Multiscale Template Matching for Multimodal Remote Sensing Image

Tian Gao, Chaozhen Lan, Wenjun Huang, Longhao Wang, Zijun Wei, and Fushan Yao

***Abstract*—Multimodal matching remains a difficult and pressing problem in the imaging processing community. Accurate and robust multimodal matching is important for the performance of applications, such as registration and fusion. Traditional image matching algorithms cannot effectively handle multimodal images with severe nonlinear radiometric distortion (NRD). In this article, a novel multiscale template matching algorithm for multimodal image matching is proposed to address this problem. We propose a novel frequency-domain convolutional map based on the wavelet transform and phase congruency to construct a feature description map that significantly reduces the NRD between multimodal images. The development of omnidirectional aggregated feature vectors with rotational invariance also helped to achieve robustness on rotated images. Finally, a multiscale template matching strategy improved the matching performance on multimodal images with displacement and scale variations. To improve the time efficiency of the algorithm, most of the complex computations in this article are performed in the frequency domain. According to the experimental findings on six multimodal image datasets, the method can obtain accurate and robust matching results between multimodal images. Through qualitative and quantitative evaluations, the method outperforms several mainstream multimodal image matching algorithms in terms of matching accuracy, success rate, and time consumption.**

***Index Terms*—Multimodal remote sensing images, multiscale strategy, template matching, wavelet transform.**

I. Introduction

With the rapid development of space and sensor technologies, remote sensing observation methods are gradually becoming more diverse. Diverse remote sensing observation data [e.g., optical, depth, light detection and ranging, infrared, and synthetic aperture radar (SAR)] and terrain data can be obtained. Combining multiple data to form multisource remote sensing data is beneficial for better change detection [1], [2], land classification [3], [4], [5], environmental simulation [6], target localization [7], and environmental and disaster detection [8], [9], [10]. Among these, multimodal image matching is essential for achieving the benefits of mutual complementarity and using data from multiple sources.

Manuscript received 17 June 2023; revised 2 September 2023 and 3 October 2023; accepted 16 October 2023. Date of publication 23 October 2023; date of current version 14 November 2023. *(Corresponding author: Chaozhen Lan.)*

The authors are with the PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China (e-mail: gt_2995330760@163.com; lan_cz@163.com; 13273718438@163.com; 848832204@qq.com; 664350477@qq.com; yaofushan123@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3326959

Fig. 1. Example of showing the differences between multimodal images.

Remote sensing images typically have higher resolution than natural images, and processing and analyzing these images may require more computing resources and time. The acquisition of remote sensing images is limited by lighting and weather conditions, which may lead to a decrease in image quality, affecting the visibility and recognition accuracy of features. Remote sensing images are usually acquired from platforms, such as satellites, airplanes, or drones, and their viewing angles and scales differ from natural scene images. This can lead to changes in the shape, size, and scale of figures in the image, adding complexity to image interpretation and analysis.

The purpose of image matching is to find the corresponding image points with the same name in two or more images containing the same scene and then find the transformation relationship between the images to achieve the registration of the reference frame of the two images. As shown in Fig. 1, there may be differences in displacement, rotation, scale, radiation, and noise between remote sensing observations from different sources [11]. The same feature may correspond to different locations and exhibit significant nonlinear radiometric distortion (NRD) due to different sensors causing the same feature to exhibit vast differences in appearance. These huge differences pose significant difficulties for the existing image-matching algorithms, which cause significant degradation in most algorithms' matching performance to meet the requirements of multisource data fusion. Therefore, it is important to develop a reliable multimodal image-matching algorithm to provide accurate multimodal observations of the same scene. To design a reliable multimodal image-matching algorithm, the following three issues must be addressed: the algorithm should: 1) be highly adaptive to NRD;

2) have a more accurate and stable outlier removal model; and 3) have strong adaptability to geometric differences.

Template matching [12] is a common image processing method in the field of computer vision that determines the matching region in the target image that most closely resembles the template image. In particular, it consists of the following three steps: 1) it designs a search strategy to determine the candidate matching window regions corresponding to all the template images; 2) it selects the feature vectors of the corresponding window regions on the designed feature description map; and 3) it calculates the similarity between the feature vectors of the template windows and those of the candidate matching windows using the designed similarity metric. The highest score candidate matching region is considered the final candidate matching window. Template matching is important for facial recognition [13], visual localization [7], and other processes.

Therefore, this study aims to investigate a template matching algorithm that can be used for multimodal image registration. Although slow and poorly adaptive to geometric differences, template matching has high accuracy. In addition, it is difficult to determine the mapping relationship between the template windows and candidate matching windows. Therefore, developing a multimodal image template matching algorithm that is faster and more adaptable to geometric differences is crucial. Based on the above analysis, this study developed an effective template matching search strategy for multimodal image registration, overcame difficulties in multisource remote sensing image matching and proposed a novel template matching framework for multisource remote sensing image registration. It maintains the advantages of the high matching accuracy of the template matching method while essentially solving the disadvantages of slow matching speed and poor geometric differences adaptability and can automatically complete multimodal remote sensing image registration under most conditions. The main contributions of this study are as follows.

1) A multiscale template matching (MSTM) algorithm is designed to solve the multimodal image registration problem with severe NRD. MSTM has strong robustness and reliability and can handle general multisource remote sensing image registration problems, including images from different sensors, time phases, views, and resolutions.
2) A frequency-domain convolutional map (FDCM) based on the wavelet transform and phase congruency (PC) is proposed, which can more accurately describe the feature information of multimodal images and significantly reduce the NRD of multimodal images. An omnidirectional aggregated feature vector with rotational invariance was designed to accomplish the automatic registration for poor quality multimodal images with significant geometric differences and NRD.
3) To address the problem of geometric differences of multimodal image data and of determining the mapping relationship between the template window and the candidate matching window, a multiscale template matching strategy based on the scale space was applied to MSTM. Good results were obtained by solving geometric differences and optimizing the template matching speed.

The rest of this article is organized as follows. Section II reviews the related work on multimodal remote sensing image registration. Section III describes the typical framework and the key processes of the proposed MSTM. The experimental results and performance analysis of MSTM for various multimodal image datasets are discussed in Section IV. Finally, Section V concludes this article.

II. Related Work

This section divides the matching methods into three categories, learning, feature, and template based, to briefly review them for multisource remote sensing images.

Although learning-based methods [14] are currently significantly advanced techniques, with many related studies having achieved good results on the corresponding datasets [15], they are difficult to use. The lack of a general multimodal image dataset for model training makes it impossible to train reliable and stable parameters; however, the trained models are poorly adapted to different real data and have insufficient generalization ability.

Feature-based methods [16] start with an image feature extraction and then perform matching using the similarity between the feature descriptors of the image feature points. The extracted features should be robust, stable, and repeatable, and they can be points [17], lines [18], and faces [19]. Ma et al. [20] proposed the PSO-SIFT algorithm, which is an improved method for calculating image gradients to improve the robustness to NRD. Li et al. [11] proposed the RIFT algorithm, which uses a PC and innovatively introduces a maximum index map descriptor (MIM). Yao et al. [21] proposed the MOTIF algorithm, which, with better performance, establishes a diffusion tensor model using the image gradient direction information and obtains multidirectional index map descriptors using this algorithm. Yao et al. [22] proposed the HAPCG algorithm, which uses anisotropic filtering for image nonlinear diffusion and constructs an anisotropic weighted moment scale space. The absolute PC direction gradient was established using the PC model and combined with the log-polar coordinate description template, which greatly enhanced the robustness. However, the matching accuracy of feature-based methods is typically lower than that of template-based methods because of the unstable localization accuracy of the extracted feature points.

