion is divided into more detailed regions.

Suppose S_0 denotes the central circular region; $S_j^i (i = 1, 2, j = 1, \ldots, N_S)$ denotes the sector subregion jth in the outer ring region ith, N_S denotes the number of subregions in each outer ring region, θ_0 denotes the principal direction of the feature points, and R_1, R_2, and R_3 denote the radii of the central and outer regions, respectively. In order to ensure the stability of the descriptor, the area of each subregion needs to be set consistently; thus, the relationship between R_1, R_2, and

R_3 can be defined by the following equation:

$$N_S \cdot \pi R_1^2 = \pi\left(R_2^2 - R_1^2\right) = \pi\left(R_3^2 - R_2^2\right). \tag{7}$$

On the second-order gradient of the image orientation feature map, the descriptor vector is calculated through the GLOH-like framework, where the direction values within $(-\pi/2, \pi/2)$ are quantized by No, as shown in Fig. 4(b), where $\varphi_k(k = 1, 2, \ldots, No)$ are quantized angles. A histogram of No in each region is computed.

For each key point, the second-order tensor orientation map value at its location is the main direction (reference direction θ_0). Then, all second-order tensor orientation map values in the GLOH-like local area are also based on θ_0 (0°), that is, all angle values minus θ_0, and the excess angle values $(-\pi/2, \pi/2)$ are flipped to their opposite angles. It is, however, unavoidable that in MRSIs, the image rotation and NRD may cause the main direction jump or direction mutation of some feature points near $-\pi/2$ and $\pi/2$. In response to this problem, PIIFD, Chen et al. [20] proposed corresponding improvement strategies. We use a similar strategy to deal with GLOH-like descriptors within the GLOH-like feature neighborhood. The features of the upper and lower parts are generated by adding and subtracting the main direction axis so as not to change the statistical order of the subregions, and the last one $204[(2 \times N_S + 1) \times o]$-dimensional feature vectors are generated.

C. Bidirectional Matching

After the feature descriptors being constructed, the initial matching of MRSI needs to be performed. In this article, Euclidean distance is used as the similarity measure for nearest neighbor matching. Images at each layer scale are matched. The matching results are then merged step by step. In order to ensure that image matching has a one-to-one correspondence, a bidirectional matching strategy is implemented. Finally, the feature point pairs after each layer of bidirectional matching are merged as the initial matching result.

D. Error-Removal Global Local Iterative Optimization

After the initial matching is completed, the matching points that still contain some errors need to be eliminated. The higher the rate of outliers, the more difficult it is to obtain inliers [52]. The commonly used RANSAC method [53] requires feature points to contain fewer outliers will work. Based on this, it is necessary to design an algorithm suitable for a high rate of outliers to extract the correct corresponding points. The fast sample consensus (FSCs) algorithm proposed by Wu et al. [54] could extract the correct corresponding points, but it is a must to artificially set a fixed initial threshold for the algorithm, and the algorithm cannot converge when there are few interior points, which limits the flexibility of the algorithm.

In order to solve the above problems, an EGIO method is proposed, which includes four steps: 1) designing the SMPI function and performing FSC calculation according to the function to obtain global seed points; 2) dividing image subregions to screen candidate points; 3) iterative optimization of candidate points; 4) global transformation model calculation, as shown in Fig. 5. Implementation Details and Pseudocode for EGIO is shown in Fig. 6.

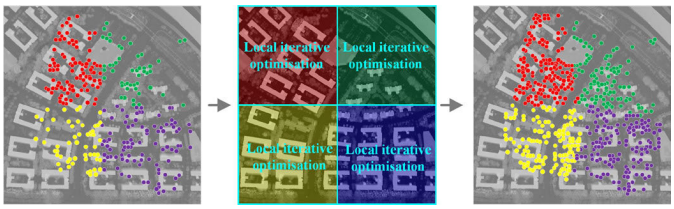

Fig. 5. Schematic of iterative optimization of partitioning. (a) Subregion of image. (b) Global seed point filtering. (c) Iterative optimization of the corresponding points in the subregion.

Algorithm 1 Error Removal Global Local Iterative Optimization

```
procedure EGIO(im1, im2, cor1, cor2, Thresh_max)
    dst1 ← grayscale(im1); dst2 ← grayscale(im2)
    mean_1 ← mean(dst1); median_1 ← median(dst1)
    std_1 ← std(dst1)
    A1 ←√|median_1 − mean_1|; B1 ←√|std_1 − mean_1|
    dim1 ← floor((A1 + B1) / 1)
    Repeat steps 3–5 for dst2 to get dim2
    SAPI ← (dim1 + dim2) / 4
    H ← calculate_affine_transform(cor1, cor2, SAPI)
    Y ← H × [cor1; 1], Normalize Y
    E ← distance(Y, cor2)
    seed_points ← { points | E ⩽ SAPI }
    refined_points ← local_refinement(seed_points, cor1, cor2,
    Thresh_max)
    if refined_points is empty then
        return "Insufficient points."
    end if
    transformation ← fit_transformation(refined_points)
    final_points ← remove_outliers(refined_points, transformation)
    return final_points, transformation
end procedure
```

Fig. 6. Implementation details and Pseudocode for EGIO.

Step 1 (Constructing SMPI Function and Extracting Seed Points): Inspired by the asymmetry measurement algorithm for the probability distribution of real-valued random variables [55], we try to use the information between MRSIs to automatically complete the algorithm's outlier filtering. Therefore, an SMPI function is designed. The function is applied to statistically match the distribution of the SMPI between the matching pairs. The mathematical expression of the SMPI function is shown in the following equation:

$$\begin{cases} \rho_1 = \sqrt{\dfrac{1}{M_1 \cdot N_1}\displaystyle\sum_{i=1}^{M_1}\sum_{j=1}^{N_1}(P_1(i,j) - \mu_1)} \\ \rho_2 = \sqrt{\dfrac{1}{M_2 \cdot N_2}\displaystyle\sum_{i=1}^{M_2}\sum_{j=1}^{N_2}(P_2(i,j) - \mu_2)} \\ \text{SMPI} = \left(\sqrt{|\text{ME}_1 - \mu_1|} + \sqrt{|\text{ME}_2 - \mu_2|} \right. \\ \qquad \left. + \sqrt{|\rho_1 - \mu_1|} + \sqrt{|\rho_2 - \mu_2|}\right) \cdot \Delta\Phi. \end{cases} \tag{8}$$

In (8), SMPI represents the SMPI of MRSI; P_1 and P_2 represent the averaged intensity of the pixels of the left and right images, respectively; ρ_1 and ρ_2 represent the standard

deviation of the left and right images, respectively; ME_1 and ME_2 represent the left image, respectively, and the median of the right image; $|\cdot|$ represents the absolute value symbol; $\Delta\Phi$ represents the weight coefficient (set to 1/4 in this article).

