International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02023-9

Robust Heterogeneous Model Fitting for Multi-source Image Correspondences

Shuyuan Lin[1] · Feiran Huang[1] · Taotao Lai[2] · Jianhuang Lai[3] · Hanzi Wang[4] · Jian Weng[1]

Received: 2 April 2023 / Accepted: 27 January 2024

Abstract
Traditional feature detection and description methods, such as scale-invariant feature transform, are susceptible to nonlinear radiation distortions (NRDs) and geometric distortions (GDs), which in turn generate a large number of outliers or incorrect correspondences. To address this issue, this paper proposes a simple yet effective heterogeneous model fitting (MIMF) for multi-source image correspondences. First, a multi-orientation phase consistency model is constructed, which fuses phase consistency, image amplitude and orientation to detect the correct correspondences of feature points. This model effectively reduces the influence of NRDs. Second, sub-region grids and orientation histograms are exploited to construct the log-polar descriptors with variable-size bins, which are robust to GDs. Finally, a heterogeneous model fitting method is proposed, which can effectively estimate the parameters of the transformation model for alleviating the influence of outliers. Experiments are performed on six public datasets and one constructed dataset containing ten types of multi-source images, and the experimental results show that the proposed MIMF method outperforms several state-of-the-art competing methods in terms of matching performance.

Keywords Model fitting · Heterogeneous model · Multi-source data · Image correspondence · Geometric matching

1 Introduction

Multi-source image correspondence (i.e., image-matching) refers to the process of establishing correspondence between two or more images with overlapping regions captured by different time phases, viewing angles, or different modal sensors (Jiang et al., 2021). As a fundamental and challenging task, multi-source image correspondence can provide supplementary information for remote sensing and geospatial observations, and it has been widely used in computer vision-related applications such as image stitching (Ma et al., 2019), image fusion (Ma et al., 2021), land cover analysis (Hu et al., 2023), and scene matching guidance (Jin et al., 2021). However, multi-source image pairs may contain nonlinear radiation distortions (NRDs) and geometric distortions (GDs) accompanied by scale, rotation, noise, blur, or temporal variations (Li et al., 2017), significantly decreasing the accuracy and speed of geometric correspondences. Although image correspondence technology has made great progress over the past decades, it still cannot meet the requirements

Communicated by Paolo Rota.

✉ Hanzi Wang
hanzi.wang@xmu.edu.cn

✉ Jian Weng
cryptjweng@gmail.com

Shuyuan Lin
swin.shuyuan.lin@gmail.com

Feiran Huang
fuangfr@jnu.edu.cn

Taotao Lai
laitaotao@gmail.com

Jianhuang Lai
stsljh@mail.sysu.edu.cn

[1] College of Cyber Security/College of Information Science and Technology, Jinan University, Guangzhou 510632, Guangdong, China

[2] College of Computer and Data Science, Minjiang University, Fuzhou 350108, Fujian, China

[3] School of Computer Science and Engineering, Sun Yat-sen University; Guangdong Key Laboratory of Information Security Technology; Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006, Guangdong, China

[4] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, Fujian, China

Published online: 23 February 2024

of practical applications in multi-source image registrations. Therefore, it is necessary to develop efficient, general, and robust image correspondence methods.

With the diversity of sensors and applications, different sensors such as the optical, infrared, light detection and ranging (LiDAR), and synthetic aperture radar (SAR) have significantly different imaging mechanisms (i.e., image pairs of the same object have different representations), which leads to a large difference in the nonlinear intensity between image pairs, especially NRDs (Li et al., 2019). Meanwhile, remote sensing images are susceptible to the influence of atmospheric effects and lighting conditions, resulting in the phenomenon of "different objects with the same spectrum" or "the same object with different spectra", which may reduce the accuracy of image correspondence and make image registration difficult. Traditional image-matching methods usually exploit intensity or gradient information to realize feature detection and description, but they are very sensitive to NRDs. Therefore, the difficulty in multi-source image correspondence is how to effectively avoid the influence caused by NRDs and GDs.

Several recent studies (Ye et al., 2017, 2019; Zhang et al., 2023) have shown that the structural texture and shape properties of images are not easily corrupted by different modal sensors, which contributes to the extraction of similar features from multi-source images. However, the processing of the structural texture and shape in the scale space is limited by Gaussian blur constraints, and the extracted features contain noise and weakening parts of the texture information (Yao et al., 2022). Generally, the structural texture features of an image can be represented by gradient information, but it is sensitive to the radiation changes in an image. In contrast, feature representations based on phase consistency have been proven to be robust to both illumination and contrast changes, and insensitive to radiative changes (Kovesi, 1999) (see Fig. 1). However, traditional phase consistency models can only obtain the amplitude of an image, so they fail to describe the complex structural features of an image (Ye et al., 2017).

In practice, sensor or measurement errors are inevitable during data acquisition and pre-processing, and multi-source data usually contain outliers, resulting in incorrect initial correspondences (Lin et al., 2021; Lai et al., 2023). Traditional feature-based methods directly use the least squares algorithm or the random sample consensus (RANSAC) algorithm to estimate the parameters of a basic transformation model and then remove outliers. However, the accuracy of these methods varies with the proportion of outliers, and the inlier noise scale and the maximum number of random samples of these methods need to be set manually.

In contrast, robust model fitting aims to estimate the parameters of a transformed model from data contaminated by outliers and then distinguish inliers from outliers (Lin et al., 2019). In robust model fitting, the geometric information in an image pair can be represented by a transformation model, and then a "hypothesis-verification" framework is employed to estimate the transformation model (Lin et al., 2023). Specifically, this framework mainly involves two steps: (1) model hypothesis generation, where minimal subsets are sampled from data to generate model hypotheses; (2) model selection, where the generated model hypotheses are validated, and the model hypotheses hitting the real model instances are selected. Additionally, traditional model fitting methods usually evaluate image correspondences in terms of a single basic transformation model, which has some limitations. For example, similarity transformation models suffer from scale size ambiguity; affine transformation models do not have sufficient ability to describe shapes; perspective transformation models may incorrectly merge feature information from small projection regions. Therefore, it is significant to investigate how to integrate the advantages of different types of basic transformation models, further reduce the influence of the limitations of a single basic transformation model, and finally improve the performance of model fitting in image correspondence tasks.

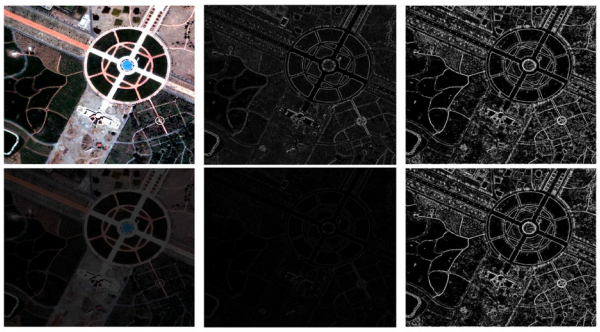

Fig. 1 Comparison of gradient with phase consistency. The 1^{st} row shows the original images, and the 2^{nd} row shows the images with illumination changes

Though some progress has been made on multi-source image correspondence, they still have the following defects: (1) there are significant intensity and NRDs differences between multi-source images, resulting in an insufficient number of feature points due to poorly detecting structural textures and shape information; (2) traditional descriptors cannot describe feature points well, leading to sparse or even failed detection of initial correspondences; (3) directly using the least squares or RANSAC to estimate the parameters of a transformation model is susceptible to outliers, leading to inaccurately estimated parameters and false correspondences. Therefore, it is an important task to preserve or enhance structural textures and shape information to reduce the sensitivity to image gradients and further integrate the

advantages of heterogeneous models to establish the robust correspondence of multi-source images.

This paper starts from feature detection and description to effectively describe common features of multi-source images by avoiding the influence of illumination/intensity differences, NRDs, and GDs. Then, a robust heterogeneous model fitting method (called MIMF) is constructed to optimize multi-source image correspondences for accurate feature matching.

The key contributions of this paper are summarized as follows.

- A multi-orientation phase consistency model is constructed, which combines the phase consistency, amplitude, and orientation information of an image. The model can not only preserve the structural texture and shape information of an image but also enhance the reliability of feature detection.
- A variable-size bin strategy is proposed to quantify the location and orientation of a descriptor structure, which in turn improves the discriminative ability of the descriptor against local geometric distortions for establishing high-quality initial correspondences.
- A heterogeneous model fitting method is proposed to estimate the parameters of transformation models, which integrates the advantages of multiple types of transformation models to efficiently alleviate the influence of outliers for rejecting false correspondences.

Besides, this paper constructs a representative multi-source dataset of real images, which contains ten different types of modalities and covers various remote and indoor application scenarios. These image pairs are derived from different types of sensors with different image pixels, rotation orientations, intensities, textures, nonlinear radiation distortions, etc.

The rest of this paper is organized as follows. Section 2 gives a review of the related work. Section 3 provides the details of the methodology. Section 4 presents and discusses the experimental results on representative datasets. Section 5 concludes this paper.

2 Related Work

In this section, some representative works on multi-source image correspondence are briefly reviewed, including region-intensity-based methods, feature-based methods, deep-learning-based methods, and outlier-removal-technique-based methods.

2.1 Region-Intensity-Based Methods

Region-intensity-based methods perform image-matching by computing the similarity of image intensities in the spatial or frequency domain (Le Moigne et al., 2002; Zeng et al., 2020). For instance, the normalized correlation coefficient (NCC) is used to measure the similarity of images by normalizing the correlation coefficient to address linear intensity variations (Uss et al., 2016). However, NCC cannot accurately handle multi-source images with complex intensity variations. Mutual information (MI) introduces information theory to statistically evaluate the intensity dependence between images, thus effectively addressing nonlinear intensity differences of images (Ma et al., 2010). However, MI ignores the spatial information of neighboring pixels in an image and is computationally inefficient. Phase correlation exploits the Fourier shift theorem to quickly estimate translations and scale changes between images (Reddy and Chatterji, 1996), and it is widely used in remote sensing image registration. For instance, HOPC (Ye et al., 2017) exploits intensity and orientation corresponding to phase information instead of gradient information to construct descriptors, but its sparse sampling grid makes it difficult to capture the structural information of images. AWOG (Fan et al., 2021) deploys gradient values to correlation directions and uses 3D phase correlation as a similarity measure to improve matching results. However, the intensity information and spatial similarity measures used by the phase correlation still cannot effectively solve the registration problem of multi-source images with significant orientation and intensity differences. Besides, since the gray level and gradient information of multi-source images usually have large differences, region-intensity-based methods may lose some similarity features.

2.2 Feature-Based Methods

Feature-based methods first extract the features from an image pair and then match them based on their similarity. These methods rely on the features extracted from the reference and target images. For example, the classical scale-invariant feature transform (SIFT) (Lowe, 2004) and its variants (e.g., SURF (Bay et al., 2006) and ORB (Rublee etal., 2011)) are robust to scale/rotation/linear intensity variations, but they are very sensitive to nonlinear intensity differences. Some improved variants of SIFT (e.g., uniformly robust SIFT (Sedaghat et al., 2011) and scale-constrained SURF (Teke and Temizel, 2010)) address this issue by improving local features. However, features extracted from multi-source images usually have low local reproducibility due to large differences in intensity and texture (Kelman et al., 2007). Thus, these methods are only applicable to specific types of images, so their application scope is limited. For multi-source

image matching, PSO-SIFT (Ma et al., 2016) improves the descriptor structure of SIFT and introduces an enhanced matching strategy to increase the number of correct matching points. OS-SIFT (Xiang et al., 2018) adopts multi-scale Sobel operators to construct robust descriptors, thus improving the robustness of SIFT against radiative distortions. RIFT (Li et al., 2019) combines the phase consistency and the maximum index map to resist nonlinear radiation difference and image rotation, but it cannot handle scale variations. HAPCG (Yao et al., 2021) utilizes anisotropic weighted moment maps to construct a histogram of absolute phase orientation gradients for feature description. 3MRS (Fan et al., 2022) combines a coarse-to-fine two-stage feature description strategy and a 3D phase correlation strategy to perform multi-source image matching. Although these methods have good resistance to nonlinear radiation differences, they are usually designed for a specific application.

In contrast, the proposed method combines the phase consistency and the amplitude/orientation information of images from a multi-orientation perspective, and it can handle the structural texture and shape information of multi-source images well. Besides, the proposed method alleviates the influence of local geometric distortions through descriptor structure quantization.

2.3 Deep-Learning-Based Methods

Deep-learning-based methods (Litjens et al., 2017) perform image-matching by training the targeted model on a large number of annotated samples. For example, PSO-SIFT-A (Ye et al., 2018) summarizes the defects of SIFT for mid- and high-level information and avoids them by fusing a deep convolutional neural network (CNN) and SIFT for remote sensing image registration. The modified CycleGAN (Fuentes Reyes et al., 2019) generates SAR-like patches from optical images by pre-training a conditional generative adversarial network (cGAN) and registers them with artificially generated patches. IVAT (Zhang et al., 2019) eliminates the differences between image pairs by joining a deep image analogy and the depth semantics of images and then registering the generated images using local features. SuperPoint (DeTone et al., 2018) employs a self-supervised learning method to extract feature points and compute descriptors by training a full convolutional neural network. However, the robustness of its interest point detection is unstable when dealing with interference factors such as occlusion and noise. SuperGlue (Sarlin et al., 2020) utilizes an end-to-end learning method based on graph neural networks to handle challenges such as occlusion and illumination changes in images, thus enhancing the robustness of feature matching. It is worth noting that the graph neural network used by SuperGlue may increase computational cost when processing large-scale data and high-resolution images. LoFTR (Sun et al., 2021) introduces transformer-based self-attention and mutual-attention layers to obtain feature descriptors of image pairs, but it is difficult to guarantee the matching results using only end-to-end network structures. Though deep-learning-based methods have a strong feature learning ability, their generalization ability and applicability are limited due to the large object differences in multi-source images and the difficulty in obtaining training samples.