The template-based [23] approach compares the similarity metrics of the two selected regions to accomplish matching. The sum of squares (SSD), normalized correlation coefficient (NCC), and mutual information (MI) are the typically used similarity metrics. The SSD is quick but less robust and sensitive to noise [24]. The NCC is less suitable as a similarity metric for multimodal image matching because it is highly adaptable to linear distortion but less adaptable to NRD [25], [26]. Although MI has good adaptability to NRD, it is more sensitive to matching windows and is computationally inefficient [27]. Recent studies have found that the geometric structure remains stable between multimodal data. Ye et al. [28] achieved good results in generating HOPC descriptors based on HOG [29] using PC maps. To describe the geometric structure features in

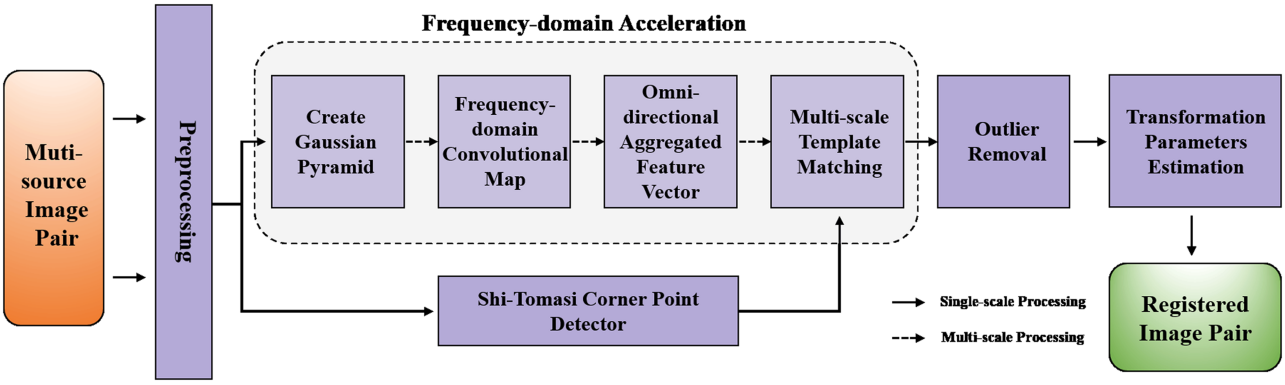

Fig. 2. Proposed MSTM multimodal remote sensing image registration framework mainly includes the Shi-Tomasi feature point detection, FDCM, omnidirectional aggregated feature vector, and multiscale template matching strategy using the scale space and the outlier removal method. The acceleration strategy converts the primary function into frequency-domain processing.

more detail, Ye et al. [30] further proposed the CFOG algorithm, which addressed the issue of not providing a detailed description of HOPC geometric structure features using pixel-by-pixel description and substantially improved the matching speed using frequency-domain template matching. However, regardless of how good the template-based method is, it has extremely high requirements for the initial position; for example, the CFOG uses the point position calculated using the satellite rational polynomial coefficient parameters as the initial matching reference control point. However, data from different sources lack spatial mapping relationships, and if the initial reference control point position deviates too much or there is no initial reference control point, the matching performance drops significantly. This is because it is sensitive to geometric differences, and the size of the selected template limits its speed.

Therefore, this study aimed to investigate a practical multimodal image registration method using template matching, which is quick, suitable for NRD, and adapts to geometric differences. We also generate the FDCM using the wavelet transform and PC in the feature description, which enhances the robustness of NRD. An omnidirectional aggregated feature vector with rotational invariance was designed that was robust to rotating multimodal images. By designing a multiscale matching strategy using the scale space, the template-based method is also robust to geometric differences, and a better performance is obtained in comparison experiments with CFOG, HAPCG, RIFT, SuperGlue, and Ms-HLMO on multiple multimodal image datasets.

III. Methodology

The framework of the proposed MSTM algorithm is shown in Fig. 2. The input image pair to be aligned is first preprocessed, primarily for basic image denoising. The Shi-Tomasi corner point detection algorithm is used to extract feature points from the multimodal image pair to provide initial control points for subsequent template window determination. The key to the proposed MSTM is the feature description map construction and template matching from a multiscale strategy. The multiscale space of the feature description map from the multiscale strategy is created by building a Gaussian pyramid, first constructing the FDCM at the lowest level and then downsampling the FDCM several times. The scale space is top down, and the feature templates of the feature points corresponding to windows are extracted at each level of the FDCM, the feature templates are rotated several times, and all the rotated feature templates are aggregated to generate an omnidirectional aggregated feature vector. Subsequently, the window with the highest similarity at the same position in the target image is selected as the correct match, and outlier removal is performed to remove incorrect matches. The correct corresponding scale level is determined using a multiscale strategy, and the parameters of the exact transformation matrix are transformed stepwise. Finally, the exact spatial transformation relationship between the image pairs is determined using the final computed exact transformation model.

A. Shi-Tomasi Feature Point Detection

The Harris corner point detector [31] is one of the most stable corner point extraction algorithms, and the Shi-Tomasi corner point extraction algorithm [32] is an improved version of the Harris corner point detector. This algorithm has a simpler calculation of corner point response values and typically produces better results than the Harris operator. The Shi-Tomasi corner point response values were calculated for each image, as follows:

$$\text{Corner} = \min(\lambda_1, \lambda_2) \tag{1}$$

$$\boldsymbol{M} = \begin{bmatrix} \sum_{\mathbf{W}_\sigma} \mathbf{G}_x^2 & \sum_{\mathbf{W}_\sigma} \mathbf{G}_x \mathbf{G}_y \\ \sum_{\mathbf{W}_\sigma} \mathbf{G}_x \mathbf{G}_y & \sum_{\mathbf{W}_\sigma} \mathbf{G}_y^2 \end{bmatrix} \tag{2}$$

where λ_1 and λ_2 are two eigenvalues of the matrix $\boldsymbol{M}$. $\mathbf{G}_x$ and $\mathbf{G}_y$ are the gradient values of the image along the x- and y-directions, respectively, and $\mathbf{W}_\sigma$ is a Gaussian sliding window with a variance σ. The pixel points with corner response values greater than the threshold are considered as reliable feature points for multimodal images.

According to previous studies, feature points with a uniform distribution reduce the local error of the interimage transformation model and improve image matching accuracy. Conversely, the Shi-Tomasi corner point detector uses image intensity to

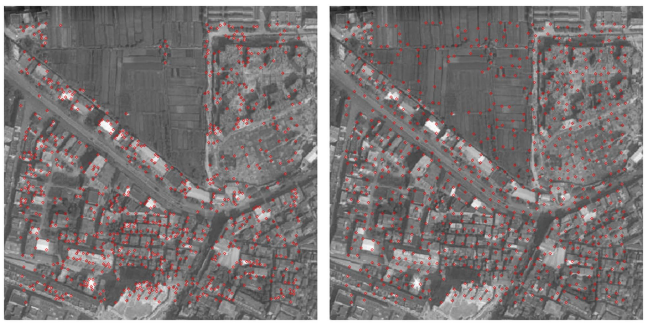

Fig. 3. (a) Effect before homogenization when the feature point is 800. (b) Effect after homogenization when the feature point is 800.

extract feature points; therefore, most feature points extracted using this algorithm are clustered in structural information rich regions. To address the problem of the uneven distribution of feature points, an efficient adaptive nonmaximal suppression algorithm [33] was used, which can quickly divide the extracted feature points evenly and avoid using the chunking method for feature point extraction. This significantly improves the algorithm performance. The suppression region edge length threshold a_h was calculated as follows:

$$a_h = -\frac{H_I + W_I + 2m - \sqrt{\Delta}}{2(m-1)} \tag{3}$$

$$\Delta = 4(W_I + m + H_I m) + (H_I - W_I)^2 + 4W_I H_I m \tag{4}$$

where H_I and W_I denote the height and width of the image, respectively, and m denotes the number of feature points to be extracted.

Therefore, when the input image size changes, the feature points can be redivided according to the image size and the number of feature points to be extracted, making the feature points more evenly distributed, as shown in Fig. 3.