The SMPI is used as the initial threshold of FSC for initial outlier filtering, and the seed points of the left and right images are obtained. The initial matching points for the left and right images are noted as: $\boldsymbol{P}_{\text{match1}}$, and $\boldsymbol{P}_{\text{match2}}$.

Step 2 (Dividing Image Subregions to Filter Candidate Points): First, the global subregion is evenly divided into four subregions (k), where $k \in [1, 2, 3, 4]$. Then, the number of seed points falling into each subarea is counted. When the number of points is greater than 4, calculate the local perspective transformation model ($\boldsymbol{H}$) of each subarea; when the number of points is less than 4, the subarea is judged to be empty and returns 0.

Then, the $\boldsymbol{H}$ matrix of each subregion is used to calculate the residual of the left image matching point and the corresponding right image matching point falling in each subregion, and its mathematical expression is shown in the following equation:

$$\begin{cases} \Delta\boldsymbol{\varepsilon}^k = \left\| \boldsymbol{P}^k_{\text{match1}} * \boldsymbol{H}^k - \boldsymbol{P}^k_{\text{match2}} \right\| \\ \Delta\boldsymbol{\varepsilon}'^k = f_{\text{sort}\uparrow}\left(\Delta\boldsymbol{\varepsilon}^k\right) \end{cases} \tag{9}$$

where in (9), $\Delta\boldsymbol{\varepsilon}^k$ represents the residual set of the left image matching point and the right image matching point in the kth subregion; $\boldsymbol{P}^k_{\text{match1}}$ and $\boldsymbol{P}^k_{\text{match2}}$ represent the left image matching point and the right image matching point falling in the kth subregion, respectively; $\boldsymbol{H}^k$ represents the perspective transformation matrix of the kth subregion; $*$ represents matrix multiplication; $f_{\text{sort}\uparrow}(\cdot)$ represents the sort function; $\Delta\boldsymbol{\varepsilon}'^k$ represents the residual set after sorting the left image and right image matching points in the kth subregion.

The standard deviation of the residual of the matching point is calculated according to (9), and compare it with the maximum pixel residual threshold (F_T). The expression is shown in the following equation:

$$\begin{cases} \Delta\tilde{\varphi}^k = \min\left(f_{\text{std}}\left(\Delta\boldsymbol{\varepsilon}'^k\right), F_T\right) \\ \boldsymbol{P}^k_{c1} = \boldsymbol{P}^k_{\text{match1}}\left\|_{\Delta\boldsymbol{\varepsilon}'^k < \Delta\tilde{\varphi}^k} \boldsymbol{P}^k_{c2} = \boldsymbol{P}^k_{\text{match2}}\right\|_{\Delta\boldsymbol{\varepsilon}'^k < \Delta\tilde{\varphi}^k} \end{cases} \tag{10}$$

where (10), $\Delta\tilde{\varphi}^k$ represents the standard deviation of the kth subregion; $f_{\text{std}}(\cdot)$ represents the standard deviation calculation function; F_T represents the maximum pixel residual threshold (this article is set to 6); $\boldsymbol{P}^k_{c1}$ and $\boldsymbol{P}^k_{c2}$ represent the left image candidate point and the right image candidate point of the kth subregion, respectively.

Step 3 (Candidate Point Iterative Optimization): The fitting points corresponding to $\boldsymbol{P}_{c2}$ are denoted by $\boldsymbol{P}'^1_{c2} = \boldsymbol{P}^1_{c1} \cdot \boldsymbol{H}^1$. The residuals of $\boldsymbol{P}'^1_{c2}$ and $\boldsymbol{P}^1_{c2}$ are denoted in $\Delta\Gamma(i) = \|\boldsymbol{P}'_{c2}(i) - \boldsymbol{P}_{c2}(i)\|_2$, $i = 1, \ldots, N_c$. N_c is the number of correct matches (NCMs). In accordance with a number of experimental observations, a larger size fitting point is more likely to be a point close to the correct position. Therefore, the points are sorted in ascending order, and the last one-fourth of the elements are found. Then, the points corresponding to the last element in $\boldsymbol{P}_{c2}$ are replaced by the fitting points. All the points are updated in P_{c2} until the sum of the residuals is equal to the threshold (0.01). The position of P_{c2} is updated.

Step 4 (Global Transformation Model Calculations): The candidate point optimization in each subarea is completed through Step 3). Then, the optimized points of the four subregions are used as the final matching corresponding points. Finally, by performing the least squares calculation again, the refined global transformation modulus is obtained.

E. Complexity and Efficiency Analysis

The computational complexity of the proposed matching algorithm involves four key stages: feature detection, feature description, feature matching, and outlier filtering (EGIO). Feature detection uses the Harris corner detector with a complexity of $O(n)$, where n is the number of pixels in the image. This provides robustness but incurs a high computational cost. Feature description uses our proposed descriptor with a complexity of $O(m * l)$, where m is the number of keypoints and l is the number of pyramid layers, making this step a major computational component. Feature matching employs bidirectional matching with a complexity of $O(m * n)$, which is optimized through parallel processing. The EGIO outlier filtering step includes global seed point calculation $O(n)$, subregion filtering $O(k\text{^}2)$, local optimization $O(m * \log(m)$, and transformation fitting $O(p\text{^}3)$. The number of iterations is limited to 5 to balance efficiency and accuracy. Feature description and outlier filtering are the primary computational bottlenecks, but parallel processing provides significant improvements for handling large-scale datasets.

IV. Experiments

Eight state-of-the-art traditional methods, i.e., PIIFD [19], PSO-SIFT [20], ASS [24], RIFT2 [26], RIFT [25], MS-HLMO [31], OSS [22], HOSS [22], and one deep learning-based method, RoMa [48], were used for comparison. During the tests, the image scale difference was set to 1.6 and 56 pixels concerning the neighborhood window. The maximum pixel FT was set to 6. The parameters of the compared methods were adjusted to the optimal stage accordingly. The proposed SOFT method, PIIFD, PSO-SIFT, ASS, RIFT2, RIFT, MS-HLMO, OSS, and HOSS were implemented in MATLAB-R2018a. The RoMa method was implemented using Python. In those methods, the number of key points was kept under 3000. The experiments were performed on a Lenovo-R9000K2021H laptop with an Intel[1] Core[2] i7-5900HX CPU, 32GB RAM, and Windows 11 × 64 operating system. The image-space affine transformation was used to model the geometric relationships of image pairs. For each pair, over 15 well-distributed ground truth points were manually collected to calculate the affine transformation as the ground truth, which is used to measure the location accuracy of the automatically matched points.

The SR is used as a measure of the probability of successful image matching. The NCMs represent the number of corresponding points in the reference and the sensing images. Rate of correct matches (RCMs): it is a rate that reflects the NCMs

[1]Registered trademark.
[2]Trademarked.