2.4 Outlier-Removal-Technique-Based Methods

Recently, robust outlier removal techniques have been proposed to enhance the accuracy of correspondences in the fine registration stage. For instance, LLT (Ma et al., 2015) removes outliers from putative matches and estimates rigid and nonrigid transformation models via local linear transformations. RLSS (Xiong et al., 2019) combines the detected features in frequency and spatial domains to improve the accuracy of registration. OSIR (Paul and Pati, 2020) introduces a bootstrap matching strategy to deal with most of the outliers in the detected features. CFOG (Ye et al., 2019) utilizes a feature descriptor based on channel features of oriented gradients to register the corners of multi-source images and improves its computational efficiency by performing a fast Fourier transform. LMR (Ma et al., 2019) treats the outlier removal problem as a binary classification problem and learns a generic classifier to determine correct correspondences. MTOPKRP (Jiang et al., 2019) employs the multi-scale top K-rank preservation based on local topological relations to achieve robust feature matching. CBG (Lin et al., 2022) treats the outlier removal problem as a bipartite graph partitioning problem and learns a generic bicluster to determine correct correspondences.

Though the above methods have achieved promising results, they have limitations in the applicability and accuracy of multi-source image correspondence. For example, to extract the geometric relationships between images, the transformation relationships in image correspondences usually involve rigid transformation models, similar transformation models, affine transformation models, and projection transformation models. Traditional outlier removal techniques usually specify one of these models and then use RANSAC to find the features satisfying the maximum consistent set rule as candidate models and perform registration. In practice, there are different types of transformation models, and it is difficult to reflect the geometric relationships between images with only one basic transformation relationship (model). Moreover, due to sensor or measurement errors, multi-source visual data can be contaminated with outliers during acquisition and pre-processing, and an example is the wrong correspondence of the same object on two views from different viewpoints.

In this paper, a novel heterogeneous model fitting method (i.e., MIMF) is proposed for multi-source image correspondence. Compared with the previous methods, the proposed MIMF can perform image-matching accurately, attributed to the advantages of model fitting technology and fused multiple transformation models.

3 Methodology

In this section, a simple but robust heterogeneous model fitting method (i.e., MIMF) is proposed for multi-source image correspondences. First, a phase-coherence-based feature space is constructed, and then a multi-orientation phase consistency feature detection model is developed (see Sect. 3.1). Next, the feature distribution of multi-source images in a polar coordinate grid is analyzed, and an improved variable-size bin log-polar descriptor is proposed (see Sect. 3.2). Finally, a heterogeneous model fitting method is provided for removing outliers in multi-source images (see Sect. 3.3).

3.1 Multi-orientation Feature Detection

Given a reference image $I(x, y)$ and a target image $I^{'}(x, y)$, the image correspondence is to find an optimal geometric transformation model by minimizing the distance (or maximizing the similarity) of feature information between the image pairs:

$$\hat{f}(x, y) = \arg\min_{f(x,y)} \left[\Phi \left(I(x, y), I^{'}(f(x, y)) \right) \right], \tag{1}$$

where $f(x, y)$, $I^{'}(f(x, y))$, and Φ are the geometric transformation model, the transformed target image, and the distance metric, respectively. The optimal geometric transformation model $\hat{f}(x, y)$ refers to the geometric transformation model that can minimize the feature information distance. For instance, feature points between two images can be used to estimate an affine transformation matrix (also known as an affine transformation model, which is a common geometric transformation model). If there are enough feature points to support the affine transformation model (i.e., minimizing the distance between feature points and the model), the model is considered an optimal geometric transformation model. However, due to the characteristics of sensors and illumination variations, multi-source images (especially remote sensing images) often contain radiation, rotation, and noise, which lead to NRDs. Therefore, it is crucial to accurately extract and describe the significant features of image pairs and estimate the relationship between them. However, traditional feature extraction methods (such as SIFT or SURF) usually rely on intensity or gradient information in the spatial domain, which makes them sensitive to NRDs and difficult to detect correct features. In contrast, extracting features in the frequency domain (e.g., phase information) can effectively avoid this problem (Li et al., 2019). Consequently, to enhance the robustness of feature extraction methods to illumination, a phase consistency ($\mathcal{PC}$) instead of an intensity or gradient histogram is constructed in this paper.

Specifically, considering the log-Gabor response of an image $I(x, y)$, a two-dimensional log-Gabor filter ($\mathcal{G}$) can be computed by a polarity separable Gaussian function as follows:

$$\mathcal{G}_{(\sigma,\mu)}(\rho, \delta) = \exp\left(\frac{(\rho - \rho_\sigma)^2}{-2\mathcal{B}_\rho^2} \right) \cdot \exp\left(\frac{\left(\delta - \delta_{(\sigma,\mu)}\right)^2}{-2\mathcal{B}_\delta^2} \right), \tag{2}$$

where $\mathcal{B}_\rho$ and $\mathcal{B}_\delta$ indicate the bandwidths in the log-polar coordinates ρ and δ, respectively; the subscripts σ and μ correspond to the scale and orientation of $\mathcal{G}$; $(\rho_\sigma, \delta_{(\sigma,\mu)})$ is the center frequency of $\mathcal{G}$, respectively. Then, the inverse Fourier transform can be used to convert $\mathcal{G}$ from the frequency domain to the spatial domain:

$$\mathcal{G}_{(\sigma,\mu)}(x, y) = \mathcal{G}^{eve}_{(\sigma,\mu)}(\sigma, \mu) + i \cdot \mathcal{G}^{odd}_{(\sigma,\mu)}(\sigma, \mu), \tag{3}$$

where $\mathcal{G}^{eve}_{(\sigma,\mu)}(\sigma, \mu)$ and $\mathcal{G}^{odd}_{(\sigma,\mu)}(\sigma, \mu)$ represent the even-symmetric (i.e., the real part) and the odd-symmetric (i.e., the imaginary part) of log-Gabor wavelets, respectively. Next, the amplitude components $\mathbb{A}_{(\sigma,\mu)}(x, y)$ and phase components $\mathbb{P}_{(\sigma,\mu)}(x, y)$ with respect to the scale σ and the orientation μ can be calculated:

$$\begin{cases} \mathbb{A}_{(\sigma,\mu)}(x, y) = \left(\mathcal{E}_{(\sigma,\mu)}(x, y)^2 + \mathcal{O}_{(\sigma,\mu)}(x, y)^2 \right)^{0.5}, \\ \mathbb{P}_{(\sigma,\mu)}(x, y) = \arctan\left(\mathcal{O}_{(\sigma,\mu)}(x, y) / \mathcal{E}_{(\sigma,\mu)}(x, y) \right), \end{cases} \tag{4}$$

where $\mathcal{E}_{(\sigma,\mu)}(x, y) = I(x, y) * \mathcal{G}^{eve}_{(\sigma,\mu)}(x, y)$ and $\mathcal{O}_{(\sigma,\mu)} = I(x, y) * \mathcal{G}^{odd}_{(\sigma,\mu)}(x, y)'$ represent the log-Gabor responses obtained by convolution operations in the specific scale σ and orientation μ. Finally, $\mathcal{PC}$ with respect to multiple scales and orientations (Kovesi, 2000) can be formulated as follows:

$$\mathcal{PC}(x, y) = \frac{\sum_\sigma \sum_\mu \omega_\sigma(x, y) \left\lfloor \mathbb{A}_{(\sigma,\mu)}(x, y) \Delta\mathbb{P}_{(\sigma,\mu)}(x, y) - \gamma \right\rfloor}{\sum_\sigma \sum_\mu \mathbb{A}_{(\sigma,\mu)}(x, y) + \epsilon}, \tag{5}$$

where $\omega_\sigma(\cdot)$ indicates a weighting factor; $\lfloor \cdot \rfloor$ indicates a truncation function that yields a zero (non-zero) value when it is negative (positive); $\Delta\mathbb{P}_{(\sigma,\mu)}(\cdot)$ denotes the phase deviation with respect to the scale σ and orientation μ; γ and ϵ indicate a noise compensation and a small value to constrain division by zero, respectively.

Although using $\mathcal{PC}$ instead of intensity or gradient histograms is robust to NRDs, it mainly contains structural texture and shape information, which may lead to noise sensitivity. Therefore, to further enhance the relationship between $\mathcal{PC}$ and orientation and to fully exploit the structural features in images, this paper proposes to construct a multi-orientation weighted moment map for representing feature information to overcome this limitation. Specifically, an independent $\mathcal{PC}$ map for each orientation at each scale is analyzed, and then the minimum and maximum moments of the $\mathcal{PC}$ maps with different orientations are calculated as follows (Kovesi, 2003):

$$\begin{cases} \mathbb{M}_\sigma = \frac{1}{2}\Big(\sum\limits_\mu (\beta_\sigma)^2 + \sum\limits_\mu (\alpha_\sigma)^2 + \\ \quad \sqrt{\Big(2\sum\limits_\mu (\alpha_\sigma)(\beta_\sigma)\Big)^2 + \Big(\sum\limits_\mu (\alpha_\sigma)^2 - \sum\limits_\mu (\beta_\sigma)^2\Big)^2}\Big), \\ \mathcal{M}_\sigma = \frac{1}{2}\Big(\sum\limits_\mu (\beta_\sigma)^2 + \sum\limits_\mu (\alpha_\sigma)^2 - \\ \quad \sqrt{\Big(2\sum\limits_\mu (\alpha_\sigma)(\beta_\sigma)\Big)^2 + \Big(\sum\limits_\mu (\alpha_\sigma)^2 - \sum\limits_\mu (\beta_\sigma)^2\Big)^2}\Big), \end{cases} \tag{6}$$

where $\mathbb{M}_\sigma$ and $\mathcal{M}_\sigma$ represent the maximum moment and the minimum moment corresponding to the scale σ, respectively; $\alpha_\sigma = \mathcal{PC}\left(\phi_{(\sigma,\mu)}\right)\cdot\cos\left(\phi_{(\sigma,\mu)}\right)$ and $\beta_\sigma = \mathcal{PC}\left(\phi_{(\sigma,\mu)}\right)\cdot\sin\left(\phi_{(\sigma,\mu)}\right)$; $\phi_{(\sigma,\mu)}$ indicates the angle of orientation μ at the scale σ. According to Eq. (6), if the maximum moment $\mathbb{M}_\sigma$ of a feature point is high, the point has a high probability of representing an edge feature. Similarly, if the minimum moment $\mathcal{M}_\psi$ is high, the point may represent a corner feature. Therefore, the minimum moment $\mathcal{M}_\sigma$ and the maximum moment $\mathbb{M}_\sigma$ can be exploited to indicate the corner features and edge features of an image, respectively. Finally, the weighted moments can be calculated as follows:

$$\mathbb{W}_\sigma = 0.5 \times (\mathbb{M}_\sigma + \mathcal{M}_\sigma + \varpi \times (\mathbb{M}_\sigma - \mathcal{M}_\sigma)), \tag{7}$$

where ϖ represents the weight coefficient. Based on this, feature points are extracted from $\mathbb{W}_\sigma$ by using the Shi-Tomasi operator (Shi, 1994) (an improved Harris operator with strong anti-noise ability), and then the feature points with lower response values are filtered out to avoid feature duplication.

3.2 Variable-Size Bin Descriptor Construction

The task of descriptor construction is to improve the distinguishability of features and describe the intensity variations/patterns of images from different sources as independently as possible to achieve robustness to GDs. However, some studies (Li et al., 2019; Ye et al., 2017) have shown that classical feature descriptors (such as SIFT and SURF) using the intensity or gradient distributions of images to construct feature vectors are sensitive to GDs. These descriptors are not suitable for multi-source image correspondence tasks.

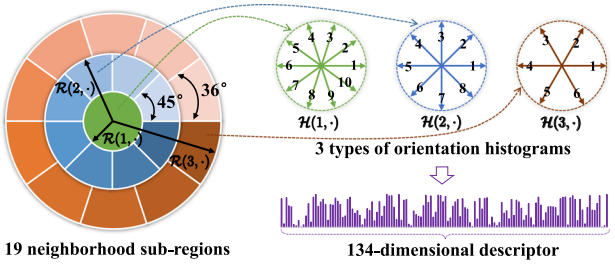

Fig. 2 A log-polar-based variable-size bin descriptor for multi-source images

To address this issue, some GLOH-like methods (Mikolajczyk and Schmid, 2005; Li et al., 2015) construct descriptors by using grid division rules to improve the robustness to GDs in log-polar coordinates. However, descriptors with different dimensions differ in stability and robustness for describing features. For instance, if the number of divided angle bins is too small, the characteristics of feature points may be insignificant; if the number of divided angle bins is too large, this may lead to high dimensions and increase computational complexity. Thus, the number of angle bins affects not only the representation of descriptors, but also affects the computational cost.