B. Frequency-Domain Convolutional Map

Because of the severe NRD between multimodal images, it is difficult to achieve good results by directly using image information for matching. Therefore, the most multimodal image matching algorithms, such as HOPC, CFOG, and RIFT, feature images using specific methods to produce a feature description map, which increases the similarity between multimodal images and significantly improves the matching performance of the algorithms. As mentioned, compared with other description methods, such as gradient information, PC is more resistant to NRD between multimodal images, and most of the above algorithms are based on the PC model for constructing the feature description graph. However, when there is considerable NRD, the PC model may be inadequate, and the generated feature description map may not be sufficiently clear for feature description or may lack some of the geometric structure. We constructed a new FDCM based on the PC model using a wavelet transform, which can decompose an image into low- and high-frequency parts [34]. The low-frequency part contains the overall geometric structure information of the image, while the high-frequency part contains the local detailed information of the image, and most of the nonlinear radiative aberrations in the image are contained in the local details of the image. After the wavelet transform, the NRD of the image's local details is used to significantly reduce the low-frequency image, which significantly enhances its robustness to NRD. The FDCM generation process is shown in Fig. 4.

The wavelet decomposition is first performed on the input image and is calculated as follows:

$$[\mathbf{C}_N, \mathbf{H}_s, \mathbf{V}_s, \mathbf{D}_s] = WT_N(I) \tag{5}$$

where I denotes the input image, N denotes the scale size in the wavelet decomposition process, s is the number of layers, $\mathbf{C}_N$ denotes the decomposed low-frequency components, and $\mathbf{H}_s$, $\mathbf{V}_s$, and $\mathbf{D}_s$ denote the horizontal, vertical, and diagonal high-frequency wavelet components, respectively.

Good noise suppression and edge extraction drive the superior performance of the feature description maps. A log-Gabor wavelet was used to construct the FDCM. It describes the geometric structure of an image in a multiscale and multidirectional manner. The log-Gabor wavelet is expressed as follows [35]:

$$L(\rho, \theta, s, o) = \exp\left(-\frac{(\rho - \rho_s)^2}{2\sigma_\rho^2}\right) \exp\left(-\frac{(\theta - \theta_{so})^2}{2\sigma_\theta^2}\right) \tag{6}$$

where (ρ, θ) denotes log-polar coordinates, s and o represent the scale and the direction of the log-Gabor wavelet, (ρ_s, θ_{so}) represents the central frequency of the log-Gabor wavelet, and σ_ρ and σ_θ represent the bandwidth of ρ and θ, respectively.

Log-Gabor wavelets can be decomposed into even- and odd-symmetric filters in the spatial domain and are defined as follows [28]:

$$L(x, y, s, o) = L^{\text{even}}(x, y, s, o) + iL^{\text{odd}}(x, y, s, o) \tag{7}$$

where the real part $L^{\text{even}}(x, y, s, o)$ represents the even wavelets, and the imaginary part $L^{\text{odd}}(x, y, s, o)$ represents the odd wavelets.

For the low-frequency part $\mathbf{C}_N$ generated by wavelet decomposition, convolving $\mathbf{C}_N$ with an even-symmetric wavelet $L^{\text{even}}(x, y, s, o)$ and an odd-symmetric wavelet $L^{\text{odd}}(x, y, s, o)$, respectively, produces $E_{so}(x, y)$ and $O_{so}(x, y)$ as follows [11]:

$$E_{so}(x, y) = \mathbf{C}_N(x, y) * L^{\text{even}}(x, y, s, o) \tag{8}$$

$$O_{so}(x, y) = \mathbf{C}_N(x, y) * L^{\text{odd}}(x, y, s, o). \tag{9}$$

Then, the magnitude of the amplitude of the image in the frequency domain is expressed as follows:

$$A_{so}(x, y) = \sqrt{E_{so}(x, y)^2 + O_{so}(x, y)^2}. \tag{10}$$

The multiscale log-Gabor features in the specified directions are

$$A_o(x, y) = \sum_{s=1}^{S} A_{so}(x, y). \tag{11}$$

The feature maps normalized and downscaled to the multidirectional log-Gabor features are then used as the reconstructed

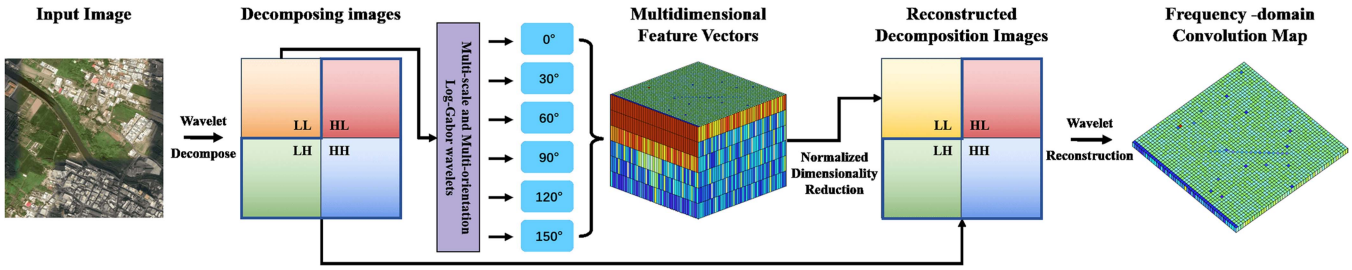

Fig. 4. Proposed framework for generating FDCMs. First, the image is wavelet decomposed, and the decomposed low-frequency image is processed using a multidirectional and multiscale log-Gabor wavelet to generate a multidimensional feature vector. Then, the multidimensional features are dimensionally reduced to decrease the number of dimensions. Finally, wavelet reconstruction is performed using the high-frequency information generated from the decomposition, which generates an FDCM.

low-frequency wavelet components of the following:

$$A_o(x,y) = \frac{A_o(x,y)}{\sqrt{|A_o(x,y)|^2 + \varepsilon}} \tag{12}$$

where ε is a small constant value

$$F(x,y) = \sum_{i=1}^{o} A_o(x,y). \tag{13}$$

Wavelet reconstruction is then performed on the four subimages to generate the final FDCM

$$f(x,y) = (IWT_N(\mathbf{F}, \mathbf{H}_s, \mathbf{V}_s, \mathbf{D}_s)) \tag{14}$$

where $f(x,y)$ is the FDCM, and $IWT_N(\bullet)$ denotes the reconstruction function.

Fig. 5 presents a comparison of the FDCM with other typical feature description maps. The original data comprised an optical map image pair with a significant NRD. For comparison, the two images were manually aligned to eliminate geometric differences. The image gradient feature description map generated using the following equation is shown in Fig. 5. Most algorithms use this approach [36], [37], which results in discontinuous and unstable feature descriptions due to the effect of NRD, making them less robust

$$\begin{cases} G_x(x,y) = \frac{\partial I(x,y)}{\partial x} \\ G_y(x,y) = \frac{\partial I(x,y)}{\partial y} \end{cases} \tag{15}$$

$$\begin{cases} G_\rho = \sqrt{{G_x}^2 + {G_y}^2} \\ G_\varphi = \arctan \frac{G_y}{G_y} \end{cases}. \tag{16}$$

The PC map shown in Fig. 5 describes the geometric information of the image but is susceptible to noise, with blurring and ghosting at the edges. The MIM shown in Fig. 5 is the feature description map of the RIFT algorithm, and it selects the main direction of the local features for a description based on the PC map. Although it is more robust to NRD, the geometric structural information of the image is lost, and the robustness is poor for some poor quality images. The FDCM shown in Fig. 5 is robust to NRD, clearly describes the geometric structure of the image, and performs well on complex and poor quality images. It is used in the MSTM to describe the multiscale local feature information of the feature point matching window for template matching.

C. Omnidirectional Aggregated Feature Vectors

The template-based matching approach has the disadvantage of being poorly rotatable. The strategy used in this study to solve the rotation invariance of template matching is to extract multidirectional feature description maps for one image of the image pair and aggregate these directional feature description maps to construct matching window omnidirectional feature description vectors that adapt to different directions. Multiple directions are matched using the template-matching algorithm, and the direction with the best matching effect is the main direction of rotation.

To extract the matching window feature description vectors in different directions, the original image-generated feature description map FDCM must be rotated n times to achieve full directional coverage, which constitutes the directional feature description map $\widehat{\text{FDCM}} = \{\text{FDCM}^0, \text{FDCM}^1, \ldots, \text{FDCM}^{n-1}\}$. For the r-direction, the rotation transformation equation is expressed as follows:

$$\begin{cases} \text{FDCM}^r = Rot_m(\text{FDCM}^0) \\ m = \frac{2\pi}{n} r, \quad r \in \{0, 1, \ldots, n-1\} \end{cases} \tag{17}$$

where Rot denotes the rotation operation on the image, FDCM_0 is the original nonrotated feature description map, and m is the rotation angle of the r feature description map.