TABLE I
INTRODUCTION TO THE MRSI DATASET

Type	Num	Size	Source
Multi-temporal	293	512x512	Google Map
Infrared-optical	228	512x512	Landsat, Google Map
Depth-optical	48	512x512	Depth Map
Map-optical	111	512x512	Google Map
SAR-optical	63	512x512	Sentinel-1, Gaofen-3
Night-Day	2	500x500	NASA

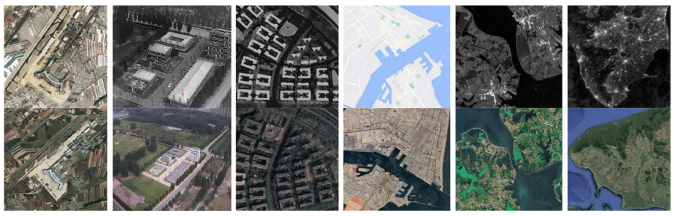

Fig. 7. Partial multi-MRSIs. (a) Multitemporal images. (b) Infrared-optical. (c) Depth-optical. (d) Map-optical. (e) SAR-optical. (f) Night-day.

to the total number of matches. The root of the mean-squared error (RMSE) of the correct matches

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}\left[\left(x_i - x_i'\right) + \left(y_i - y_i'\right)\right]\right)} \tag{11}$$

where, N represents the number of ground truth points; (x_i', y_i') is the coordinate of the ith ground truth point converted by matching the correspondence. (x_i, y_i) is the coordinate of the ith predicted point.

A. Data Sources

The experimental dataset consisting of 744 sets of MRSI was collected. These images include multitemporal images, infrared-optical, LiDAR deep image-optical, map-optical, SAR-optical and night-day, and the majority of them are 512 × 512 pixels. These data cover spaceborne, airborne and ground remote sensing data. Detailed information can be found in Table I. Each type of data contains images with large rotation differences. At the same time, some images also have scale differences, as shown in Fig. 7. In order to better verify the performance of the proposed method, we performed random rotations ranging from 0° to 180° at 30° intervals on 744 sets of test images to enhance sample randomness.

B. Results of Matching

In order to evaluate the matching accuracy of the proposed SOFT method, the SOFT method with nine state-of-the-art methods (PIIFD, PSO-SIFT, ASS, RIFT2, RIFT, MS-HLMO, OSS, HOSS, and RoMa) are compared. Fig. 8 shows the quantitative evaluation results of ten methods in three metrics. Table II shows the average results of the ten methods in the four metrics.

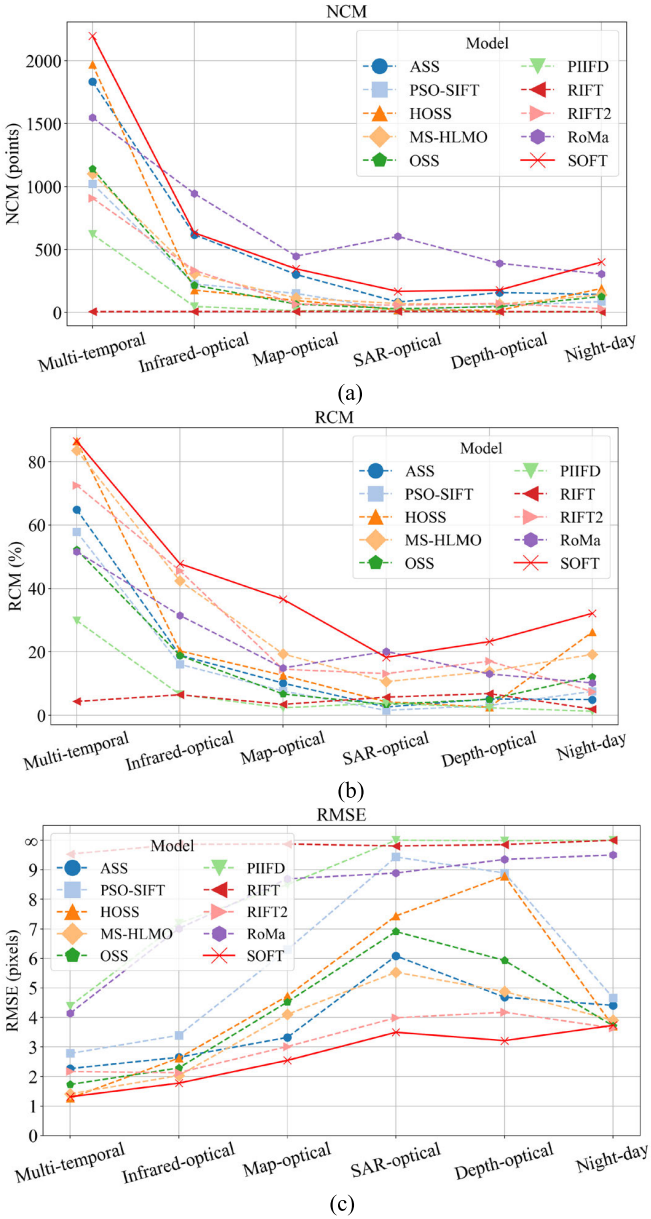

Fig. 8. Results in three metrics for SOFT and other nine methods. (a) NCM results. (b) RCM results. (c) RMSE results.

The NCM and RCM of the matching failure results are set to 0. The RMSE is obtained by ten algorithms (matching failure or RMSE greater than 10 pixels are set to $+\infty$).

1) Quantitative Evaluation: From Fig. 8 shows that the unit of SR is %; the unit of NCM is points; the unit of RCM is %; the unit of RMSE is pixel. Quantitative results as shown in Fig. 8. The results of all three metrics are average values.

The light green dashed lines from Fig. 8(a)–(c) represent the NCM, RCM, and RMSE results of the PIIFD method, respectively. In the 744 sets of images, the SR is only 48.25%, the NCM is 262.52, the RCM is only 14.60%, and the RMSE is as high as 6.70 pixels, which shows that this method is not suitable for multimodal matching. The light blue dashed lines in Fig. 8(a)–(c) represent the NCM, RCM, and RMSE results of the PSO-SIFT method, respectively. Its SR result is better than that of the PIIFD method, but only 73.66%

TABLE II
RESULTS IN FOUR METRICS OF THE SOFT AND THE COMPARED METHODS

	PIIFD	PSO-SIFT	ASS	RIFT2	RIFT	MS-HLMO	OSS	HOSS	RoMa	SOFT
SR	48.25	73.66	95.43	94.49	3.90	95.56	91.00	82.26	41.40	100
NCM	262.52	507.04	974.90	480.14	7.70	556.46	531.88	848.40	1042.74	1138.35
RCM	14.60	29.71	33.43	46.94	5.13	50.62	27.99	42.76	34.76	57.28
RMSE	6.70	4.45	3.02	2.57	9.73	2.58	3.03	3.21	6.45	1.95

(see Table II), and the obtained NCM and RCM are not high; the NCM and RCM are 507.04 point and 29.71%, respectively (see Table II). The RMSE of the ground truth computation is 4.45 pixels. It can be seen that although the PSO-SIFT method uses the second-order Sobel operator to calculate the new image gradient feature, it still has a certain effect on the brightness difference of the image, this method is still sensitive to NRD.