Note that multi-source images obtained from different viewpoints, especially high-resolution wide-baseline images, usually have significantly different appearances and local geometric distortions. This indicates that the geometric distortion levels of the corresponding local regions from two heterogeneous images increase with the deviation from the feature center. Therefore, to improve the stability and robustness of descriptors, as shown in Fig. 2, this paper proposes a novel variable-size bin strategy to construct distinctive descriptors, which divides circular neighborhoods centered on local feature points and constructs histograms by computing gradient amplitudes and orientations. Specifically, given a set of angular quantizations, each circular neighborhood is uniformly divided into sub-regions $\mathcal{R}(i,\iota)$. Then, variable-size bins for the gradient grids and orientation histograms are proposed, and the orientation histogram $\mathcal{H}(i,j)$ of each sub-region $\mathcal{R}(i,\iota)$ is calculated as the descriptors of the variable-size bins. The proposed descriptor $\mathcal{D}_d$ for each feature point in histogram quantization orientations can be represented as:

$$\begin{aligned} \mathcal{D}_d = \{&\mathcal{R}(1,1)\cdot\mathcal{H}(1,1),\ldots,\mathcal{R}(i,\iota)\cdot\mathcal{H}(i,j), \\ &\ldots,\mathcal{R}(n,\kappa)\cdot\mathcal{H}(n,m)\}, \\ &\forall i \in\{1,\ldots,n\},\forall j=\{1,\ldots,m\},\forall \iota=\{1,\ldots,\kappa\}, \end{aligned} \tag{8}$$

where n denotes the amount of the radial quantization, m denotes the amount of the histogram quantization, and κ denotes the amount of the angular quantization. Thus, the dimension of each descriptor can be described as $d = \sum_{i=1}^{n} m_i \cdot \kappa_i$. Finally, the descriptor vectors are normalized to reduce the influence of illumination variations. Compared with existing GLOH-like descriptors, the performance of the proposed descriptor is significantly improved by applying the variable-size bin strategy to divide local regions and compute gradient histograms.

With the proposed grid division strategy, the neighborhood region of feature points is divided into three-level sector neighborhoods (where the first, second, and third circular neighborhoods are divided into one, eight, and ten parts, respectively), thus generating a log-polar coordinate grid containing nineteen neighborhood sub-regions (see Sect. 4.2 for the detailed experimental results and analysis). So, this grid division strategy effectively compensates for the instability of traditional descriptors uniformly dividing grids (e.g., both the second and third circular neighborhoods are divided into eighteen parts), and increases the flexibility and scalability of descriptors. Meanwhile, to increase the distinctiveness of descriptors, for different levels of circular neighborhoods, this paper uses gradient histograms of different sizes as local descriptors according to the distance from the local feature center to the neighborhood, instead of using gradient histograms of the same size. Note that the level of geometric distortion increases with deviation from the feature center, and regions with a small geometric distortion contribute significantly to the descriptor structure. Therefore, the external location bins are assigned small orientation histograms to reduce the influence on the descriptor center.

To sum up, the proposed descriptor can significantly increase the robustness and distinctiveness of the descriptors through the circular neighborhood log-polar grid with variable-size bins, and it is robust to GDs of images with different viewpoints (as the experimental results shown in Sect. 4.2). Once the descriptors are obtained, the initial matching pairs $\mathcal{S}$ can be estimated by computing the distance ratio between the descriptors. However, traditional distance ratio metrics (such as Hamming or Euclidean distance) are usually difficult to accurately describe the relationships between features of multi-source images. By considering this, a robust model fitting method is proposed to deal with the issue in the following section.

3.3 Robust Heterogeneous Model Fitting

Given a set of initial matching pairs $\mathcal{S} = \{(s_i, s_i')\}_{i=1}^{N}$, where N is the number of matching pairs, and $s_i = (x_i, y_i)$ and $s_i' = (x_i', y_i')$ represent the coordinates of two feature points from two heterogeneous images, respectively. First, for each transformation model v, M model hypotheses $\boldsymbol{\theta}^{(v)} = \{\theta_i^{(v)}\}_{i=1:M}$ are generated for every two heterogeneous images using the transformation model $v \in \mathcal{V}$, where $\mathcal{V}$ is a set of transformation models of different types (e.g., $\boldsymbol{\theta}^{(s)} = \{\theta_i^{(s)}\}, \boldsymbol{\theta}^{(a)} = \{\theta_i^{(a)}\}$, and $\boldsymbol{\theta}^{(p)} = \{\theta_i^{(p)}\}$ corresponding to the similarity transformation model, the affine transformation model, and the perspective transformation model, respectively). These model hypotheses are generated through random sampling from a set of minimal subsets (e.g., at least three feature points are required as a minimal subset p to construct an affine transformation model). Then, the transformation error $\tau_{\theta_i^{(v)}}(s_i, s_i')$ of two feature points (s_i, s_i') is computed with respect to the model hypotheses $\theta_i^{(v)}$ by using the Sampson distance (Hartley and Zisserman, 2003) and form an ascending permutation:

$$\boldsymbol{\lambda}_i^{(v)} = \left[\lambda_{i,1}^{(v)}, \lambda_{i,2}^{(v)}, ..., \lambda_{i,M}^{(v)}\right], \tag{9}$$

which satisfies $\tau_{\theta_i^{(v)},\lambda_{i,1}^{(v)}} \leq ... \leq \tau_{\theta_i^{(v)},\lambda_{i,M}^{(v)}}$. Here, $\boldsymbol{\lambda}_i^{(v)}$ represents the preference of feature points with respect to a transformation model. Its value is small if the feature point belongs to the inlier of a transformation model, and vice versa.

In robust model fitting, an objective function is usually used to determine whether there is a structure in data. For example, RANSAC takes the sample consensus as the objective function in the parameter space to quantify the maximum consensus set of a structure by random sampling. Note that RANSAC prefers a larger structure if there are structures of different sizes in data. In contrast, the least k-th-order statistics (LkOS) estimator is widely used owing to its stability and breakdown bounds (Rousseeuw and Leroy, 2005). LkOS avoids the preference of the objective function for a large structure by minimizing the k-th-order statistics of squared residuals (Bab-Hadiashar and Hoseinnezhad, 2008). Therefore, to evaluate the quality of the model hypotheses generated by random sampling, this paper introduces a modified cost function for selecting the least k-th-order statistics of the squared transformation errors as follows:

$$\mathcal{C}(\theta^{(v)}) = \sum_{j=k-p+1}^{k} \boldsymbol{\lambda}_j^{2(v)}, \tag{10}$$

where $\boldsymbol{\lambda}_j^{2(v)}$ denotes the j-th sorted squared transformation error, and k represents the acceptable size of a structure, which is greater than that of a minimal subset ($k \gg p$).

With the above cost function, the significant transformation model can be effectively quantified as the minimal cost of the k-th-order statistics. However, for multi-source images with local distortions and homonymous points, the inliers constrained by the significant transformation model

Algorithm 1 Heterogeneous model fitting for multi-source image correspondence (MIMF)

Input: Image pairs, the heterogeneous models $\mathcal{V}$, the number of model hypotheses M, and the acceptable size of a structure k;
Output: The final matching pairs $\mathcal{S}^*$;
Extract feature points via Section 3.1;
Construct descriptors and calculate the initial matching pairs $\mathcal{S} = \{(s_i, s_i')\}_{i=1}^{N}$ via Section 3.2;
for each type of models $v \in \mathcal{V}$ **do**
  $\mathcal{C}(\theta_{best}^{(v)}) \leftarrow \infty, \varphi \leftarrow 3\ pixels$;
  Generate M hypotheses $\{\theta_j^{(v)}\}_{j=1:M}$;
  for $j = 1 : M$ **do**
    $\tau_{\theta_j^{(v)}}(s_j, s_j') \leftarrow$ Calculate the transformation error according to $\theta_j^{(v)}$;
    $\boldsymbol{\lambda}_j^{(v)} \leftarrow \text{Sorted}\left(\tilde{\tau}_{\theta_j^{(v)}}(s_j, s_j')\right)$;
    $\mathcal{C}(\theta_j^{(v)}) \leftarrow$ Evaluate the cost of $\boldsymbol{\lambda}_j^{(v)}$ by Eq. (10);
    if $\mathcal{C}(\theta_j^{(v)}) < \mathcal{C}(\theta_{best}^{(v)})$ **then**
      $\mathcal{I}^{(v)} \leftarrow [\boldsymbol{\lambda}_j^{(v)}]_{j=k-p+1}^{k}$;
      $\mathcal{C}(\theta_{best}^{(v)}) \leftarrow \mathcal{C}(\theta_j^{(v)})$;
    end if
  end for
  $\tilde{\theta}^{(v)} \leftarrow \text{LeastSquareFit}(\mathcal{I}^{(v)})$;
  $\tau_{\tilde{\theta}^{(v)}}(s_j, s_j') \leftarrow$ Update $\tau_{\theta_j^{(v)}}(s_j, s_j')$ according to $\tilde{\theta}^{(v)}$;
  $\tilde{\mathcal{S}}^{(v)} \leftarrow \tau_{\tilde{\theta}^{(v)}}(s_j, s_j') < \varphi$;
  $\tilde{\mathcal{S}}^{(v)} \leftarrow$ Update $\tilde{\mathcal{S}}^{(v)}$ from the descriptors according to Eq. (11);
end for
$\mathcal{S}^* \leftarrow$ Fusion $\tilde{\mathcal{S}}^{(v)}$ by Eq. (12).

are difficult to cover all matching pairs. To obtain more correct matching pairs, inspired by Li et al. (2009), this paper combines horizontal and vertical displacements as constraint criteria to extract more correct matching pairs. Specifically, the estimated significant transformation model is first utilized to obtain a small number of reliable feature-matching pairs. Then, the offsets of these matching pairs in the horizontal and vertical directions are computed as position transformation errors to constrain the feature descriptors. Finally, a new joint position offset transformation error is defined as:

$$\mathcal{J}^{(v)}(s_i, s_j') = \left(1 + \mathbb{E}^{(v)}(s_i, s_j')\right) \cdot \mathbb{D}(s_i, s_j'), \tag{11}$$

where $\mathbb{D}(s_i, s_j')$ indicates the inverse cosine similarity of the descriptors corresponding to s_i and s_j'; $\mathbb{E}^{(v)}(s_i, s_j') = \|s_i - \tau_{\theta_j^{(v)}}(s_j, s_j')\|$ represents the position transformation error between matching pairs (s_i, s_j'). Note that incorrect matching pairs can hardly satisfy minimal transformation errors and the same horizontal and vertical displacements at the same time, while correct matching pairs usually have relatively small transformation errors and the same horizontal and vertical displacements (Li et al., 2009). It implies that a smaller joint position offset transformation error indicates a correct matching pair and otherwise an incorrect matching pair. Therefore, this paper identifies those matching pairs with the minimum $\mathcal{J}^{(v)}(s_i, s_j')$ as the candidate matching pairs $\tilde{\mathcal{S}}^{(v)}$. After this, the correct matching pairs corresponding to different transformation models are accumulated, and thus all the correct matching pairs from multi-source images can be determined as:

$$\mathcal{S}^* = \mathcal{F}\left(\sum_{v \in \mathcal{V}} \tilde{\mathcal{S}}^{(v)}\right), \tag{12}$$

where $\mathcal{F}(\cdot)$ denotes a refinement operation that removes duplicate matches. The procedure of the proposed MIMF is shown in Algorithm 1.

4 Experiments

In this section, to evaluate the effectiveness of the proposed MIMF method, it is compared with nine state-of-the-art methods, including SIFT (Lowe, 2004), UR-SIFT (Sedaghat et al., 2011), PSO-SIFT (Ma et al., 2016), OS-SIFT (Xiang et al., 2018), RIFT (Li et al., 2019), HAPCG (Yao et al., 2021), LPSO (Yang et al., 2022), MSHLMO (Gao et al., 2022), and COFSM (Yao et al., 2022). Also, the robustness of the proposed MIMF and these competing methods is investigated on seven multi-source datasets.

4.1 Datasets and Evaluation Criteria

In this subsection, the multi-source datasets used in this work are listed below:

(1) The UR-SIFT dataset (UR-SIFT-DS) (Sedaghat et al., 2011) contains 14 images with resolutions ranging from 400×400 to 799×799, and they have two types: inter-band images with simulated distortions and multi-sensor images. These images are scaled by at least 2.5 times and rotated by 20° and cover a spatial resolution from 1 to 30 ms with illumination differences and scene changes.

(2) The CFOG dataset (CFOG-DS) (Ye et al., 2019) contains 20 images with resolutions ranging from 512×512 to 1074×1080 in four modalities (i.e., optical-infrared, LiDAR-optical, optical-SAR, and optical-map) obtained at different times with significant differences in intensity and texture. The registration of optical images and map images (or SAR images) is challenging due to text labels in map images (or significant speckle noises in SAR).

(3) The RIFT dataset (RIFT-DS) (Li et al., 2019) contains 12 multi-source images with resolutions ranging from 400×400 to 500×500 from multi-sensor, multi-temporal, and artificially generated images. These image pairs cover remote sensing images, satellite images, and close-range images with serious radiation distortions.

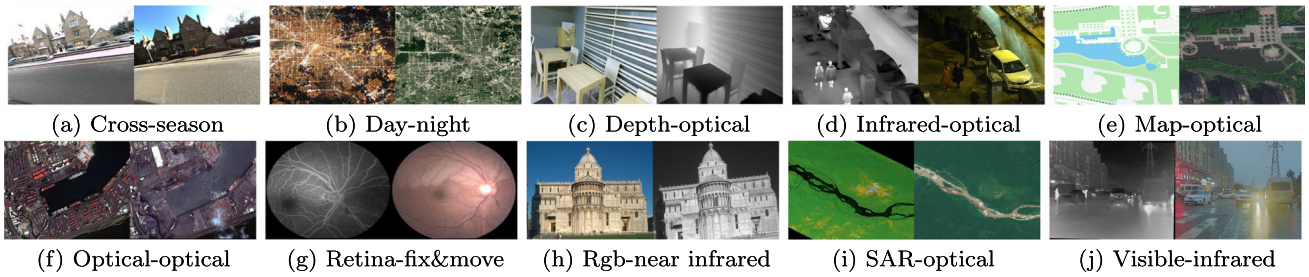

Fig. 3 Illustrations of some sample image pairs on the constructed TENM-DS dataset

(4) The HAPCG dataset (HAPCG-DS) (Yao et al., 2021) contains 8 heterogeneous remote sensing images with resolutions ranging from 400×400 to 500×500 from NASA, Google Earth, Landsat, and Sentinel satellites. These images include illumination, contrast, rotation, and comprehensive differences.