The Shi-Tomasi algorithm can be used to obtain the keypoints for the rotated directional feature description map FDCM_r. The keypoint at position (i,j) in the feature description map FDCM_r can be represented as p_{ij}^r, and the directional feature description vector of the corresponding matching window can be represented as f_{ij}^r. The set of directional feature description vectors of all the feature points in image FDCM_r can be described as $(\widehat{p}^r, \widehat{f}^r)$, which is defined as the r-directional feature description vector of the image. In the aggregation model, the feature description map FDCM aggregates feature keypoints and matches the window omnidirectional feature description vector $(\widehat{p}, \widehat{f})$ for the feature description map FDCM.

After obtaining the matching point pairs in each direction, the number of matching points $score^r$ obtained in the r-direction

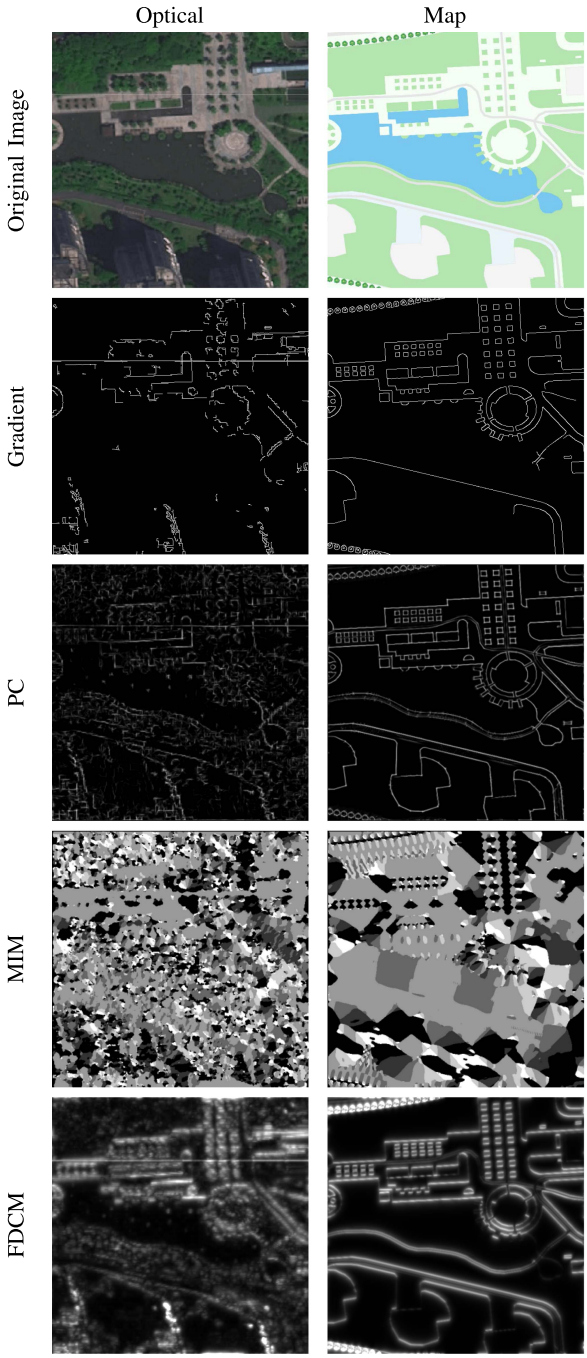

Fig. 5. Comparison of multiple feature description maps of the selected optical-map image pairs, including image gradient, PC, MIM, and FDCM.

was recorded as the matching score. The direction with the highest matching score $score^r$ was determined as the main feature direction r^{main}, and the formula is as follows:

$$r^{\text{main}} = \arg\max_{r\in\{0,1,\ldots,n-1\}} score^r. \tag{18}$$

Because the rotational differences of the matched images are typically not strictly in the main direction r^{main}, correct matches may occur in the main, front, and back directions. Therefore, the proposed algorithm merges the matching results of adjacent direction features with the main direction to obtain the best matching result.

Instead of the position on the original feature description map, the keypoint position in the matching result is the coordinate on the rotated feature description map, FDCM^r. Therefore, it must still be inversely rotated and transformed to map back to the coordinate on the original feature description map

$$\begin{cases} \widehat{p}_0^r = Rot_m^{-1}(\widehat{p}^r) \\ m = \frac{2\pi}{n} r, \quad r \in \{0, 1, \ldots, n-1\} \end{cases} \tag{19}$$

where $\widehat{p}^r$ is the directional feature keypoint after rotation, Rot^{-1} is the rotational transformation inverse to the feature description map FDCM^r, and the output result $\widehat{p}_0^r$ is the coordinate on the original feature description map.

D. Fast Frequency-Domain Template Matching

After extracting the feature points and generating the feature description map, the template-based matching strategy requires selecting a suitable similarity metric and performing template matching using the feature point positions in the search region of the target image to determine the corresponding points with a similarity greater than a threshold value as the correct matching points. To increase the matching speed of the algorithm, we adopted a concept similar to that in [30], i.e., that of converting the SSD-based similarity metric to the frequency domain and using a fast Fourier transform (FFT) for template matching.

The matching image pairs consisted of two images, and their corresponding 2-D feature description maps [frequency-domain convolutional maps (FDCMs)] are denoted by M_1 and M_2. The SSD between the corresponding template windows in the feature description map is defined as follows:

$$T_i(s) = \sum_p [M_1(p) - M_2(p-s)]^2 \tag{20}$$

where p denotes the position representation of the pixel in the 2-D feature description map, and $T_i(s)$ denotes the SSD similarity metric function between two template windows after the template window on M_1 is displaced by s at the corresponding position on M_2. The best match between the template windows on M_1 and M_2 is achieved when T reaches its minimum value. The matching function is defined as follows:

$$s_i = \arg\min_s \left\{ \sum_p [M_1(p) - M_2(p-s)]^2 \right\} \tag{21}$$

where s_i denotes the displacement vector between M_1 and M_2 when the SSD similarity metric function of the corresponding template window reaches its minimum value.

Expanding and simplifying the above equation, the displacement vector s is only related to the parameters of the following equation:

$$s_i = \arg\max_s \left[\sum_p M_1(p) \cdot M_2(p-s) \right]. \tag{22}$$

The convolution operation in the spatial domain is equal to multiplication in the frequency domain. We used the 2-D FFT