The red dashed line from Fig. 8(a)–(c) represents the matching results of the RIFT method. This method is designed with maximum index map descriptors based on the PC modal, which improves the matching performance of the algorithm. It achieves a better SR result of 3.90%. Its NCM, RCM, and RMSE are 7.70, 5.13%, and 9.73 pixels, respectively (see Table II). The main reason for the poor results of the RIFT method on the three metrics may be the lack of a rotation module in the published RIFT code by the authors. The pink dashed line represents the results of the RIFT2 method. From Fig. 8(a)–(c), we found that it achieves 94.49% SR results, and its NCM, RCM, and RMSE results are 480.14, 46.94%, and 2.57 pixels, respectively (see Table II). Among them, the RCM and RMSE metrics showed high, but the NCM performance is poor.

The orange dashed line represents the results of the HOSS method. From Fig. 8(a)–(c), we found that it achieves 82.26% SR results, and its NCM, RCM, and RMSE results are 848.40, 42.76%, and 3.21 pixels, respectively (see Table II). The HOSS method, however, has fluctuating results and unstable matching in the three metrics. The green dashed line from Fig. 8(a)–(c) represents the result of the OSS method. It achieves MRSI matching by constructing an orientation self-similar feature descriptor. The NCM is 531.88, the RCM is 27.99%, and the RMSE result is 3.03 pixels. The overall effect of the OSS method is better than PIIFD, PSO-SIFT, RIFT, and HOSS but inferior to RIFT2.

The blue dashed line represents the results of the ASS method. From Fig. 8(a)–(c), we found that it achieves 95.43% SR results, and its NCM, RCM, and RMSE results are 974.90, 33.43%, and 3.02 pixels, respectively (see Table II). As an optimized version of the OSS method, the ASS method is slightly better than the OSS method in SR, RCM, and RMSE, and the improvement in NCM is more significant. The light yellow dashed line represents the results of the MS-HLMO method. From Fig. 8(a)–(c), we found that it achieves 95.56% SR results, and its NCM, RCM, and RMSE results are 556.46, 50.62%, and 2.58 pixels, respectively (see Table II). The MS-HLMO method gives better results than the other seven conventional methods, but it performs poorly and less efficiently in terms of scale differences and large modal differences (SAR-optical) in MRSI. The purple dashed line represents the results of the RoMa method (deep learning method). From Fig. 8(a)–(c), we found that it achieves 41.40% SR results, and its NCM, RCM, and RMSE results are 1042.74, 34.76%, and 6.45 pixels, respectively (see Table II). It is worth noting that the RoMa method is capable of obtaining a large number of corresponding points in successfully matched images, but it is poorly adapted to modality, and the lack of a multimodal image dataset is the main reason for this problem.

The red solid line from Fig. 8(a)–(c) represents the result of the proposed SOFT method, which is successfully matched in 744 sets of images, where the NCM is 1138.35, the RCM is 57.28%, and the RMSE is 1.95 pixels, of which the result is significantly better than the other nine methods.

2) Qualitative Match Results: In order to further evaluate the performance of the SOFT method, we tested the qualitative matching results by selecting a set of images from each of the six MRSIs types, as shown in Fig. 9. And the matching results of the SOFT method in the remaining images are shown in Fig. 10.

Fig. 9(a) and (b) shows the partial matching results of the PIIFD and PSO-SIFTs respectively. They are based on image gradient features for matching, so they are more sensitive to the NRD of MRSI and prone to matching failure. Fig. 9(d) and (e) shows the partial matching results of the RIFT2 and RIFT methods. They use a PC model to complete the feature description, where the matching performance of their maximum index map descriptor is more stable; however, they do not support scale differences, and their performance is also limited under rotational differences. Fig. 9(c), (g), and (h) shows the partial matching results of the ASS, OSS, and HOSS methods, we found that they achieve successful matching among most image types, but there are still some limitations of the method. Fig. 9(d) shows the partial matching results for the MS-HLMO method, which was found to perform better for rotational disparity but poorly for larger NRD and scale disparity. Fig. 9(i) shows the partial matching results for the RoMa method, which can obtain a large number of correspondence points in the successfully matching images, but performs poorly in terms of large NRD differences (e.g., map-optical and night-day).

Fig. 9(j) shows the partial matching results of the method proposed in this article. It can be seen that the SOFT method demonstrates robust matching capabilities under conditions of NRD, lighting variations, and contrast differences. Moreover, it performs well in handling scale variations and displacement

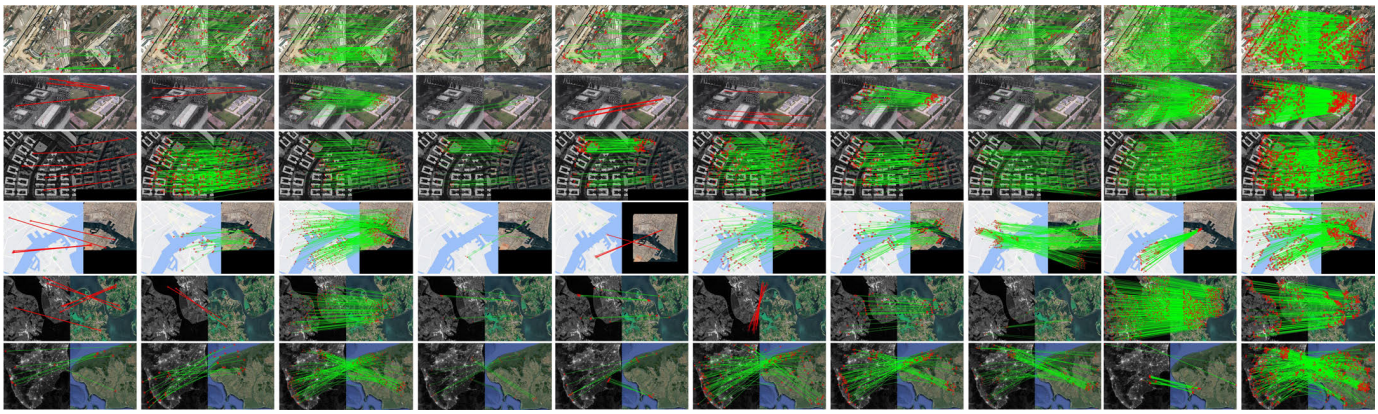

Fig. 9. Matching results of SOFT and the other nine methods. (a) PIIFD. (b) PSO-SIFT. (c) ASS. (d) RIFT2. (e) RIFT. (f) MS-HLMO. (g) OSS. (h) HOSS. (i) RoMa. (j) SOFT.