(5) The LPSO dataset (LPSO-DS) (Yang et al., 2022) contains 16 multi-source images with resolutions from 500×500 to 648×648. These images include intensity, scale, rotation, and translation characteristics from Landsat and Google.

(6) The COFSM dataset (COFSM-DS) (Yao et al., 2022) contains 92 remote sensing images with resolutions ranging from 450×450 to 661×661 covering six different types. These images have significant nonlinear radiation distortions and multiple application scenarios, such as multi-source data interpretation, multi-structure data registration, and multi-spectral data fusion.

(7) The constructed dataset (TENM-DS). This paper constructs a multi-source image dataset, which contains ten types of modality image pairs from Google and several public datasets for qualitative and quantitative evaluation, including cross-season, day-night, RGB-depth, infrared-optical, map-optical, optical-optical, retina-fix and move, rgb-near infrared, SAR-optical, and visible-infrared. The TENM-DS dataset contains a total of 48 multi-modal images with resolutions from 256×256 to 1280×1024, and each type of dataset contains two to four image pairs. Some sample image pairs are shown in Fig. 3. It can be seen from Fig. 3 that there are apparent nonlinear radiation distortions, geometric distortions, texture differences, and contrast variations between these multi-source image pairs, accompanied by illumination differences and scene changes.

To evaluate the performance of the ten competing methods, four evaluation criteria are used for quantitative comparison, including the number of correct matches (NCM), the success rate (SR), the root mean square error (RMSE), and the running time. Following (Yao et al., 2022), the SR of a set of matching pairs (s_i, s_i') is given below:

$$SR = \frac{\Sigma_i \mathcal{L}(s_i, s_i')}{TNI} \times 100\%, \tag{13}$$

where $\mathcal{L}(s_i, s_i') = \{1|\text{NCM} \geq p \ \& \ ((\text{NGT} \geq th_1)/\text{NCM}) \geq th_2\}$, otherwise $\mathcal{L}(s_i, s_i') = 0$; NGT indicates the number of correctly matched image pairs computed by the ground-truth model; p indicates the minimal subset required to solve a transformed model (e.g., four feature points for a perspective transformation model). Similar to Yao et al. (2022), th_1 and th_2 are the thresholds, and they are set to 3 and 20%, respectively; TNI indicates the total number of image pairs. Correspondingly, the RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{NCM} \sum_{i=1}^{NCM} (s_i - \tau_{gt}(s_i, s_i'))^2}, \tag{14}$$

where $\tau_{gt}(\cdot, \cdot)$ represents the ground-truth transformation error between s_j and s_j'. Here, those matches whose residuals are less than three pixels are defined as the correct matches (Li et al., 2019). It is worth noting that robust estimation techniques may fail if the number of correct matches for an image pair is too small (Li et al., 2022). Therefore, this paper considers an image pair containing more than ten correct matches as a correctly matched image pair, and considers other image pairs as incorrectly matched image pairs and fixes their RMSE to ten (Li et al., 2022). Additionally, if the estimated number of correct matches for each image pair is less than two or empty, the matching result is marked as NaN. Moreover, according to Tennakoon et al. (2016), the acceptable size of a structure k is set to ten, representing at least ten inliers. The experiments are conducted on a workstation equipped with dual Intel Xeon 4210R/256GB/RTX 3090, and then the inference time consumed by various components is recorded as the running time.

4.2 Influence of Variable-Size Bins

The robustness and discriminativeness of the proposed variable-size bin descriptor rely on the division of sub-region grids and orientation histograms. Thus, the number of sub-region grids and orientation histograms in the neighborhood of feature points is the key to constructing the log-polar-based variable-size bin descriptor. To evaluate the influence of different numbers of sub-region grids and orientation

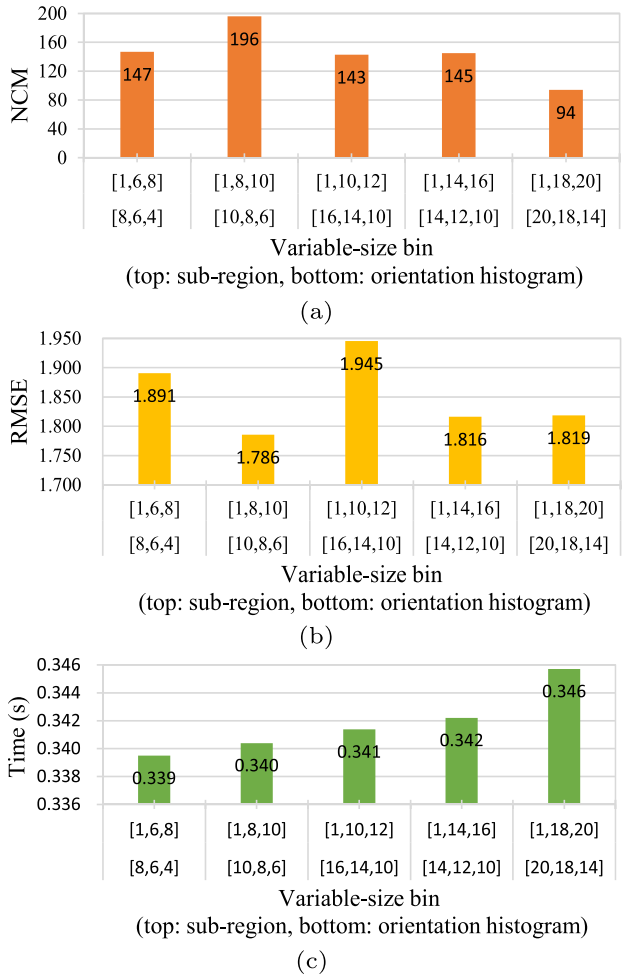

Fig. 4 Influence of variable-size bins on the proposed MIMF

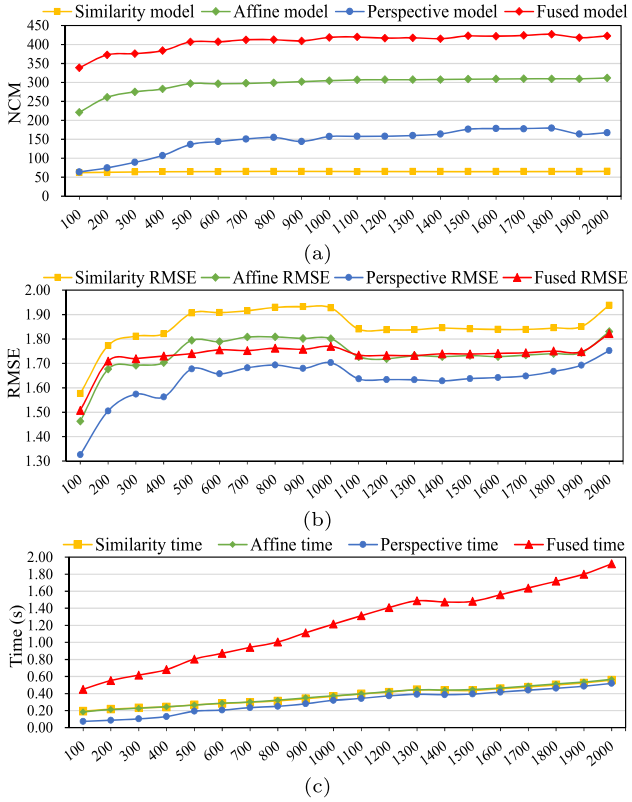

Fig. 5 Influence of sampling frequency on the proposed MIMF

histograms on the proposed descriptor, NCM, RMSE and running time are used. Similar to the GLOH-based method (Mikolajczyk and Schmid, 2005), the number of the radial quantization is fixed to 3. Figure 4 shows the NCM, RMSE and running time obtained by the proposed method for different sub-region grids and orientation histograms on the RGB-depth image pairs from the TENM-DS dataset. As illustrated in Fig. 4, when the number of sub-region grids and orientation histograms are set to [1,6,8] and [8,6,4], the NCM, RMSE, and runtime obtained by the proposed method are 147, 1.891, and 0.339s, respectively. As the numbers of sub-region grids and orientation histograms increases to [1,18,20] and [20,18,14], the NCM, RMSE and running time obtained by the proposed method are 94, 1.819 and 0.346s, respectively. It can be seen that fewer sub-region grids and orientation histograms can reduce the running time, but the RMSE is higher; while more sub-region grids and orientation histograms consume more running time and yield less NCM. In contrast, when the numbers of sub-regions and orientation histograms are set to [1,8,10] and [10,8,6], the proposed method obtains the best NCM (i.e., 195) and RMSE (i.e., 1.786) and competitive running time (i.e., ranked second). Therefore, these parameters will be used for the following experiments.

4.3 Influence of Sampling Frequency

Model fitting usually requires sampling a minimal subset from data to generate model hypotheses that hit the real model instances. Sampling more minimal subsets leads to better fitting results, but it also consumes more time. Therefore, the number of minimal subsets (i.e., sampling frequency) is a critical factor in evaluating the performance of a model fitting method. To evaluate the influence of sampling frequency on the performance of the proposed MIMF method (i.e., the fused model), three basic transformation models (i.e., the similarity model, the affine model, and the perspective model) are taken as benchmarks. This paper gradually increases the sampling frequency from 100 to 2000 and then reports the average NCM, RMSE, and running time on TENM-DS in Fig. 5.

As illustrated in Fig. 5, when the sampling frequency is increased from 100 to 500, the NCM obtained by the three models (i.e., the affine model, the perspective model, and the fused model) gradually increases, while the NCMs obtained

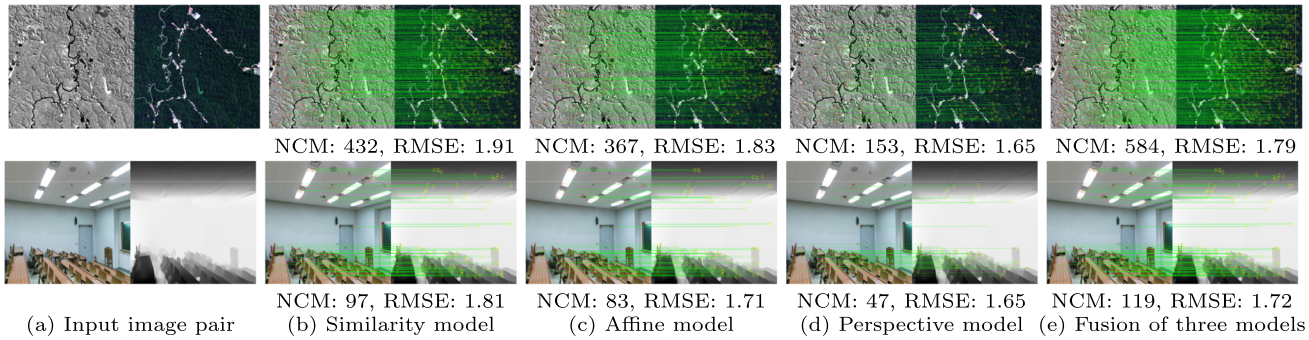

Fig. 6 The NCM and RMSE obtained by the proposed MIMF and three basic transformation models on the infrared-optical (top) and RGB-depth (bottom) image pairs from the TENM-DS dataset

by the similarity model hardly changes. This is because more sampling frequencies can generate more model hypotheses to increase the number of correct matches, but some invalid model hypotheses yield incorrect matches and thus a higher RMSE. When the sampling frequency is increased from 500 to 1000, the NCMs obtained by all four models changes slightly, while the RMSEs obtained by these models are all at a high level. When the sampling frequency is increased from 1000 to 1900, the RMSEs obtained by the four models decrease to a stable level. This indicates that higher NCMs and lower RMSEs are generated. When the sampling frequency exceeds 1900, the RMSEs obtained by the four methods start to increase. This is because when the sampling frequency exceeds a threshold (e.g., 1900 in this case), some data are repeatedly sampled thus increasing the RMSEs. Besides, the running time consumed by all four models increases with the sampling frequency. Based on the above analysis, when the sampling frequency is 1100, the NCM obtained by the proposed MIMF is better than that of the other three basic transformation models, and the RMSE obtained by the proposed MIMF is lower than that of the similarity model and similar to that of the affine model. Therefore, this paper sets the sampling frequency at 1100 in the following experiments to balance the NCM, RMSE, and running time.

4.4 Influence of Heterogeneous Models

In this subsection, the influence of the proposed MIMF method (i.e., the fused model) and the three basic transformation models (i.e., the similarity model, the affine model, and the projection model) on the performance of image correspondences is evaluated on two multi-source image pairs (i.e., infrared-optical and RGB-depth) from the TENM-DS dataset. Figure 6 presents the visualization results. It can be seen that for the infrared-optical image pairs, the NCM obtained by the proposed MIMF method (i.e., the fused model) is 584, while the NCMs obtained by the other three basic models are 432, 367, and 153, respectively. Compared with each independent basic model (i.e., the similarity model, the affine model, and the projection model), the proposed fused model improves the NCMs by 35.2%, 59.2%, and 281.7%, respectively. Meanwhile, the RMSE obtained by the proposed MIMF ranks second. For RGB-depth image pairs, the NCMs obtained by the proposed MIMF is 22.7%, 43.4%, and 153.2% higher than those of the three independent basic models, and the obtained RMSE ranks second. It can be seen that the proposed MIMF method (i.e., the fused model) can increase the correspondence to multi-source image pairs (up to 281.7% in the infrared-optical image pairs and 153.2% in the RGB-depth image pairs) with lower RMSEs. This indicates that the MIMF method effectively integrates the advantages of different types of basic models to improve the number of correspondences.