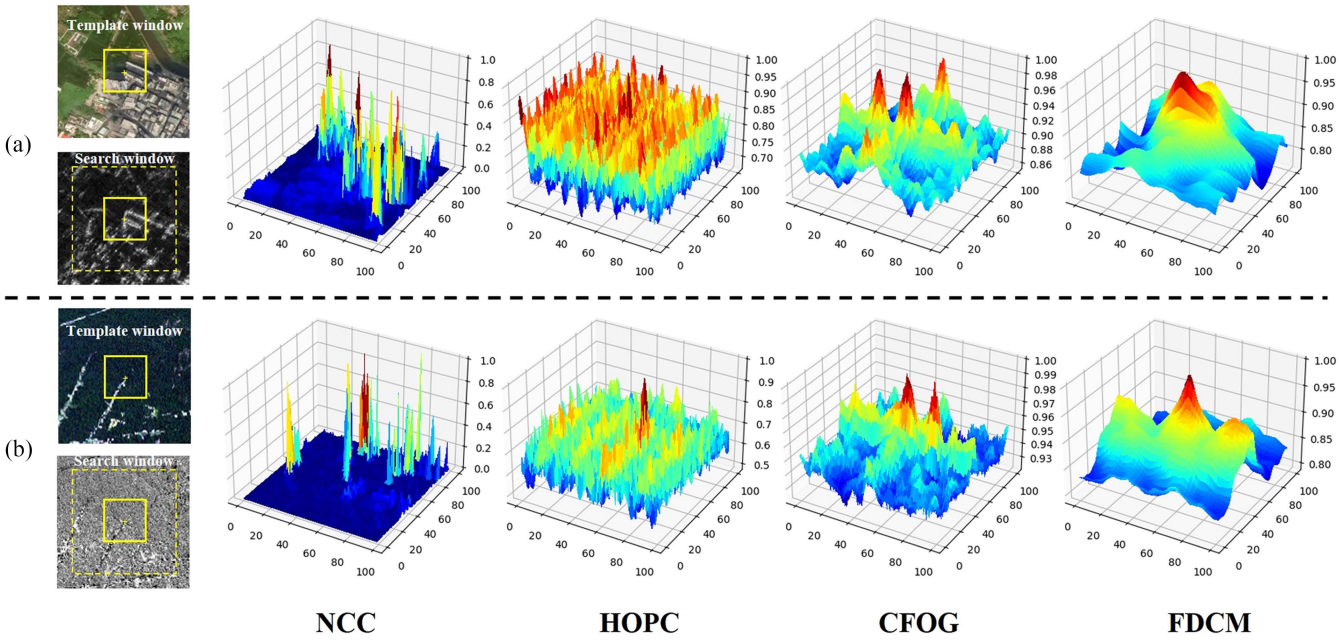

Fig. 6. Feature similarity maps. The similarity maps of NCC, HOPC (Reproduction version), CFOG, and the proposed FDCM were obtained for the (a) optical-SAR and (b) optical-infrared images with a template window of 70 × 70 pixels and a search window of 30 × 30 pixels. In each plot, the center of the plot corresponds to the correct matching position.

to accelerate the spatial-domain convolution calculation. The above equation can be defined as follows:

$$s_i = \arg\max_{s} \left\{ F_{2\mathrm{D}}^{-1} \left[F_{2\mathrm{D}} \left(M_1(p) \right) \cdot F_{2\mathrm{D}}^{*} \left(M_2(p-s) \right) \right] \right\} \quad (23)$$

where $F_{2\mathrm{D}}$ and $F_{2\mathrm{D}}^{-1}$ denote the forward and reverse 2-D FFT, respectively, and $F_{2\mathrm{D}}^{*}$ represents the complex conjugate of the 2-D FFT.

Fig. 6 shows the similarity plots of various similarity metrics. The test images were optical-SAR (a) and optical-infrared (b), and the similarity metrics used were NCC, HOPC, CFOG, and FDCM. Of all the metrics, the NCC method failed to determine a correct match. Both CFOG and FDCM could accurately detect the corresponding region, but CFOG had interference from other peaks and suffered from mismatching, and the peaks of FDCM were smoother and more robust.

E. Outlier Removal

Template matching was performed on the extracted feature points in the feature description map following the above principle, and numerous matching results were obtained. However, many outliers in the matching results must still be eliminated; otherwise, parameter estimation of the transform model will be negatively affected. Therefore, we adopted a more accurate MAGSAC++ [38] method to eliminate outliers.

F. Multiscale Matching Strategy

It is well known that the most challenging aspect of the template-based matching method is determining the correct template matching search area corresponding to the feature points. For a feature point in the reference image, we looped through the entire image search to be matched. Although it is possible to determine the correct corresponding search area this way, the process is time consuming. In particular, when faced with many feature points, the search time increases exponentially, significantly reducing the matching performance. Moreover, when processing two images with inconsistent scales, the local feature content described within the template window is inconsistent even when the template matching windows correspond correctly, leading to unreliable matching results or matching failure. A multiscale template matching strategy was designed to solve these two key problems, which eliminates the dependence on the initial control points and achieves robustness to scale.

A Gaussian pyramid of image feature description maps was adopted using a scale-space theory [39]. As shown in Fig. 7, since generating FDCM feature description maps for different scale images separately would consume a lot of time, and the original images are not high-dimensional features to generate FDCM feature description maps first, they are directly downsampled layer by layer to obtain a series of feature description maps with different resolutions.

After establishing the feature description scale space, template matching on the low-level feature map using a template with a larger window is time consuming, and it is difficult to search for the correct corresponding region using a template with a smaller window. However, template matching with a larger window on a high-level feature map can reduce the matching time and improve the matching success rate (SR). Although the location of the correctly matched search points may not be very accurate, a more reliable transformation model can still be estimated using the remaining points after MAGSAC++ eliminated the outliers. The model is passed to the next level to

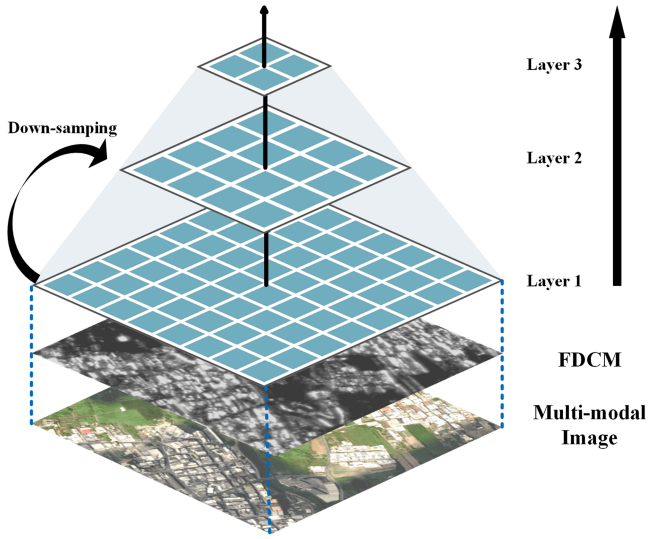

Fig. 7. Scale space pyramid designed in MSTM.

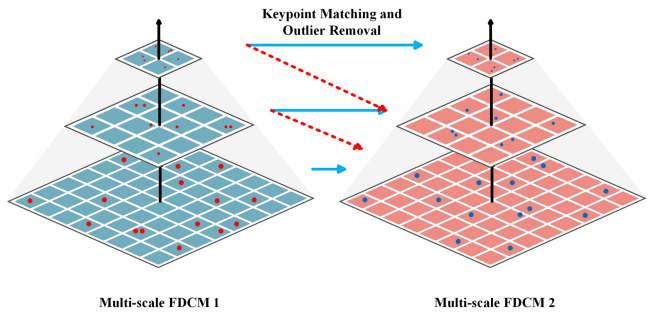

Fig. 8. Multiscale template matching strategy in MSTM.

continue template matching, and with the initial transformation model, the template matching search window can be appropriately narrowed, further reducing the search time and improving the matching accuracy. This continues until it reaches the lowest level, where the entire multiscale matching is complete. When there is scaling, the layers of the reference image and the image to be matched may not be the same. First, the highest level of the reference image is matched with all the levels of the image to be matched by templates separately to determine the correct corresponding level. Then, the entire multiscale matching is completed by passing them in order according to the conventional matching process. The complete matching process is shown in Fig. 8.

The resultant feature description scale space is similar to the classical and widely used Gaussian scale space in SIFT. The main difference is that SIFT provides scale attribute information for feature points based on the scale space and uses it for feature extraction and multiscale feature description. Finally, all the scale features participate in matching simultaneously. The scale space in this study does not involve scale information in feature matching but is used to search for the corresponding layers, which more effectively solves the problem of scale differences.

IV. Experiments

A. Datasets and Evaluation Indicators

In order to verify the matching performance of this framework, several sets of typical multisource remote sensing image data are selected, and various effective matching algorithms, such as CFOG, SuperGlue [40], RIFT, HAPCG, and Ms-HLMO [41], are compared. The experimental computer is a Lenovo Y9000K notebook with i7-10875H CPU, GeForce RTX 2080 graphics card (8-GB video memory), 64-GB RAM, and Windows 10 64-bit.

To ensure that the comparison experiments were fair, we used the default optimal parameters for the parameter settings of various algorithms. The SuperGlue, RIFT, HAPCG, and Ms-HLMO parameters were consistent with the original published versions. CFOG cannot be applied to multimodal data, which lacked known information because the published version requires known initial control point information; therefore, we removed this part and kept the rest unchanged.