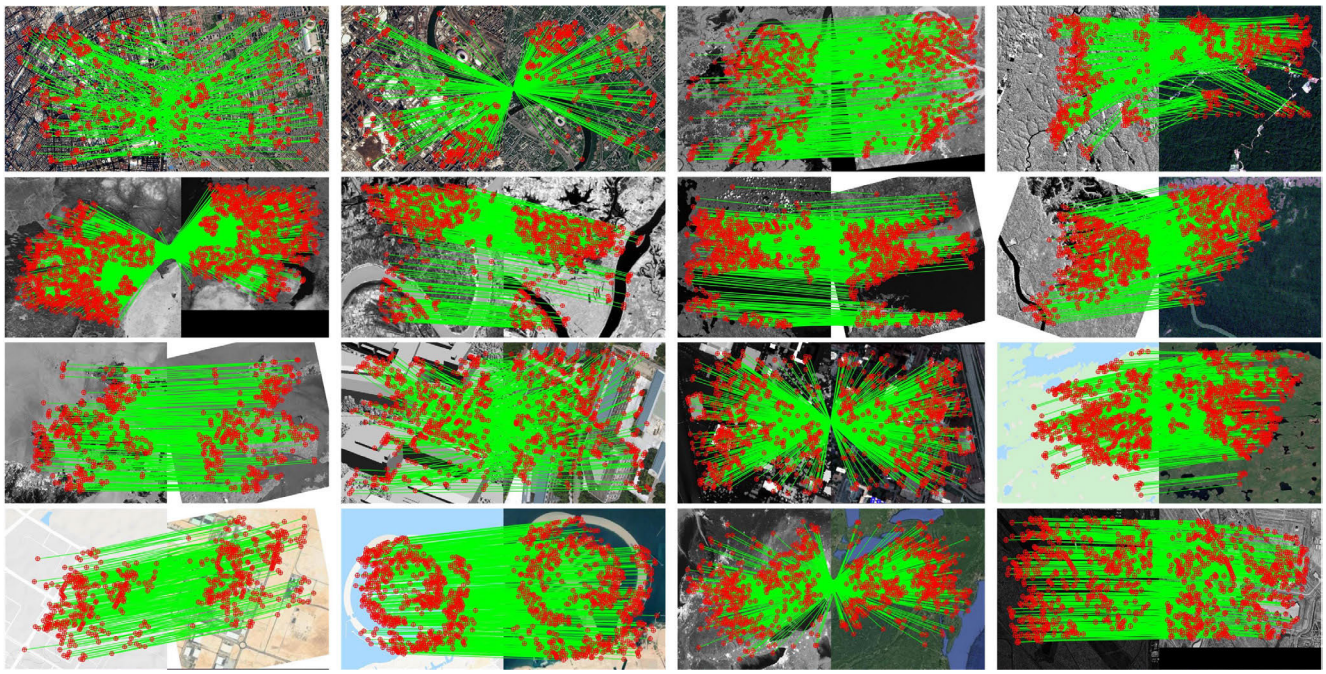

Fig. 10. Part matching results of SOFT method.

differences, highlighting its strong applicability. In order to further demonstrate the matching effect of the SOFT method, the matching results of 16 groups of typical images are also shown (see Fig. 10). It can be seen from Fig. 10 that the proposed SOFT method has good stability against MRSI, which can be obtained enough NCM in translation, scale, and rotation difference.

C. Rotational Invariance Test

In order to verify the rotation invariance of the SOFT method, we selected 22 typical image pairs for matching testing, rotating every 45°, as shown in Fig. 11. Fig. 11 shows that MRSI images can be successfully matched under different rotation angles. The NCM of each group of image pairs is higher than 50 points, and most of the RCM are above 20%. At the same time, it is not difficult to find that the performance of the SOFT method does not decrease with the change of the rotation angle. The quantitative results are shown in Fig. 12.

TABLE III
PARAMETER SETTINGS OF THE PROPOSED SOFT MODEL

Experiment	Variable	Fixed Parameters
Parameter N_W	N_W = [24,32,40,48,56,64,72,80,88,96,104]	$F_T = 6$
Parameter F_T	F_T = [2,3,4,5,6,7,8,9,10]	$N_W = 56$

D. Discussion

In order to comprehensively evaluate the matching robustness of the SOFT method, three parts of the SOFT method, the parameter setting, the construction of second-order tensor-orientation feature descriptors and EGIO are further analyzed.

1) Analysis of Parameter Settings: The proposed SOFT method has parameters, such as Nw and F_T, whose different values will affect its matching performance. Therefore, we quantitatively analyzed how the SOFT method functioned under different settings. More details of parameter settings are given in Table III.

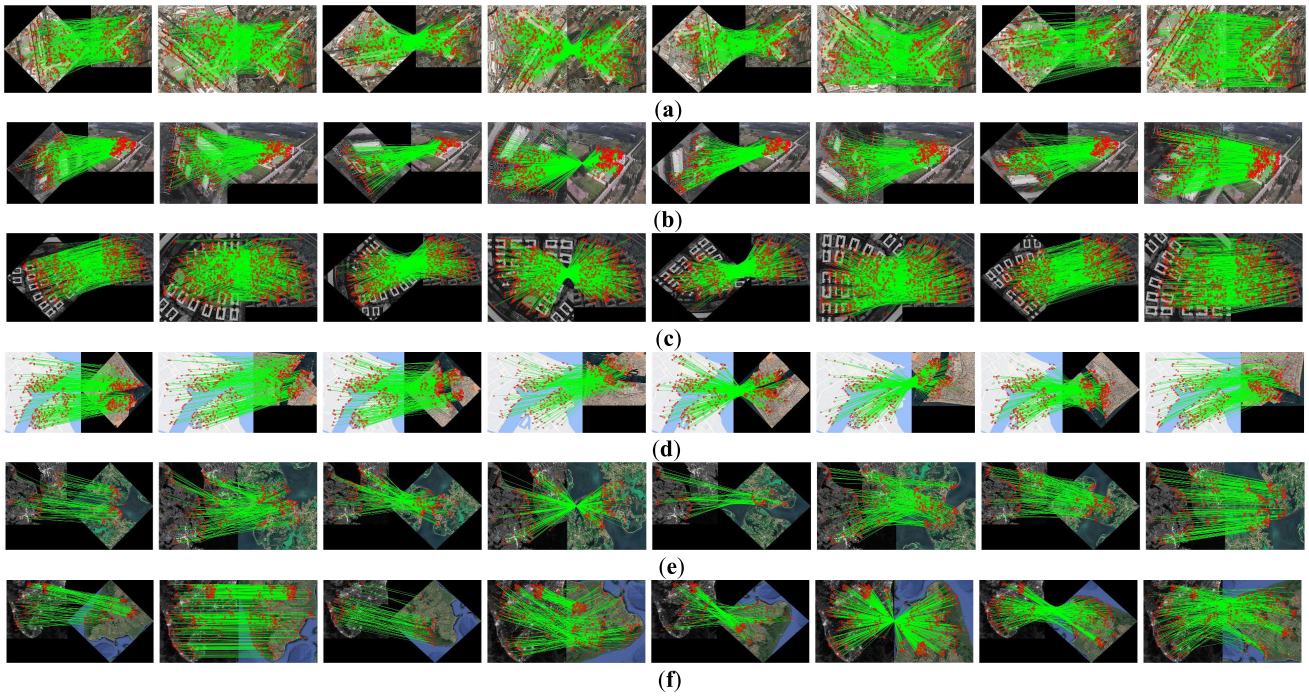

Fig. 11. Matching results for the SOFT method in rotational transformation. (a) Multitemporal images. (b) Infrared-optical. (c) Depth-optical. (d) Map-optical. (e) SAR-optical. (f) Night-day.