4.5 Results on the Public Datasets

To intuitively evaluate the matching performance, experiments are first conducted on six public multi-source datasets.

Table 1 shows the quantitative results obtained by the ten competing methods on the six public multi-source datasets, where the higher the values of NCM and SR, and the lower the values of RMSE and the running time, the better. Figure 7 shows some representative visualization results obtained by SIFT, UR-SIFT, PSO-SIFT, OS-SIFT, RIFT, HAPCG, LPSO, MSHLMO, COFSM, and the proposed MIMF method, respectively. As shown in Table 1 and Fig. 7, SIFT-based methods (i.e., SIFT, UR-SIFT, POS-SIFT, and OS-SIFT) obtain unsatisfactory matching results.

For example, the number of total average NCMs obtained by SIFT, PSO-SIFT, and OS-SIFT is less than 30; the SRs obtained by SIFT, UR-SIFT, and POS-SIFT are zeros on the CFOG-DS dataset, and the SR obtained by OS-SIFT is zero on the HAPCG-DS dataset (i.e., a sufficient number of correct matches are not obtained). Additionally, OS-SIFT fails on the UR-SIFT-DS dataset (i.e., there are no correct matching results in all the image pairs). This is because SIFT relies on the feature description of gradient histograms, and the NRDs

Table 1 Quantitative results (i.e., NCM, SR, RMSE and Time) obtained by the ten competing methods on the six public datasets

Method		UR-SIFT-DS	CFOG-DS	RIFT-DS	HAPCG-DS	LPSO-DS	COFSM-DS	Total mean
SIFT (2004)	NCM	36.67	7.70	8.67	8.50	62.50	8.69	22.12
	SR	28.57	0.00	16.67	25.00	50.00	21.74	23.66
	RMSE	3.70	10.00	8.43	7.63	5.29	7.87	7.15
	Time	1.99	5.48	2.96	**2.04**	2.74	**2.25**	**2.91**
UR-SIFT (2011)	NCM	455.29	59.40	75.33	107.75	327.00	52.22	179.50
	SR	71.43	0.00	16.67	25.00	50.00	13.04	29.36
	RMSE	4.98	10.00	8.83	8.25	6.44	9.08	7.93
	Time	30.77	32.84	27.93	30.14	30.50	26.53	29.78
PSO-SIFT (2016)	NCM	15.33	12.80	37.33	22.75	71.63	11.29	28.52
	SR	14.29	0.00	33.33	25.00	62.50	13.04	24.69
	RMSE	3.60	3.41	3.74	5.29	2.94	5.59	4.10
	Time	**1.08**	15.31	7.77	2.68	8.15	3.92	6.49
OS-SIFT (2018)	NCM	NaN	10.10	12.60	3.00	24.00	6.54	11.25
	SR	NaN	20.00	16.67	0.00	50.00	15.22	16.98
	RMSE	NaN	8.86	8.78	10.00	6.16	8.90	8.54
	Time	NaN	11.33	6.61	4.26	4.98	4.76	6.39
RIFT (2019)	NCM	140.57	449.50	509.67	162.75	253.13	254.70	295.05
	SR	28.57	60.00	**100.00**	25.00	50.00	60.87	54.07
	RMSE	7.49	4.80	**1.23**	7.83	5.61	4.69	5.28
	Time	3.35	**4.73**	3.51	3.45	3.64	3.32	3.67
HAPCG (2021)	NCM	228.00	765.30	**677.67**	325.00	**537.38**	361.00	482.39
	SR	14.29	10.00	16.67	0.00	25.00	4.35	11.72
	RMSE	8.65	9.19	8.64	10.00	7.93	9.64	9.01
	Time	14.63	20.23	9.88	11.05	12.15	10.52	13.08
LPSO (2022)	NCM	107.43	78.20	291.67	87.50	251.75	69.17	147.62
	SR	14.29	0.00	33.33	0.00	37.50	6.52	15.27
	RMSE	8.81	10.00	7.21	10.00	6.83	9.46	8.72
	Time	5.27	6.79	5.40	5.61	5.91	5.66	5.77
MSHLMO (2022)	NCM	133.00	59.40	76.33	32.00	101.13	22.07	70.66
	SR	**100.00**	50.00	**100.00**	50.00	60.87	**95.65**	76.09
	RMSE	2.81	4.23	1.62	3.89	2.62	4.16	3.22
	Time	41.00	55.08	40.64	37.95	43.37	41.02	43.18
COFSM (2022)	NCM	265.14	792.88	666.83	322.00	252.00	**372.24**	445.18
	SR	71.43	20.00	66.67	75.00	**75.00**	43.48	58.60
	RMSE	4.10	3.52	1.87	3.87	1.71	2.39	2.91
	Time	60.63	165.83	24.21	19.86	36.49	21.11	54.69
MIMF	NCM	**693.00**	**818.63**	494.20	**450.00**	378.50	366.14	**533.41**
	SR	85.71	**80.00**	83.33	**100.00**	**75.00**	89.13	**85.53**
	RMSE	**1.75**	**1.89**	1.80	**1.81**	**1.66**	**2.00**	**1.82**
	Time	3.71	5.04	**2.74**	2.69	**2.47**	2.44	3.18

NaN indicates insufficient matching results. The best results are boldfaced

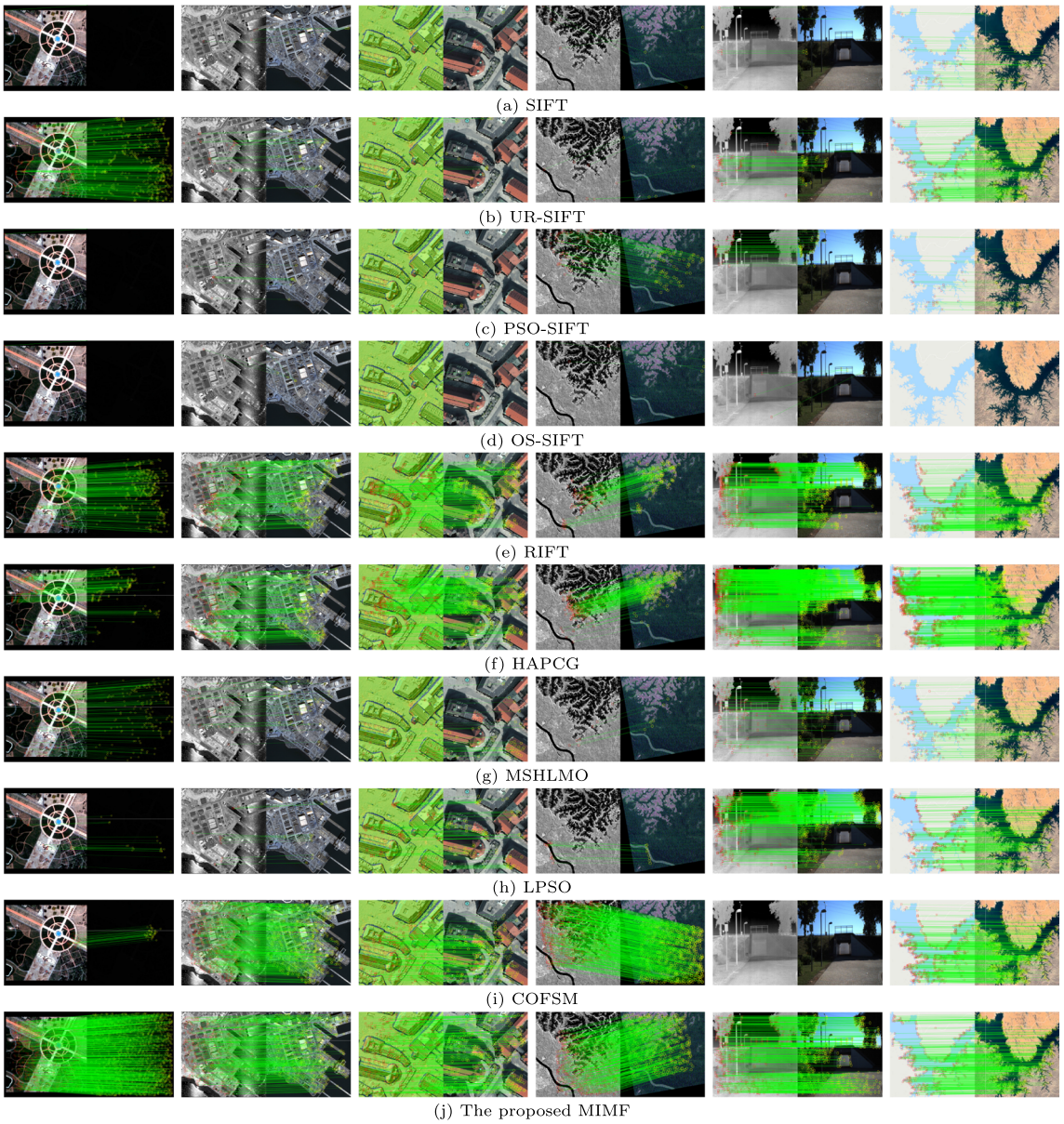

Fig. 7 Some image-matching results obtained by the ten competing methods on the six public datasets (i.e., the image pairs in the 1^{st} to 6^{th} columns are from the UR-SIFT-DS, CFOG-DS, RIFT-DS, HAPCG-DS, LPSO-DS, and COFSM-DS datasets, respectively)

in multi-source images make it difficult to correctly calculate the similarity of image pairs; PSO-SIFT and OS-SIFT only overcome the gradient orientation and intensity differences of images. In contrast, UR-SIFT increases the number of NCMs by introducing initial cross-matching to reduce the influence caused by the location and scale distributions, but the SR obtained by it is still zero on the CFOG-DS dataset.

The numbers of NCMs obtained by RIFT, HAPCG, LPSO, MSHLMO, and COFSM are significantly improved compared to the SIFT-based methods. For instance, RIFT obtains the best SR and the best RMSE on the RIFT-DS dataset. Though RIFT has better performance in contrast difference, rotation difference, and displacement difference, it does not support scale differences, so it performs poorly in data with scale differences. HAPCG obtains two best NCMs and the second best total average NCM by introducing a phase-consistent orientation histogram to alleviate the influence of nonlinear radiometric differences, but its matching performance drops, and it even fails in larger rotations. LPSO does not obtain better results because it uses only local phase sharpness features instead of gradient images, but it still obtains a better total average NCM than some SIFT-based methods (e.g., SIFT, PSO-SIFT, and OS-SIFT). MSHLMO obtains the three best SRs by introducing local principal ori-

entation maps with generalized gradient locations for feature extraction. However, the total average NCM obtained by it is still much lower than that of the SIFT-based method (i.e., UR-SIFT). COFSM obtains the best NCM, the best SR, and the second-best total average SR because its proposed co-occurrence filter reduces the effect of NRDs and extracts more edge features. However, it has limited matching success for the case of both NRDs and GDs. In contrast, the proposed MIMF method obtains three best NCMs, three best SRs, and five best RMSEs among all competing methods. Meanwhile, it also obtains the best total average NCM, SR, and RMSE. On the one hand, the proposed MIMF method combines multi-orientation phase coherence and image magnitude information to better detect feature points. On the other hand, the proposed variable-size bin strategy improves the discrimination of descriptors against local geometric distortions.

In terms of running time, SIFT achieves the lowest total average running time on two out of the six datasets, while the proposed MIMF method obtains the lowest total average running time on two out of the six datasets. Note that the number of the correct matching pairs (i.e., NCMs) detected by SIFT is very small due to NRDs, which reduces its time cost. Although the total average running time consumed by the proposed MIMF ranks second (only slightly slower than SIFT), the number of total average NCM obtained by the MIMF method is about 23.1 times higher than that of SIFT. Therefore, the proposed MIMF method is competitive in terms of running time.

The visualization results obtained by the ten competing methods on some representative image pairs are shown in Fig. 7. Among the six representative image pairs, the numbers of successful matches obtained by SIFT-based methods, i.e., SIFT, UR-SIFT, POS-SIFT, and OS-SIFT, are 3, 4, 5, and 1, respectively. RIFT, HAPCG, LPSO, MSHLMO, and COFSM can correctly estimate the matching pairs from the six representative image pairs (except COFSM in LPSO-DS), but the number of correct matching pairs estimated by them is still not significant on some image pairs (e.g., UR-SIFT-DS). Overall, the proposed MIMF obtains more correct matching results than these competing methods on most image pairs.

4.6 Results on the Constructed Dataset

This subsection presents the performance evaluation of the ten competing methods on the 48 real images from the constructed TENM-DS dataset, including ten types of modalities. These multi-modal image pairs have different imaging mechanisms and contain severe nonlinear radiation distortions. Therefore, performing feature matching on these image pairs is a challenging task. Table 2 shows the quantitative results obtained by the ten competing methods on the constructed TEMM-DS dataset, in which the higher the values of NCM and SR, and the lower the values of RMSE and the running time, the better. Figures 8 and 9 illustrate some representative visualization results obtained by SIFT, UR-SIFT, PSO-SIFT, OS-SIFT, RIFT, HAPCG, LPSO, MSHLMO, COFSM, and the proposed MIMF method, respectively.

As shown in Table 2, some SIFT-based methods (i.e., SIFT, UR-SIFT, and OS-SIFT) obtain the best SRs (i.e., all image pairs are correctly matched) on D8, and UR-SIFT also obtains the best SR on D6, but they obtain poor SRs on other multi-source image pairs. Meanwhile, PSO-SIFT and OS-SIFT fail to detect features on D7 and D5, respectively. From Table 2, it can be seen that the number of NCMs obtained by SIFT is less than 10 on eight out of ten image pairs, and this is because the gradient descriptor used by SIFT is sensitive to NRDs. UR-SIFT uses the entropy-based feature selection strategy to improve the uniform distribution of SIFT so that it obtains the best SRs on D6 and D8 and the best NCM on D8. Although PSO-SIFT also improves the gradient calculation of SIFT, the SRs obtained by it are still not significant, and the total average SR is only 8.33. OS-SIFT performs slightly better than PSO-SIFT with a total average SR of 36.67, and it obtains the best SR on D8. OS-SIFT uses a multi-scale Harris function to detect features, and its SR is improved on D8, but it still fails on D5.