We collected six real multimodal image datasets for our experiments: optical-depth, optical-infrared, optical-map, optical–optical, optical-SAR, and day–night datasets. The dataset image sizes ranged from 400×400 to 1000×1000 pixels. The optical–optical dataset included different seasons, views, and other elements. These multimodal remote sensing images covered almost all the multimodal image matching scenarios, and each dataset contained ten image pairs, for a total of 60 pairs of multimodal image data. Most images contained a significant NRD, which was highly representative and experimentally valuable for fully validating the performance of multimodal matching algorithms. Fig. 9 shows several sample pairs for each dataset.

The images of the six datasets were first separately matched with features, and a correspondence with an error of less than three pixels was considered the correct matching relationship [11]. In this study, the number of correct matches (NCM), SR, and root-mean-square error (RMSE) were used as evaluation metrics. These metrics were calculated as follows:

$$\mathrm{NCM} = \left| \left\{ \left\| p_i^1 - \mathbf{H} p_i^2 \right\| < 3 \right\}_{i=1}^{N} \right| \tag{24}$$

where p_i^1 and p_i^2 are the coordinates of a matched pair of points, H is the ground truth spatial transformation between image pairs (calculated from manually selected points), and N is the total number of all the matches

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{NCM}} \sum_{i=1}^{\mathrm{NCM}} \left[\left(x_2^{i'} - x_2^i \right)^2 + \left(y_2^{i'} - y_2^i \right)^2 \right]} \tag{25}$$

where NCM is the number of correct matches, $(x_2^{i'}, y_2^{i'})$ is calculated from $(x_1^i, y_1^i,)$ on the reference image by the transformation matrix H, and $(x_2^i, y_2^i,)$ is the true coordinates of the feature points on the image to be matched

$$\left[x_2^{i'}, y_2^{i'}, 1 \right]^T = H \cdot \left[x_1^i, y_1^i, 1 \right]^T \tag{26}$$

$$\mathrm{SR} = \frac{M_{\mathrm{correct}}}{M_{\mathrm{correct}} + M_{\mathrm{error}}} \times 100\% \tag{27}$$

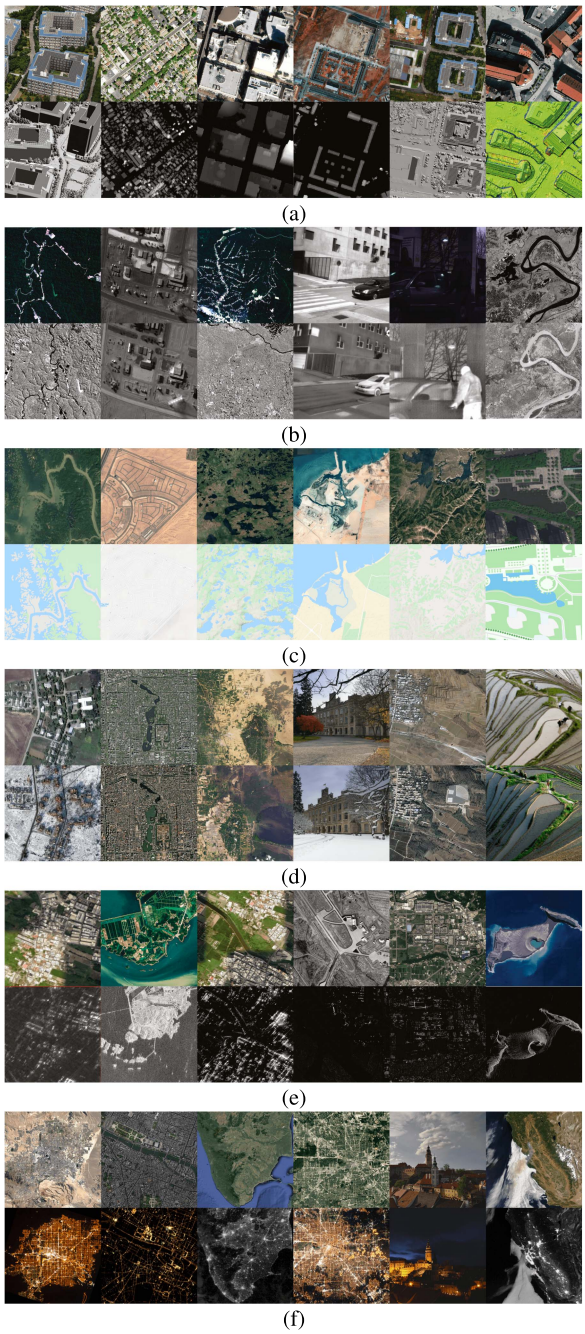

Fig. 9. Sample data of our four multimodal image datasets. (a) Dataset 1: optical-depth. (b) Dataset 2: optical-infrared. (c) Dataset 3: optical-map. (d) Dataset 4: optical–optical. (e) Dataset 5: optical-SAR. (f) Dataset 6: day–night.

where SR represents the matching SR, M represents the total number of image pairs of a multimodal image sets, M_{correct} represents the number of successfully matched pairs of images, and M_{error} represents the number of failed pairs of images.

B. Parameter Setting

The main parameters of the proposed MSTM algorithm included the template sizes of the highest and lowest levels and the number of rotation directions in the omnidirectional aggregated feature vector. To determine the template size for each level and select the best parameters, a comparison experiment was conducted, and the results are shown in Fig. 10. According to Fig. 10(a), the matching SR was higher at the highest-level template size of 200 pixels and at the lowest-level template size larger than 100 pixels. As shown in Fig. 10(b), the matching time increased exponentially as the template size increased. Using this analysis, we determined the highest-level template size J1 as 200 pixels and the lowest-level template size J2 as 100 pixels. When a difficult-to-match image was encountered, the lowest template size can be increased to 200 pixels.

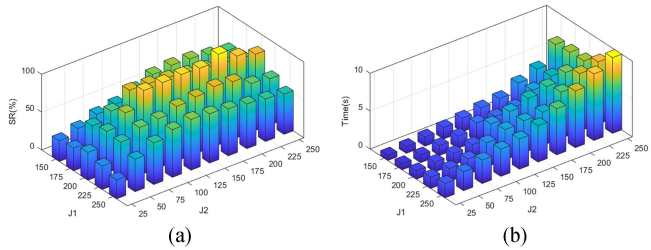

Fig. 10. (a) Matching SR statistics for different template sizes at each level. (b) Matching time consumption statistics for different template sizes at each level.

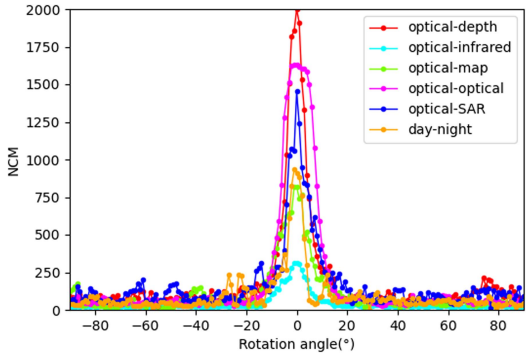

Fig. 11. Statistics on the NCM on multiple types of rotated image pairs between -90° and 90° for the MSTM algorithm without rotation invariance.

To determine the number of directions in the all-directional aggregated feature vector of the MSTM algorithm, we selected a random set of image pairs in each of the six sets of data for the test experiment. The results are shown in Fig. 11. One of the images remained unchanged as we rotated the other image from -90° to 90° to calculate the number of correct matching points after the rotation. According to Fig. 11, the MSTM algorithm without rotation invariance can adapt to the rotated image pairs above and below $\pm 15^\circ$. The number of directions n was set to 24. Theoretically, this parameter can be rotated in all the directions.

C. Invariance Tests

Determining whether the template matching based on the constructed feature description map is robust to NRD is important for template-based multimodal image matching. According to the analysis, rotation and scale differences are

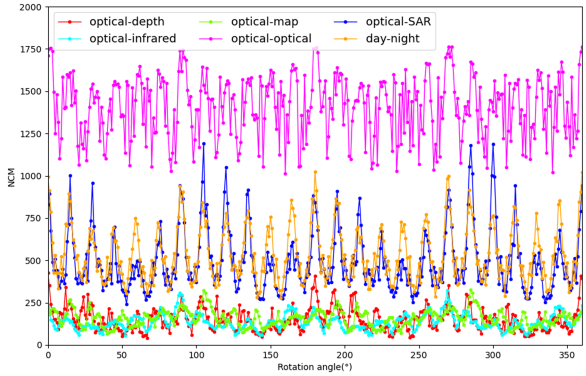

Fig. 12. NCMs of MSTM on different types of multisource remote sensing scenes as the rotation angles from 0° to 359°.