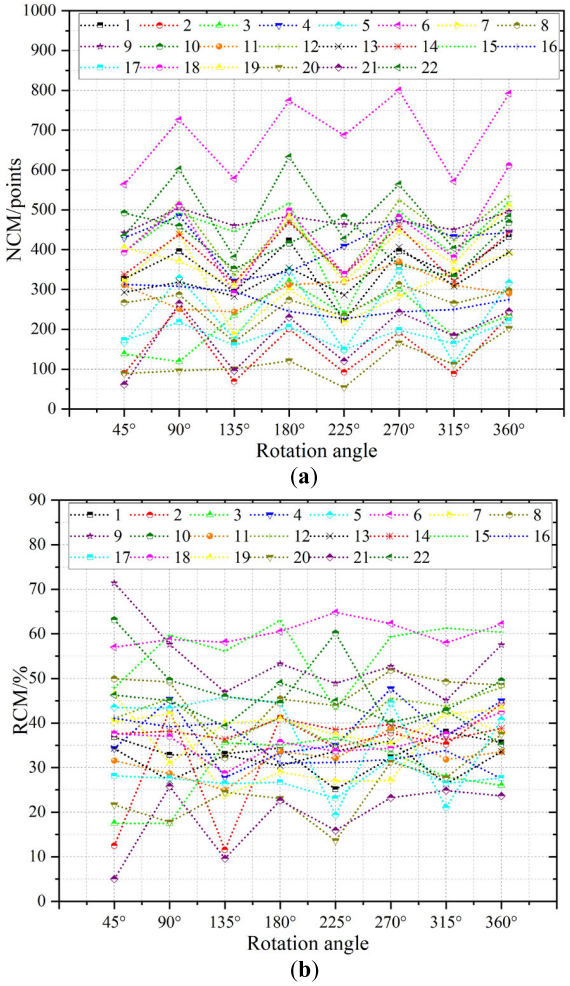

Fig. 12. Matching results of the SOFT method for different rotation angles. (a) NCM results of SOFT. (b) RCM results of SOFT.

We tested 744 sets of MRSIs according to the parameter settings given in Table III to evaluate the impact of different parameters on the SOFT method by observing the average NCMs and average RMSEs of images, which are shown in Figs. 13 and 14.

Figs. 13 and 14 show the effect of different settings of the two parameters on the SOFT method. As can be seen from Fig. 13, the SOFT method results in a gradual increase in NCM and a gradual decrease in RMSE results for $24 < N_W < 56$. When $N_W = 56$, optimal results were obtained for both NCM and RMSE. Subsequently, the NCM results start to decrease gradually and the RMSE results increase gradually; thus, an N_W setting of 56 presents optimal results. As can be seen from Fig. 14, the RMSE tends to decrease and then increase with the increase of F_T. When $F_T = 6$, the RMSE of the SOFT method reaches the best result. Therefore, setting F_T to 6 is optimal.

2) Descriptor Analysis of Nine Methods: The proposed descriptors for the construction of second-order tensor-orientation features play an important role in this article. In order to compare the performance of the descriptors more fairly, the image feature point extraction, matching methods and coarse difference rejection methods are all used in the same way, where the outlier filtering method module of the SOFT method uses the FSC method, denoted as SOFT-FSC. Three metrics, NCM, RCM, and RMSE, are used to quantitatively evaluate the results of several descriptors, as shown in Fig. 15.

As Fig. 15 shows, when other matching links remain unchanged, the proposed descriptor in this article can still achieve overall best results. Among them, the SOFT-FSC method achieved successful matching in all sets of MRSI,

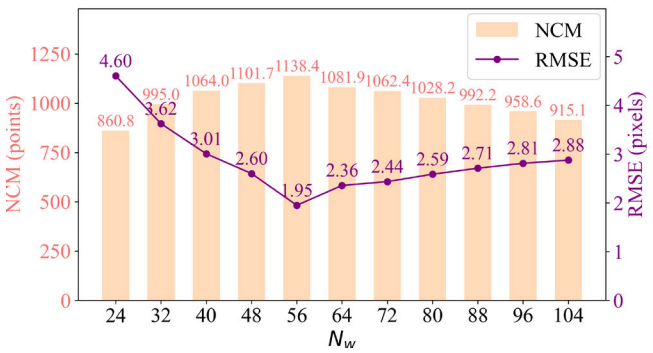

Fig. 13. Matching results of different N_w values.

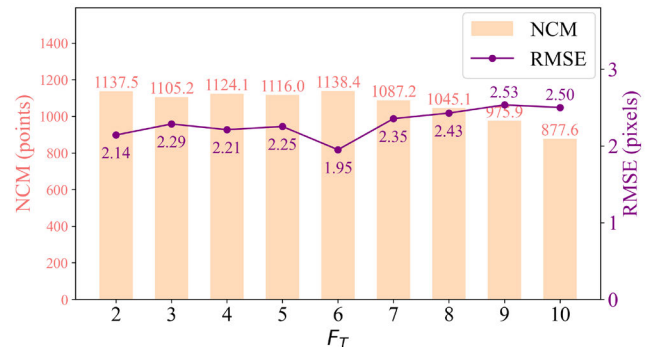

Fig. 14. Matching results of different F_T values.

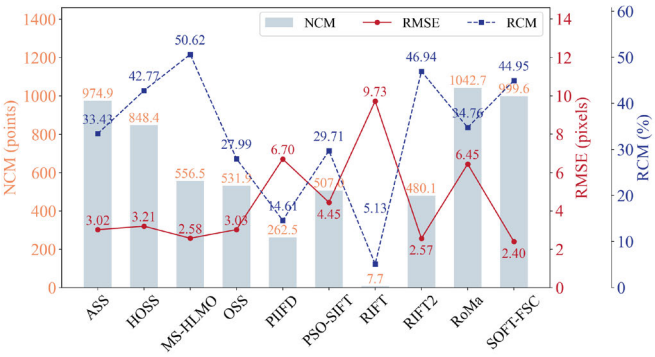

Fig. 15. Quantitative comparison results of descriptors for different methods.

its NCM is 999.6, the RCM is 44.95%, and the RMSE is 2.40 pixels. The RMSE of the SOFT-FSC method is at least 0.63 pixels lower than the OSS method and 4.30 pixels lower than the PIIFD method. Compared with MS-HLMO and RIFT2 with lower RMSE, the NCM of SOFT-FSC is 1.8 and 2.08 times higher than theirs, respectively. It can be seen that the descriptors in the SOFT-FSC method have significant effects on MRSI matching.