Compared to the above methods, RIFT achieves better results, and it obtains five best SRs and one best NCM. RIFT uses consistent maps and maximum index to characterize features to reduce the influence of NRDs, but it is sensitive to speckle noise. For example, the SR obtained by RIFT is zero on D5. One reason is that severe speckle noise leads to inaccurate edge structure information. HAPCG obtains the best NCMs on D9 and D10, and the best SR on D8, but zero SRs on other multi-source image pairs (i.e., D2-D7 and D9-D10). LPSO obtains lower NCMs than HAPCG (except D8), but it still performs better than some SIFT-based methods (e.g., SIFT, POS-SIFT, and OS-SIFT) on D1-D10, and it also obtains two best RMSEs on D1 and D4. LPSO reduces the impact of noises by highlighting the contour features of the image pairs, thus improving the description of similarity features of multi-source images. Note that HAPCG and LPSO perform poorly on SR. This is because the initial matches estimated by them contain a large number of incorrect matches, which has a relatively large impact on the success rate. MSHLMO obtains the five best SRs and the four best RMSEs, attributed to its multi-scale feature extraction and matching strategy. However, the NCMs obtained by it is still smaller than that of RIFT, HAPCG, and LPSO. COFSM obtains two best NCMs and four best SRs. COFSM optimizes gradients by co-occurring scale spaces to improve robustness to NRDs, but in the case of serious NRDs and geometric distortions, its matching success rate is still limited. In contrast, the proposed MIMF method obtains the best total average NCM, SR, and RMSE among the ten competing methods.

Table 2 Quantitative results (i.e., NCM, SR, RMSE and Time) obtained by the ten competing methods on the TEMM-DS dataset

Method		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Total mean
SIFT (2004)	NCM	9.00	9.67	3.00	4.67	6.50	15.00	4.00	716.50	3.00	4.00	77.53
	SR	0.00	33.33	0.00	0.00	0.00	50.00	0.00	**100.00**	0.00	0.00	18.33
	RMSE	5.24	6.89	10.00	10.00	10.00	5.26	10.00	**0.45**	10.00	10.00	7.78
	Time	5.29	**2.76**	3.74	**2.91**	**1.26**	**1.26**	**0.78**	5.97	**0.59**	**1.56**	**2.61**
UR-SIFT (2011)	NCM	106.00	86.33	6.75	30.33	23.00	155.00	18.00	**1735.00**	8.50	33.50	220.24
	SR	50.00	0.00	0.00	0.00	0.00	**100.00**	0.00	**100.00**	0.00	50.00	30.00
	RMSE	2.98	5.29	10.00	5.17	6.48	2.97	2.91	2.99	6.41	6.46	5.17
	Time	20.07	29.34	14.12	27.70	25.37	17.21	32.37	25.52	20.28	21.89	23.39
PSO-SIFT (2016)	NCM	9.50	8.67	22.00	31.33	5.50	13.50	NaN	78.00	8.00	6.00	20.28
	SR	0.00	0.00	0.00	33.33	0.00	50.00	NaN	0.00	0.00	0.00	8.33
	RMSE	10.00	10.00	10.00	6.87	10.00	5.28	NaN	0.59	10.00	10.00	8.08
	Time	13.38	7.24	8.92	5.24	1.34	1.67	NaN	25.04	0.74	2.52	7.34
OS-SIFT (2018)	NCM	4.00	7.00	2.00	7.33	NaN	6.50	3.00	9.00	6.00	2.00	5.20
	SR	0.00	33.33	0.00	33.33	NaN	50.00	50.00	**100.00**	50.00	50.00	36.67
	RMSE	10.00	10.00	10.00	7.71	NaN	7.57	10.00	3.86	7.30	10.00	8.49
	Time	9.73	5.30	3.29	13.77	NaN	5.10	2.02	9.74	2.06	8.29	6.59
RIFT (2019)	NCM	96.50	270.67	193.75	233.33	133.00	275.50	**735.50**	1538.00	139.00	181.50	379.68
	SR	50.00	**66.67**	75.00	**66.67**	0.00	**100.00**	**100.00**	**100.00**	50.00	50.00	65.83
	RMSE	2.22	2.35	2.03	4.43	2.39	2.33	2.06	2.23	2.14	2.29	2.45
	Time	**3.89**	3.49	2.33	3.91	3.72	2.15	4.80	**4.46**	2.25	3.16	3.42
HAPCG (2021)	NCM	176.00	233.00	216.00	545.00	191.00	569.00	503.00	615.50	**277.50**	**632.00**	395.80
	SR	50.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	15.00
	RMSE	1.93	4.56	1.90	1.88	1.98	1.89	1.90	1.65	1.98	1.88	2.16
	Time	22.26	10.28	7.48	20.83	12.39	7.24	13.91	18.44	11.96	14.01	13.88
LPSO (2022)	NCM	84.50	48.33	42.50	40.67	13.50	322.00	276.50	1653.50	81.50	24.00	258.70
	SR	0.00	0.00	50.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	15.00
	RMSE	**1.75**	7.27	3.79	**1.74**	10.00	1.78	1.87	1.38	5.87	5.95	4.14
	Time	6.51	5.80	4.59	7.60	5.84	5.04	6.31	7.26	5.72	5.49	6.02
MSHLMO (2022)	NCM	14.50	25.00	9.00	12.00	14.00	74.50	35.00	790.50	29.00	29.00	103.25
	SR	50.00	**66.67**	75.00	**66.67**	**100.00**	**100.00**	**100.00**	**100.00**	50.00	50.00	75.83
	RMSE	5.95	**1.69**	7.82	7.27	5.95	**1.53**	**1.72**	1.04	**1.80**	5.88	4.07
	Time	67.05	42.40	28.22	55.15	38.88	36.39	44.96	57.22	53.38	49.25	47.29
COFSM (2022)	NCM	105.00	396.00	198.75	557.00	**391.00**	**653.50**	170.00	403.00	51.00	178.00	310.33
	SR	50.00	**66.67**	25.00	**66.67**	50.00	**100.00**	0.00	**100.00**	50.00	50.00	55.83
	RMSE	3.12	5.03	2.38	2.80	2.78	2.35	2.82	1.89	6.51	2.37	3.21
	Time	146.11	27.22	17.91	94.84	25.55	14.74	10.71	72.78	10.61	35.03	45.55
MIMF	NCM	**315.50**	**456.33**	**236.25**	**591.00**	226.00	437.50	143.50	1279.50	165.50	201.50	**405.26**
	SR	**100.00**	**66.67**	**100.00**	**66.67**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**93.33**
	RMSE	1.76	1.90	**1.76**	1.77	**1.85**	1.77	1.89	1.30	1.88	**1.79**	**1.77**
	Time	4.62	2.84	**1.89**	3.27	2.10	2.28	1.67	4.83	1.78	2.05	2.73

(D1: Cross-season; D2:Day-night; D3: RGB-depth; D4: Infrared-optical; D5: Map-optical; D6: Optical-optical; D7: Retina-fix and move; D8: RGB-near infrared; D9: SAR-optical; D10: Visible-infrared.)
NaN indicates insufficient matching results. The best results are boldfaced

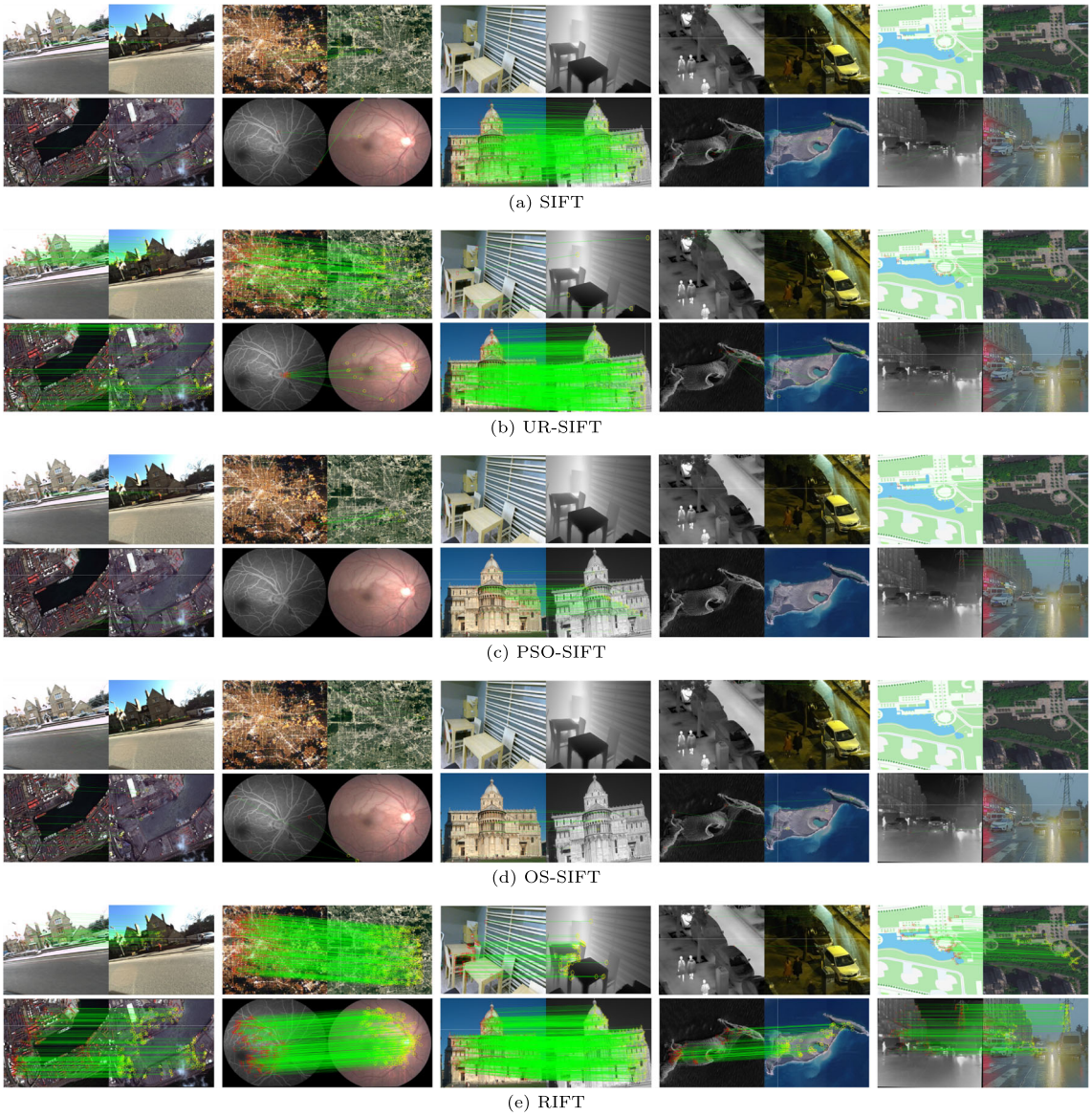

Fig. 8 Some image-matching results obtained by the five competing methods on the TENM-DS dataset (i.e., the image pairs from the top left to the bottom right correspond to the cross-season, day-night, RGB-depth, infrared-optical, map-optical, optical-optical, retina-fix and move, rgb-near infrared, SAR-optical, and visible-infrared, respectively)

For example, the proposed MIMF method obtains the best SRs among all the ten image pairs with different modalities. Also, it obtains much higher NCMs than the other competing methods. For example, the total average NCM obtained by the proposed method is about 4.2 times that of the classical SIFT and 30.6% higher than that of the state-of-the-art COFSM. The reasons may be twofold: (1) Our well-designed feature detectors and variable-size descriptors are suitable for multi-source images. (2) The use of heterogeneous model fitting for estimating the parameters of multiple transformation models effectively alleviates the influence of outliers on the matching performance.

In terms of running time, SIFT obtains the least total average running time (i.e., 2.61s), while MSHLMO obtains the most total average running time (i.e., 47.29s). In contrast, the proposed MIMF method obtains the second least total average running time (i.e., 2.73s). This is because the proposed MIMF method consumes more time to filter reliable descriptor structures when processing image pairs with a large resolution (e.g., D1 and D8), but it still outperforms the other eight competing methods.