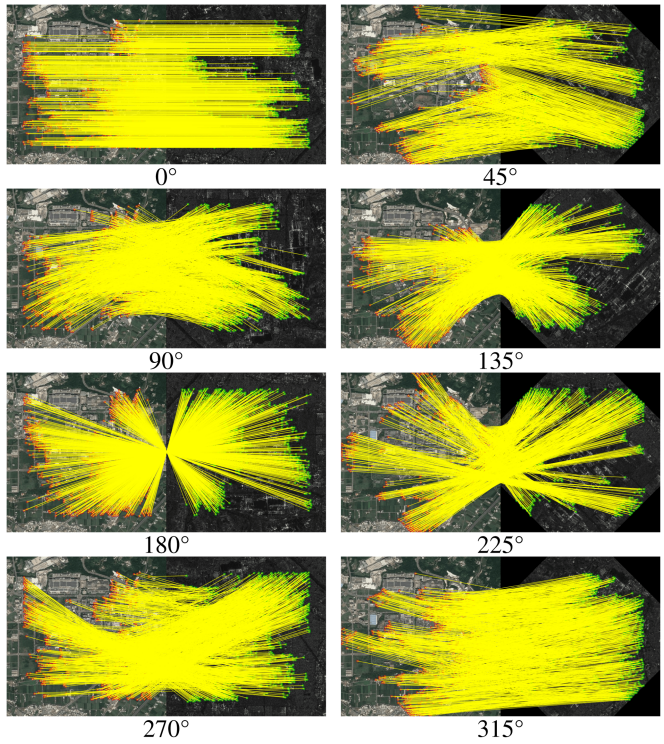

Fig. 13. Some typical visualization results of Fig. 12. The angle below an image represents the rotation angle between the image pairs.

the most common differences; therefore, we designed experiments to test whether the algorithm developed in this study has better invariance from these two aspects and whether the algorithm can handle various multimodal image data.

1) Rotation Invariance: For image pairs without rotation differences in the above dataset, one image was fixed, while the other was rotated to test the algorithm's ability to adapt well to the rotation. The rotation angle ranged from 0° to 359° at 1° intervals, and 360 image pairs were generated for each pair. The matching algorithm was then executed for each image pair, and the NCMs of the corresponding datasets are plotted in Fig. 12. The matching results for different rotation angles are shown in Fig. 13. According to Fig. 12, the NCMs of all six scenes vary with the rotation angle; however, they are stable within a certain interval. No matching failure occurred at any angle, which demonstrates that the algorithm proposed in this study has exceptional rotational adaptability. As shown in Fig. 13, the matching results of this algorithm are sufficient to accomplish other functions such as image alignment for image pairs with different rotation angles.

2) Scale Invariance: To test whether the algorithm adapts well to images with certain scale differences, the same strategy was used to fix one image in the image pair and scale the other image. The scaling ratio was 1:2, and the interval was 0.2. Five pairs of image pairs were generated for each pair of images. The matching results are shown in Fig. 14. The algorithm could still obtain accurate matching results even if the images had scale differences. This is because of the multiscale matching strategy used in the algorithm, which enhances the matching performance for images with scale differences and can adapt to multimodal image data with certain scale differences.

D. Matching Performance Test

1) Qualitative Evaluations: We selected one image pair from each of the six multimodal datasets for testing, as illustrated in Fig. 15. Among them, Fig. 15(a) and (e) contains translational and small rotation variations, Fig. 15(b) and (d) contains only translational variations, and Fig. 15(c) and (f) shows translational, small rotation, and scale variations. Because these image pairs are multimodal image data from different imaging devices, there is a significant NRD. Therefore, performing tests on these image pairs is difficult. The matching results for CFOG, Ms-HLMO, HAPCG, RIFT, SuperGlue, and the proposed MSTM are plotted in Fig. 15.

The results showed that the CFOG algorithm could not successfully match all the images, and its SR accuracy was 0. Although the CFOG algorithm provides a pixel-by-pixel feature description with powerful feature description capability, its requirement for initial reference control points leads to drastic performance degradation in images with displacement, rotation, and scale differences. The Ms-HLMO algorithm could not match the second and third image pairs, and its SR accuracy was 50%. The HAPCG algorithm could not match the first, fifth, and sixth image pairs, and its SR accuracy was 50%. Although the two image pairs failed to match, the performance of the successfully matched image pairs was satisfactory. This is because the HAPCG algorithm constructs an anisotropic scale space and uses log-polar coordinates for the feature description, which is more robust. The SuperGlue algorithm could not match the third and fourth image pairs, achieving an SR accuracy of 66.6%. Although the SuperGlue algorithm uses a deep learning approach that combines an attention mechanism and a graph neural network, it has strong robustness for multimodal data. However, because the training optical-depth dataset does not contain multimodal data, the matching performance degrades

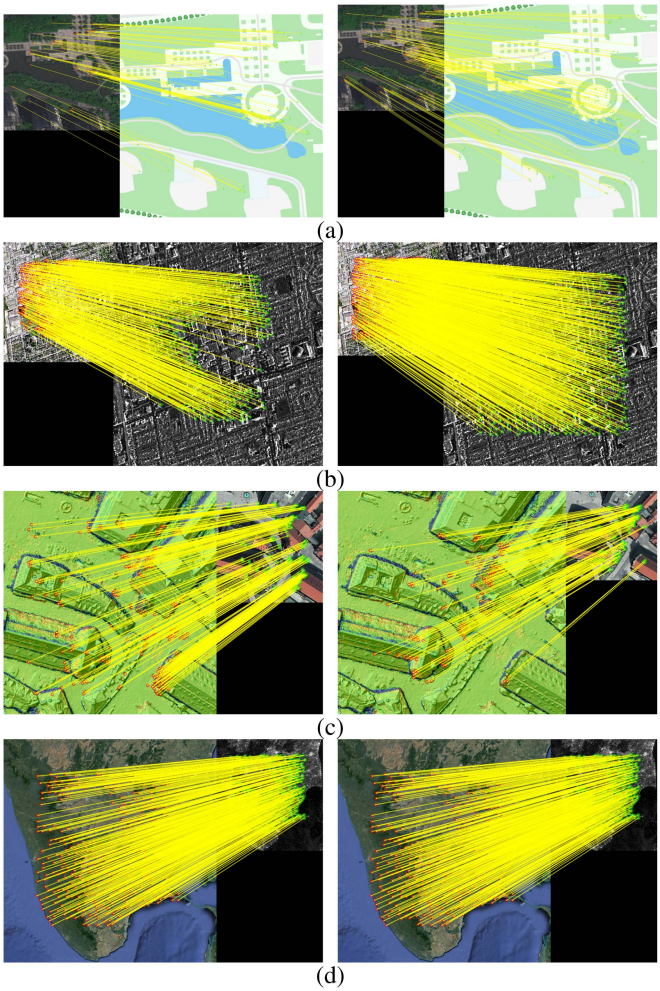

Fig. 14. (a)–(d) Examples of feature matching results of scene with scale differences. Matching results by MSTM with scale ratio of 1:2.

for some multimodal datasets. The RIFT algorithm could not match the third image pair, achieving an SR accuracy of 83.3%. Because the RIFT algorithm designs the MIM map based on the PC map, it is more robust to NRD. In addition, it analyzes the effect of rotation on the MIM and achieves the rotation invariance of the feature description, which improves the algorithm performance. Conversely, the MSTM algorithm proposed in this study successfully matched all six image pairs with 100% SR accuracy.

The MSTM outperformed the popular algorithms in terms of matching on multimodal data because: 1) MSTM designs the FDCM feature description map, which has better feature robustness than other algorithms and lays an important foundation for subsequent matching, and 2) a multiscale matching strategy was adopted, which makes the algorithm robust to multimodal images and guarantees matching accuracy.