3) Analysis of Three Outlier Filtering Methods: The proposed descriptors for the construction of second-order tensor-orientation features play an important role in this article. In order to compare our EGIO method, more fairly with FSC and RANSAC, image feature point extraction, feature description, and matching methods are all used in the same way, where the outlier filtering method module of the SOFT method uses the RANSAC method, denoted as SOFT-RANSAC. Three

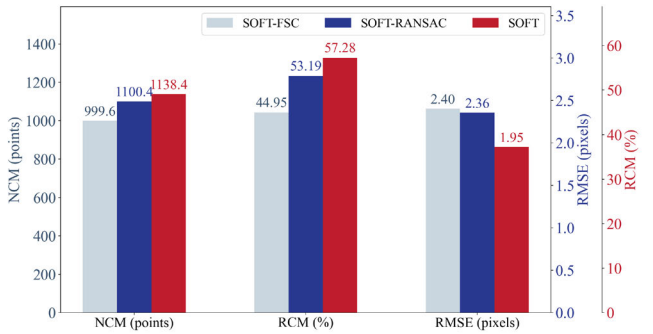

Fig. 16. Quantitative comparison results of different outlier filtering methods.

metrics, NCM, RCM, and RMSE, are used to quantitatively evaluate, as shown in Fig. 16. As Fig. 16 shows, when other matching links remain unchanged, the proposed EGIO method achieves the best results across all three metrics—NCM, RCM, and RMSE. Compared to FSC and RANSAC, our EGIO method achieved a 13.89% and 3.45% improvement in NCM, a 27.43% and 7.69% improvement in RCM, and a reduction of 0.45 and 0.41 in RMSE, respectively. It can be seen that the EGIO method have significant effects in MRSI matching.

V. Conclusion

In this article, an MRSI matching method based on second-order tensor-orientation feature descriptors is proposed to improve the rotational invariance of MRSIs and to overcome the problems of NRD, noise interference, feature direction reversal and abrupt changes of MRSI. Comprehensive experiments on 744 sets of MRSIs demonstrate that the proposed SOFT method can achieve robust matching of MRSI, by effectively guaranteeing rotational invariance and obtaining high correct matching rates, which can be concluded as follows.

1) *Enhanced Matching Performance:* The proposed SOFT method exhibits superior matching performance, with a 97% improvement in NCM, a 25.51% improvement in RCM, and a reduction in RMSE by 2.69 pixels compared to the other nine methods.
2) *Robust Rotation Invariance:* The SOFT method demonstrates strong rotation invariance, achieving robust matching results under arbitrary angular rotations in the rotation simulation experiments.
3) *Effective Error-Removal Optimization:* The introduction of an EGIO significantly enhances the correct matching rate by better filtering and optimizing corresponding points.

In summary, the proposed SOFT method offers strong applicability in MRSI matching, providing robust support for applications such as aerial triangulation, image stitching, 3-D reconstruction, and SLAM. Its capabilities extend to diverse fields, including geospatial analysis for aerial and satellite imagery, enhanced SLAM for autonomous driving, and improved image registration in medical diagnostics.

Future improvements focus on enhancing computational efficiency via GPU acceleration for faster, real-time performance. Integrating deep learning could further improve accuracy and robustness, especially in complex scenarios.

These advancements will help solidify the SOFT method as a versatile and efficient solution for MRSI matching in challenging environments.

REFERENCES

[1] 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.

[2] Y. Zhang, Z. Zhang, and J. Gong, "Generalized photogrammetry of spaceborne, airborne and terrestrial multi-source remote sensing datasets," *Acta Geodaetica et Cartographica Sinica*, vol. 50, no. 1, pp. 1–11, 2021.

[3] 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, Sep. 2021.

[4] X. Zhang, C. Leng, Y. Hong, Z. Pei, I. Cheng, and A. Basu, "Multimodal remote sensing image registration methods and advancements: A survey," *Remote Sens.*, vol. 13, no. 24, p. 5128, Dec. 2021.

[5] D. N. Joanes and C. A. Gill, "Comparing measures of sample skewness and kurtosis," *J. Royal Stat. Soc. D, Stat.*, vol. 47, no. 1, pp. 183–189, 1998, doi: 10.1111/1467-9884.00122.

[6] S. Ji, C. Zeng, Y. Zhang, and Y. Duan, "An evaluation of conventional and deep learning-based image-matching methods on diverse datasets," *Photogramm. Rec.*, vol. 38, no. 182, pp. 137–159, Jun. 2023.

[7] S. Bas and A. O. Ok, "A new productive framework for point-based matching of oblique aircraft and UAV-based images," *Photogramm. Rec.*, vol. 36, no. 175, pp. 252–284, Sep. 2021.

[8] Y. Hel-Or, H. Hel-Or, and E. David, "Matching by tone mapping: Photometric invariant template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 317–330, Feb. 2014.

[9] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[10] Y. Hel-Or, H. Hel-Or, and E. David, "Fast template matching in non-linear tone-mapped images," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1355–1362.

[11] A. A. Goshtasby, *Image Registration: Principles, Tools and Methods*. Cham, Switzerland: Springer, 2012.

[12] X. Yan, Y. Zhang, D. Zhang, N. Hou, and B. Zhang, "Registration of multimodal remote sensing images using transfer optimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2060–2064, Dec. 2020.

[13] X. Xu, X. Li, X. Liu, H. Shen, and Q. Shi, "Multimodal registration of remotely sensed images based on Jeffrey's divergence," *ISPRS J. Photogramm. Remote Sens.*, vol. 122, pp. 97–115, Dec. 2016.

[14] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu, "Robust optical-to-SAR image matching based on shape properties," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 564–568, Apr. 2017.

[15] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.

[16] Y. Ye, L. Bruzzone, and J. Shan, "A fast and robust matching framework for multimodal remote sensing image registration," 2018, *arXiv:1808.06194*.

[17] 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, Jun. 2022.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[19] H.-J. Chien, C.-C. Chuang, C.-Y. Chen, and R. Klette, "When to use what feature? SIFT, SURF, ORB, or A-KAZE features for monocular visual odometry," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2016, pp. 1–6.

[20] J. Chen et al., "A partial intensity invariant feature descriptor for multimodal retinal image registration," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1707–1718, Feb. 2010.

[21] W. Ma et al., "Remote sensing image registration with modified SIFT and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.

[22] A. Sedaghat and N. Mohammadi, "Illumination-robust remote sensing image matching based on oriented self-similarity," *ISPRS J. Photogramm. Remote Sens.*, vol. 153, pp. 21–35, Jul. 2019.