The visualization results of some representative image pairs are presented in Figs. 8 and 9. Among the representative image pairs with ten different modalities, RIFT, HAPCG,

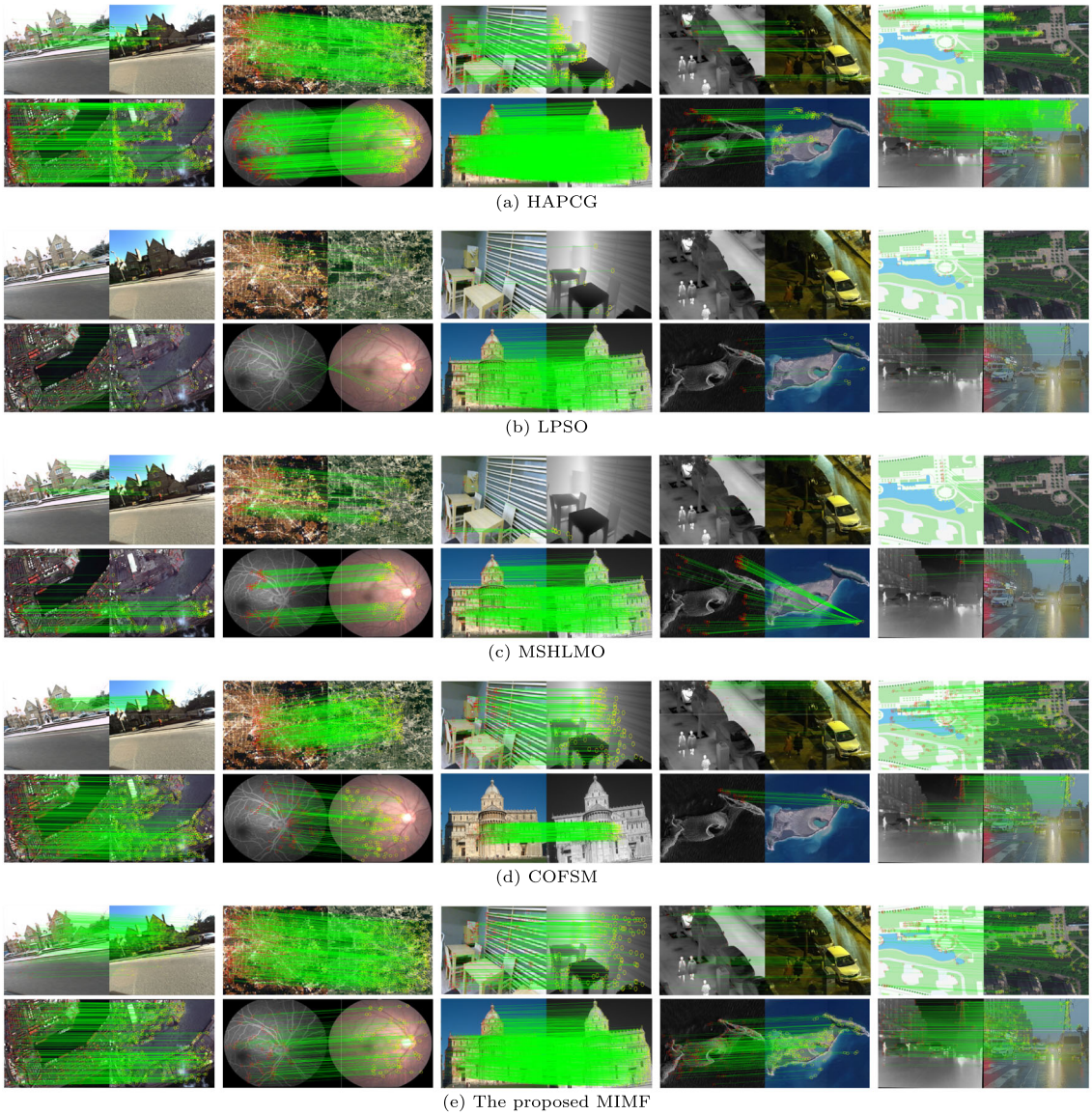

Fig. 9 Some image-matching results obtained by the five competing methods on the TENM-DS dataset (i.e., the image pairs from the top left to the bottom right correspond to the cross-season, day-night, RGB-depth, infrared-optical, map-optical, optical-optical, retina-fix and move, rgb-near infrared, SAR-optical, and visible-infrared, respectively)

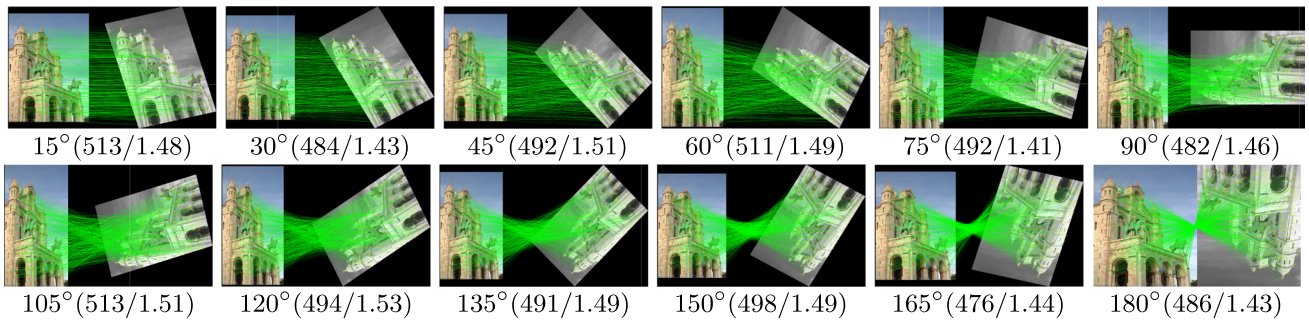

Fig. 10 Visualization results (NCM/RMSE) obtained by the proposed MIMF at different rotation angles on the rgb-near infrared image pairs from the TENM-DS dataset

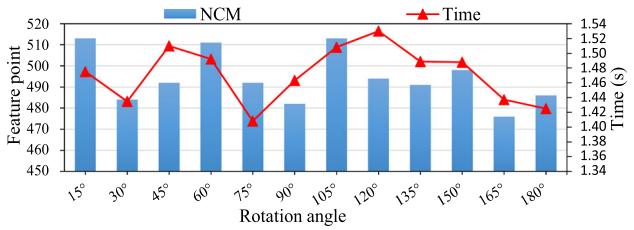

Fig. 11 The NCM and running time obtained by the proposed MIMF on the rgb-near infrared image pairs at different rotation angles

COFSM, and the proposed MIMF method can correctly estimate all the matching pairs. However, RIFT, HAPCG, and COFSM still obtain a smaller number of correct matching pairs than the proposed MIMF method on some image pairs (e.g., cross-season and RGB-depth). This shows that the proposed MIMF method is effective in processing multi-source image pairs.

4.7 Robustness Analysis of the Proposed MIMF

In this subsection, the robustness of the proposed MIMF under different rotation angles, different scales and the image registration performance under different modalities are evaluated. In addition, the performance of the proposed MIMF combined with other learning-based methods is also evaluated. For different rotation angles, the rbg-near infrared image pair from the TENM-DS dataset is first rotated from 15° to 180°, and then the visualization results obtained by the proposed MIMF are shown in Figs. 10 and 11. It can be seen from Figs. 10 and 11 that the proposed MIMF can obtain good matching results under different rotation angles. However, the rotation angles of image pairs can affect the NCM and running time obtained by the proposed MIMF. For example, when the rotation angle is 165°, the NCM obtained by the proposed MIMF is the least among all rotation angles. However, when the rotation angle is 120°, the running time obtained by the proposed MIMF is the highest among all rotation angles. It can be concluded that the rotation angles of image pairs can affect the robustness of the proposed MIMF, but it still achieves competitive results.

For different scales, the scale ratio between the retina-fix and move image pairs from the TENM-DS dataset is manually adjusted. Specifically, the scale of the image is reduced and increased according to the scale ratios 1:0.6, 1:0.8, 1:1.2, 1:1.4 and 1:1.6, and then the visualization results obtained by the proposed MIMF are shown in Fig. 12. From Fig. 12, it can be seen that the NCM exhibits variations in response to changes in the scale ratios. The closer the scale ratio approaches 1:1, the more NCM is detected. This is because some feature points are not obvious at smaller scales, and they may be offset at larger scales. However, the proposed MIMF still shows good performance under different scales.

Table 3 The NCM obtained by the four different components on the TEMM-DS dataset

Data	SP	SPMF	MVSG	MVNR
Cross-season	27.00	**85.00**	14.00	84.50
Day-night	30.00	**221.00**	1.00	75.00
RGB-depth	48.00	**607.00**	1.00	35.00
Infrared-optical	NaN	NaN	NaN	**206.33**
Map-optical	NaN	NaN	NaN	**19.50**
Optical-optical	64.50	221.50	NaN	**376.00**
Retina-fix and move	NaN	NaN	11.00	**62.50**
RGB-near infrared	918.00	**962.50**	186.50	816.50
SAR-optical	NaN	NaN	NaN	**78.00**
Visible-infrared	NaN	NaN	2.00	**102.50**
Total mean	217.50	**419.40**	35.92	185.58

NaN indicates insufficient matching results. The best results are bold-faced

Table 4 TThe RMSE obtained by the four different components on the TEMM-DS dataset

Data	SP	SPMF	MVSG	MVNR
Cross-season	1.62	1.84	**1.11**	1.80
Day-night	1.84	1.57	10.00	**1.51**
RGB-depth	**1.53**	1.69	10.00	1.85
Infrared-optical	NaN	NaN	NaN	**1.86**
Map-optical	NaN	NaN	NaN	**1.57**
Optical-optical	**1.54**	1.66	NaN	1.90
Retina-fix and move	NaN	NaN	**1.65**	1.86
RGB-near infrared	1.13	**1.12**	1.43	1.39
SAR-optical	NaN	NaN	NaN	**0.94**
Visible-infrared	NaN	NaN	10.00	**1.44**
Total mean	**1.53**	1.58	5.70	1.61

NaN indicates insufficient matching results. The best results are bold-faced

In particular, when the scale ratio reaches 1:1.6, there are still 134 matching pairs detected by the proposed MIMF.

In terms of registration robustness, the performance of the proposed MIMF is evaluated on three modality image pairs from the TENM-DS dataset (i.e. the day-night, the map-optical and the retina-fix and move). As shown in Fig. 13, the proposed MIMF can effectively detect the feature distribution of different modality image pairs, which helps to estimate the homography matrix model for accurate registration. The image registration and image fusion results show that the proposed MIMF can accurately match edges (e.g., the checkerboard registration task) and fuse images with different perspectives (e.g., the fusion registration task).

For evaluating the performance of the proposed component combined with other learning-based methods (i.e., SuperPoint and SuperGlue), four different combinations are

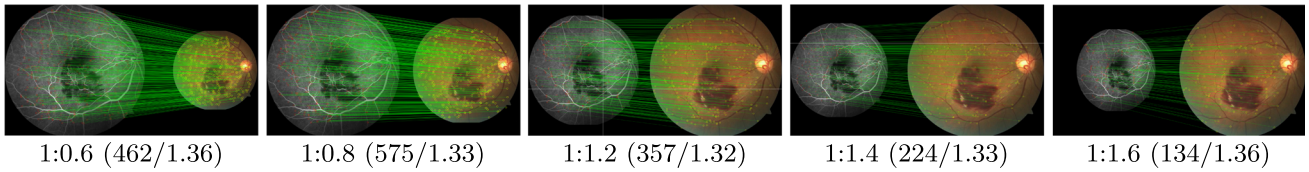

Fig. 12 Visualization results (NCM/RMSE) obtained by the proposed MIMF at different scales on the retina-fix and move image pairs from the TENM-DS dataset

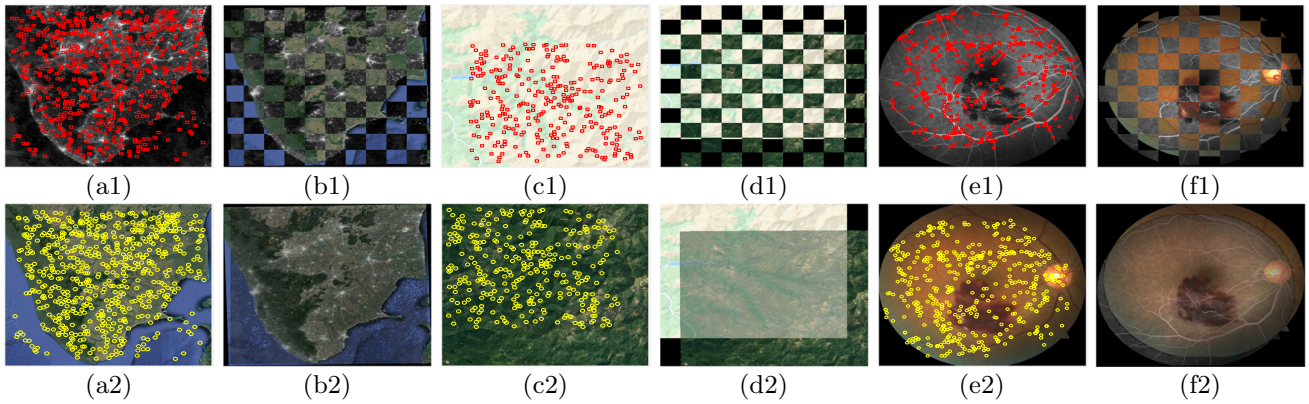

Fig. 13 Some registration results obtained by the proposed MIMF on the day-night (**a1**-**b2**), map-optical (**c1**-**d2**), and retina-fix and move (**e1**-**f2**) image pairs from the TENM-DS dataset. (**a1**-**a2**), (**c1**-**c2**), and (**e1**-**e2**) show the feature distribution of the image pairs. (**b1**), (**d1**), and (**f1**) show the checkerboard registration results. (**b2**), (**d2**), and (**f2**) show the fusion registration results

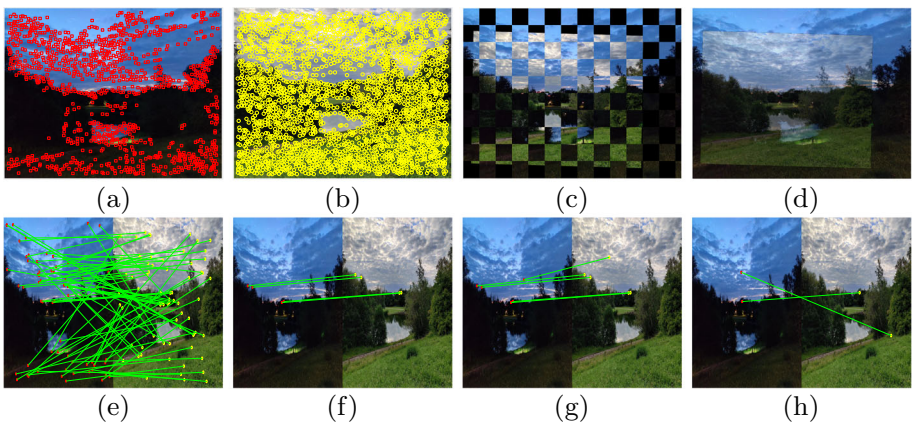

Fig. 14 A failure example obtained by the proposed MIMF. (**a**, **b**) The input image pair (i.e., day–night) with feature points (**c**) Checkerboard registration results. (**d**) Fusion registration results. (**e**) Initial correspondence. (**f**) Estimated similarity model correspondences. (**g**) Estimated affine model correspondences. (**h**) Estimated perspective model correspondences

constructed, including SuperPoint (SP) as a benchmark, SuperPoint with the proposed heterogeneous model fitting (SPMF), the proposed multi-orientation phase consistency model and variable-size bin strategy (MV) with SuperGlue (MVSG), and MV with nearest neighbor matching and RANSAC (MVNR). It is worth pointing out that SuperPoint is concerned with feature detection and description, while SuperGlue is concerned with feature matching; nearest neighbor matching and RANSAC are used to verify the reliability of the features and descriptors in MVSG. Additionally, since the official model used by SuperGlue needs to adapt a 256-dimensional descriptor (Shen et al., 2023), a 256-dimensional variable-size box descriptor is employed (i.e., the numbers of sub-region grids and orientation histograms are set to [1, 10, 10] and [16, 14, 10], respectively). Therefore, comparing SP and MVSG with SPMF and MVNR can verify the robustness of the proposed components. The NCM and RMSE obtained by these methods are then shown in Tables 3 and 4. For the NCM, the total average results obtained for SPMF and MVNR are about 0.9 and 4.2 times better than SP and MVSG, respectively. For the RMSE, although the total average results obtained by SP and SPMF are relatively similar, the total average results obtained by MVNR are reduced about 2.5 times compared to MVSG. This shows that the proposed multi-orientation phase consistency model, variable-size bin strategy and heterogeneous model fitting

components are effective in improving the performance of multi-source image correspondence.

4.8 Limitations of the Proposed MIMF

In this paper, a heterogeneous model fitting method is proposed to estimate multi-source image correspondence, which combines the advantages of three different types of basic models to achieve robust registration. However, the proposed MIMF fails in image correspondences when the model hypotheses generated by random sampling do not fit the true model hypotheses. Figure 14 shows a failure example of the proposed MIMF. As shown in Fig. 14, MIMF cannot effectively estimate enough correct matching pairs to produce a correct perspective model. This is because the detected feature points contain a large amount of noisy data, which may result in nearest matches (wrong matches), especially for regions with deformations. For instance, the sky area in Fig. 14b is covered with a large number of feature points that are prone to displacement, and it is challenging to eliminate false matches from these feature points. Therefore, the proposed MIMF fails to estimate the perspective model.

5 Conclusion

In this work, a robust heterogeneous model fitting method (i.e., MIMF) is proposed for multi-source image correspondence, which transforms the image-matching problem into a heterogeneous model fitting problem. First, a feature detection model is constructed based on the multi-orientation phase consistency to detect the structural texture and shape features of images to reduce the influence of multi-source image differences. Second, a log-polar coordinate-based descriptor operator with variable-sized bins is developed to describe the contributions of features with respect to the sub-region grids and orientation histograms, thereby enhancing the robustness to GDs. Finally, a robust heterogeneous model fitting method is proposed, which incorporates multiple types of basic transformation models for estimating multi-source image correspondences. Furthermore, a representative multi-source dataset containing ten different types of modalities is constructed. Quantitative and qualitative experimental results on six public datasets and one constructed dataset indicate that the proposed MIMF method can overcome the discrepancy between images caused by nonlinear radiation differences and geometric distortions, and it outperforms several state-of-the-art competition methods in terms of accuracy and efficiency. In the future, we will try to improve the accuracy of the proposed method in regions with deformation, and expand its application to three-dimensional images and point clouds.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U22A2095, U21A20514, U22B2028, 62272200, 62172197, 61825203), Natural Science Foundation of Guangdong Province (Grant Name: Research on key technologies of robust model fitting based on neighborhood constraints), Guangdong Key Laboratory of Data Security and Privacy Preserving (Grant No. 2023B1212060036).

Data Availability Statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Bab-Hadiashar, A., & Hoseinnezhad, R. (2008). Bridging parameter and data spaces for fast robust estimation in computer vision. In *Digital image computing: Techniques and applications* (pp. 1–8).

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *Proceedings of the European conference on computer vision* (pp. 404–417).

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 224–236).

Fan, Z., Liu, Y., Liu, Y., Zhang, L., Zhang, J., Sun, Y., & Ai, H. (2022). 3MRS: An effective coarse-to-fine matching method for multimodal remote sensing imagery. *Remote Sensing, 14*(3), 478.

Fan, Z., Zhang, L., Liu, Y., Wang, Q., & Zlatanova, S. (2021). Exploiting high geopositioning accuracy of SAR data to obtain accurate geometric orientation of optical satellite images. *Remote Sensing, 13*(17), 3535.

Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C., & Schmitt, M. (2019). Sar-to-optical image translation based on conditional generative adversarial networks-optimization, opportunities and limits. *Remote Sensing, 11*(17), 2067.

Gao, C., Li, W., Tao, R., & Du, Q. (2022). MS-HLMO: Multi-scale histogram of local main orientation for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 5626714.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press.

Hu, M., Sun, B., Kang, X., & Li, S. (2023). Multiscale structural feature transform for multi-modal image matching. *Information Fusion, 95*, 341–354.

Jiang, X., Jiang, J., Fan, A., Wang, Z., & Ma, J. (2019). Multiscale locality and rank preservation for robust feature matching of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 57*(9), 6462–6472.

Jiang, X., Ma, J., Xiao, G., Shao, Z., & Guo, X. (2021). A review of multimodal image matching: Methods and applications. *Information Fusion, 73*, 22–71.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K. M., & Trulls, E. (2021). Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision, 129*(2), 517–547.

Kelman, A., Sofka, M., & Stewart, C.V. (2007). Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–7).

Kovesi, P. (2003). Phase congruency detects corners and edges. In *Proceedings of the seventh international conference on digital image computing: techniques and applications* (pp. 1–10).

Kovesi, P. (1999). Image features from phase congruency. *Journal of Computer Vision Research, 1*(3), 1–26.

Kovesi, P. (2000). Phase congruency: A low-level image invariant. *Psychological research, 64*(2), 136–148.

Lai, T., Sadri, A., Lin, S., Li, Z., Chen, R., & Wang, H. (2023). Efficient sampling using feature matching and variable minimal structure size. *Pattern Recognition, 137*, 109311.

Le Moigne, J., Campbell, W. J., & Cromp, R. F. (2002). An automated parallel image registration technique based on the correlation of wavelet features. *IEEE Transactions on Geoscience and Remote Sensing, 40*(8), 1849–1864.

Li, J., Hu, Q., & Ai, M. (2017). 4FP-structure: A robust local region feature descriptor. *Photogrammetric Engineering & Remote Sensing, 83*(12), 813–826.

Li, J., Hu, Q., & Ai, M. (2019). RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing, 29*, 3296–3310.

Lin, S., Xiao, G., Yan, Y., Suter, D., & Wang, H. (2019). Hypergraph optimization for multi-structural geometric model fitting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 8730–8737).

Lin, S., Yang, A., Lai, T., Weng, J., & Wang, H. (2023). Multi-motion segmentation via co-attention-induced heterogeneous model fitting. IEEE *Transactions on Circuits and Systems for Video Technology* (pp. 1–13).

Lin, S., Luo, H., Yan, Y., Xiao, G., & Wang, H. (2022). Co-clustering on bipartite graphs for robust model fitting. *IEEE Transactions on Image Processing, 31*, 6605–6620.

Lin, S., Wang, X., Xiao, G., Yan, Y., & Wang, H. (2021). Hierarchical representation via message propagation for robust model fitting. *IEEE Transactions on Industrial Electronics, 68*(9), 8582–8592.

Li, Q., Qi, S., Shen, Y., Ni, D., Zhang, H., & Wang, T. (2015). Multispectral image alignment with nonlinear scale-invariant keypoint and enhanced local feature matrix. *IEEE Geoscience and Remote Sensing Letters, 12*(7), 1551–1555.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

Li, Q., Wang, G., Liu, J., & Chen, S. (2009). Robust scale-invariant feature matching for remote sensing image registration. *IEEE Geoscience and Remote Sensing Letters, 6*(2), 287–291.

Li, J., Xu, W., Shi, P., Zhang, Y., & Hu, Q. (2022). LNIFT: Locally normalized image for rotation invariant multimodal feature matching. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–14.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Ma, J., Chan, J.C.-W., & Canters, F. (2010). Fully automatic subpixel image registration of multiangle CHRIS/Proba data. *IEEE Transactions on Geoscience and Remote Sensing, 48*(7), 2829–2839.

Ma, J., Jiang, X., Fan, A., Jiang, J., & Yan, J. (2021). Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision, 129*(1), 23–79.

Ma, J., Jiang, X., Jiang, J., Zhao, J., & Guo, X. (2019). LMR: Learning a two-class classifier for mismatch removal. *IEEE Transactions on Image Processing, 28*(8), 4045–4059.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., & Liu, L. (2016). Remote sensing image registration with modified SIFT and enhanced feature matching. *IEEE Geoscience and Remote Sensing Letters, 14*(1), 3–7.

Ma, J., Zhao, J., Jiang, J., Zhou, H., & Guo, X. (2019). Locality preserving matching. *International Journal of Computer Vision, 127*(5), 512–531.

Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J., & Tian, J. (2015). Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Transactions on Geoscience and Remote Sensing, 53*(12), 6469–6481.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(10), 1615–1630.

Paul, S., & Pati, U. C. (2020). Automatic optical-to-SAR image registration using a structural descriptor. *IET Image Processing, 14*(1), 62–73.

Reddy, B. S., & Chatterji, B. N. (1996). An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing, 5*(8), 1266–1271.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. Wiley.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE international conference on computer vision* (pp. 2564–2571).

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4938–4947).

Sedaghat, A., Mokhtarzade, M., & Ebadi, H. (2011). Uniform robust scale-invariant feature matching for optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 49*(11), 4516–4527.

Shen, X., Hu, Q., Li, X., & Wang, C. (2023). A detector-oblivious multi-arm network for keypoint matching. *IEEE Transactions on Image Processing, 32*, 2776–2785.

Shi, J. (1994). Good features to track. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 593–600).

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8922–8931).

Teke, M., & Temizel, A. (2010). Multi-spectral satellite image registration using scale-restricted SURF. In *Proceedings of the IEEE international conference on pattern recognition* (pp. 2310–2313).

Tennakoon, R. B., Bab-Hadiashar, A., Cao, Z., Hoseinnezhad, R., & Suter, D. (2016). Robust model fitting using higher than minimal subset sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 350–362.

Uss, M. L., Vozel, B., Lukin, V. V., & Chehdi, K. (2016). Multimodal remote sensing image registration with accuracy estimation at local and global scales. *IEEE Transactions on Geoscience and Remote Sensing, 54*(11), 6587–6605.

Xiang, Y., Wang, F., & You, H. (2018). OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. *IEEE Transactions on Geoscience and Remote Sensing, 56*(6), 3078–3090.

Xiong, X., Xu, Q., Jin, G., Zhang, H., & Gao, X. (2019). Rank-based local self-similarity descriptor for optical-to-SAR image matching. *IEEE Geoscience and Remote Sensing Letters, 17*(10), 1742–1746.

Yang, W., Xu, C., Mei, L., Yao, Y., & Liu, C. (2022). LPSO: Multisource image matching donsidering the description of local phase sharpness orientation. *IEEE Photonics Journal, 14*(1), 1–9.

Yao, Y., Zhang, Y., Wan, Y., Liu, X., & Guo, H. (2021). Heterologous images matching considering anisotropic weighted moment and absolute phase orientation. *Geomatics and Information Science of Wuhan University, 46*(11), 1727–1736.

Yao, Y., Zhang, Y., Wan, Y., Liu, X., Yan, X., & Li, J. (2022). Multimodal remote sensing image matching considering co-occurrence filter. *IEEE Transactions on Image Processing, 31*, 2584–2597.

Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., & Zhu, Q. (2019). Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing, 57*(11), 9059–9070.

Ye, Y., Shan, J., Bruzzone, L., & Shen, L. (2017). Robust registration of multimodal remote sensing images based on structural similar-

ity. *IEEE Transactions on Geoscience and Remote Sensing, 55*(5), 2941–2958.

Ye, Y., Shen, L., Hao, M., Wang, J., & Xu, Z. (2017). Robust optical-to-SAR image matching based on shape properties. *IEEE Geoscience and Remote Sensing Letters, 14*(4), 564–568.

Ye, F., Su, Y., Xiao, H., Zhao, X., & Min, W. (2018). Remote sensing image registration using convolutional neural network features. *IEEE Geoscience and Remote Sensing Letters, 15*(2), 232–236.

Zeng, L., Du, Y., Lin, H., Wang, J., Yin, J., & Yang, J. (2020). A novel region-based image registration method for multisource remote sensing images via CNN. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 1821–1831.

Zhang, J., Ma, W., Wu, Y., & Jiao, L. (2019). Multimodal remote sensing image registration based onimage transfer and local features. *IEEE Geoscience and Remote Sensing Letters, 16*(8), 1210–1214.

Zhang, Y., Yao, Y., Wan, Y., Liu, W., Yang, W., Zheng, Z., & Xiao, R. (2023). Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing, 196*, 1–15.