[23] X. Xiong, G. Jin, Q. Xu, and H. Zhang, "Self-similarity features for multimodal remote sensing image matching," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12440–12454, 2021.

[24] X. Xiong, G. Jin, Q. Xu, H. Zhang, L. Wang, and K. Wu, "Robust registration algorithm for optical and SAR images based on adjacent self-similarity feature," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, doi: 10.1109/TGRS.2022.3197357.

[25] 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.

[26] J. Li, P. Shi, Q. Hu, and Y. Zhang, "RIFT2: Speeding-up RIFT with a new rotation-invariance technique," 2023, *arXiv:2303.00319*.

[27] Y. Yao, Y. Zhang, and Y. Wan, "Heterologous images matching considering anisotropic weighted moment and absolute phase orientation," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 46, no. 11, pp. 1727–1736, 2021.

[28] W. Yang, C. Xu, L. Mei, Y. Yao, and C. Liu, "LPSO: Multi-source image matching considering the description of local phase sharpness orientation," *IEEE Photon. J.*, vol. 14, no. 1, pp. 1–9, Feb. 2022.

[29] Y. Yao, B. Zhang, Y. Wan, and Y. Zhang, "MOTIF: Multi-orientation tensor index feature descriptor for SAR-optical image registration," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 99–105, May 2022.

[30] N. Mohammadi, A. Sedaghat, and M. Jodeiri Rad, "Rotation-invariant self-similarity descriptor for multi-temporal remote sensing image registration," *Photogramm. Rec.*, vol. 37, no. 177, pp. 6–34, Mar. 2022.

[31] 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, doi: 10.1109/TGRS.2022.3193109.

[32] W. Yang, Y. Yao, Y. Zhang, and Y. Wan, "Weak texture remote sensing image matching based on hybrid domain features and adaptive description method," *Photogramm. Rec.*, vol. 38, no. 184, pp. 537–562, Dec. 2023.

[33] J. Li, W. Xu, P. Shi, Y. Zhang, and Q. Hu, "LNIFT: Locally normalized image for rotation invariant multimodal feature matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2022.3165940.

[34] Y. Liao et al., "Refining multi-modal remote sensing image matching with repetitive feature optimization," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 134, Nov. 2024, Art. no. 104186, doi: 10.1016/j.jag.2024.104186.

[35] Y. Liao, P. Tao, Q. Chen, L. Wang, and T. Ke, "Highly adaptive multimodal image matching based on tuning-free filtering and enhanced sketch features," *Inf. Fusion*, vol. 112, Dec. 2024, Art. no. 102599, doi: 10.1016/j.inffus.2024.102599.

[36] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Computer Vision—ECCV 2016* (Lecture Notes in Computer Science), vol. 9910, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 467–483, doi: 10.1007/978-3-319-46466-4_28.

[37] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4937–4946.

[38] M. Dusmanu et al., "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8092–8101.

[39] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "XFeat: Accelerated features for lightweight image matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024.

[40] A. Moreshet and Y. Keller, "Attention-based multimodal image matching," 2021, *arXiv:2103.11247*.

[41] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8918–8927.

[42] G. Bokman and F. Kahl, "A case for using rotation invariant features in state of the art feature matchers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 5106–5115.

[43] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "MatchFormer: Interleaving attention in transformers for feature matching," 2022, *arXiv:2203.09645*.

[44] O. Wiles, S. Ehrhardt, and A. Zisserman, "Co-attention for conditioned image matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15915–15924.

[45] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023.

[46] H. Chen et al., "Learning to match features with seeded graph matching network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6281–6290.

[47] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, "DKM: Dense kernelized feature matching for geometry estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.
[48] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "RoMa: Robust dense feature matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024.
[49] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 23.1–23.6.
[50] U. Köthe, "Edge and junction detection with an improved structure tensor," in *Proc. Joint Pattern Recognit. Symp.* Berlin, Germany: Springer, 2003, pp. 25–32.
[51] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Aug. 2005.
[52] J. Li, Q. Hu, M. Ai, and S. Wang, "A geometric estimation technique based on adaptive M-estimators: Algorithm and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5613–5626, 2021.
[53] M. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[54] Y. Wu, W. Ma, M. Gong, L. Su, and L. Jiao, "A novel point-matching algorithm based on fast sample consensus for image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 43–47, Jan. 2015.
[55] D. P. Doane and L. E. Seward, "Measuring skewness: A forgotten statistic?" *J. Statist. Educ.*, vol. 19, no. 2, Jul. 2011, doi: 10.1080/10691898.2011.11889611.

Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photogrammetry from Wuhan University (WHU), Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently a Professor and the Dean of the School of Remote Sensing and Information Engineering, WHU. 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 datasets, object information extraction and modeling with artificial intelligence, integration of LiDAR point clouds and images, and 3-D city model reconstruction.

Dr. Zhang is the Coeditor-in-Chief of *The Photogrammetric Record*.

Peihao Wu received the B.Eng. degree from China Agricultural University, Beijing, China, in 2024. He is currently pursuing the M.Eng. degree with Wuhan University (WHU), Wuhan, China.

His main research interests include computer vision and multimodal image matching.

Yongxiang Yao received the Ph.D. degree in geodesy and photogrammetry from Wuhan University (WHU), Wuhan, China, in 2023.

He is currently a Laboratory Technician with the School of Remote Sensing and Information Engineering, WHU. He has published more than 30 research articles. His research interests include image matching, multimodal image fusion, and visual image relocation.

Yi Wan (Member, IEEE) received the B.S. and Ph.D. degrees from Wuhan University (WHU), Wuhan, China, in 2013 and 2018, respectively.

He is currently an Associate Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include photogrammetry, computer vision, 3-D reconstruction, satellite image interpretation, and change detection.

Wenfei Zhang received the B.Eng. degree from the Central South University, Changsha, China, in 2022. He is currently pursuing the Ph.D. degree with Wuhan University (WHU), Wuhan, China.

His main research interests include the registration and fusion of SAR and optical imagery.

Yansheng Li (Senior Member, IEEE) received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

He is currently a Full Professor and the Vice Dean of the School of Remote Sensing and Information Engineering, Wuhan University (WHU), Wuhan, China. He has authored more than 100 peer-reviewed papers such as IEEE TPAMI, CVPR, ECCV and AAAI. His research interests include knowledge graph, deep learning and their applications in remote sensing big data mining.

Dr. Li is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and a Junior Editorial Member of *The Innovation*.

Xiaohu Yan received the B.S. degree from Huazhong Agricultural University, Wuhan, China, in 2008, the M.S. degree from North China Electric Power University, Beijing, China, in 2010, and the Ph.D. degree from Wuhan University (WHU), Wuhan, in 2017.

He is an Associate Professor with the School of Undergraduate Education, Shenzhen Polytechnic University, Shenzhen, Guangdong, China. His research interests include computer vision, artificial intelligence, and optimization algorithm.