2) Quantitative Evaluations: Fig. 16 presents the quantitative results for the NCMs and the results of the six methods on the six multimodal datasets. The results demonstrate that the CFOG algorithm outperforms the other methods on the optical-depth dataset. This may be caused by the small difference in displacement and scale between the depth and optical images, making it possible to successfully match without the initial reference control point. The Ms-HLMO algorithm performs relatively well on all the datasets, and its stability is significantly enhanced because of its construction of PMOM, the use of GGLOH descriptors, and the adoption of a multiscale matching strategy. The SuperGlue algorithm outperformed the other three algorithms on the optical-depth dataset, probably because of its use of a combination of RGB images and depth data during training. However, its performance on the optical-SAR dataset is not much different from the other three algorithms because there is a significant NRD between the SAR and optical images, which makes the matching considerably more difficult. Because the gap between this dataset and the training dataset of SuperGlue is the largest, the training parameters of SuperGlue cannot be adapted to this dataset. The HAPCG algorithm outperformed all the algorithms on some datasets mainly because it uses anisotropic filtering to nonlinearly diffuse the images and constructs an anisotropic weighted moment scale space based on this. Then, the PC model is extended to establish an absolute PC directional gradient and is combined with a log-polar coordinate description template to establish a kind of absolute phase directional gradient histogram, which significantly enhances the robustness of the descriptor. The relatively balanced performance of the RIFT algorithm on all the datasets is because the RIFT algorithm designs the MIM map based on the PC map, which makes it more robust to NRD. In addition, it achieved rotational invariance of the feature description, which improved the algorithm performance. Conversely, the proposed MSTM successfully matched all the images in the six datasets, and the NCMs for almost all the image pairs were significantly larger than 250. The matching performance of this algorithm is more stable and robust, with better adaptability to NRD, and it outperformed the other algorithms.

TABLE I
COMPARISONS ON SR METRIC

Method	Optical-depth	Optical-infrared	Optical-map	Optical–optical	Optical-SAR	Day–night
CFOG	90	70	40	80	30	50
Ms-HLMO	70	90	100	100	50	70
HAPCG	70	100	90	100	100	60
SuperGlue	100	100	90	90	70	80
RIFT	100	100	100	100	90	100
MSTM	100	100	100	100	100	100

As shown in Table I, the CFOG algorithm performs the best on the optical-depth dataset with 90% SR, the Ms-HLMO algorithm performs the worst on the optical-SAR dataset with only 50% SR, the HAPCG algorithm achieves 100% SR on several datasets, the SuperGlue algorithm performs the worst on the optical-SAR dataset with only 70% SR, which is not as good as the traditional algorithm, and the RIFT algorithm performs well on all the datasets and can match successfully. The SuperGlue algorithm had the worst performance on the optical-SAR dataset, with only 70% SR, which is inferior to the

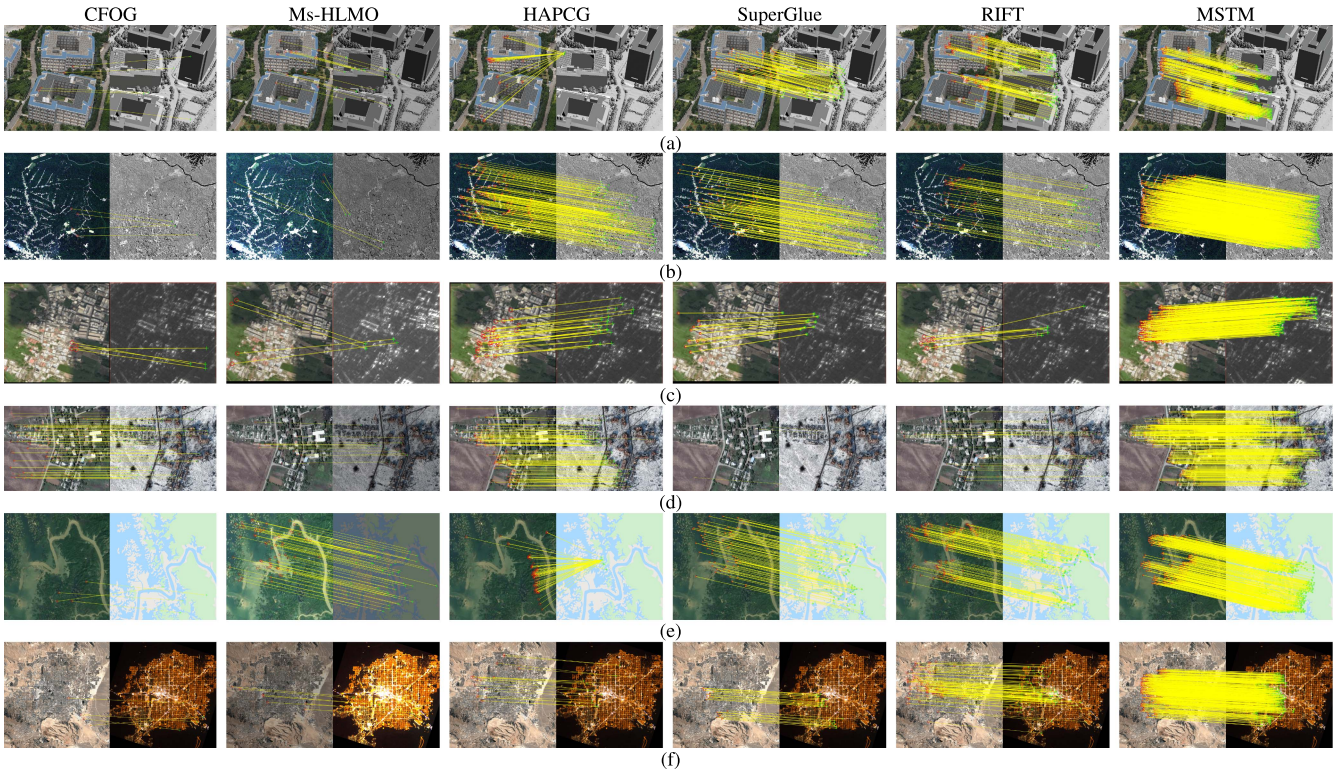

Fig. 15. Qualitative comparison results on the sample data. The red and green circles in the figure indicate the feature points on the reference and target images, respectively. (a) Matching results of optical-depth datasets. (b) Matching results of optical-infrared datasets. (c) Matching results of optical-SAR datasets. (d) Matching results of optical–optical datasets. (e) Matching results of optical-map datasets. (f) Matching results of day–night datasets.

traditional algorithm; RIFT had a balanced performance on all the datasets and could match successfully; the MSTM algorithm had 100% SR on all the datasets. The average SRs of CFOG, Ms-HLMO, HAPCG, SuperGlue, and RIFT for the six datasets were 60%, 80%, 86.6%, 88.3%, and 98.3%, respectively. Compared with the CFOG algorithm, the performance of the MSTM algorithm was improved by 40%. Fig. 17 and Table II show the RMSEs of the six methods for all the datasets. Only the RMSEs of successfully matched images were counted because of the low SR of some algorithms on some datasets. In terms of RMSE, the accuracy of feature-based matching methods was slightly lower, whereas the accuracy of both the HAPCG and RIFT algorithms was lower than that of the CFOG and MSTM algorithms. The CFOG algorithm, which is the multimodal matching algorithm with the highest accuracy, maintained the same RMSE accuracy as CFOG compared with the MSTM algorithm and is slightly higher than the CFOG algorithm on some datasets. This demonstrates that the MSTM algorithm had the same RMSE accuracy as that of CFOG and slightly higher than that of CFOG in some datasets, showing better performance.

Combining the above qualitative and quantitative test results, we can conclude that each part of the MSTM algorithm is designed for NRD, including feature point extraction, a feature description map, template matching, omnidirectional aggregated feature vectors, and a multiscale matching strategy. Therefore, the algorithm has good adaptability to NRD. Exceptional NCM and RMSE accuracies were obtained for all six datasets, exceeding those of the other algorithms. This implies that the proposed algorithm is a superior multimodal matching algorithm.

E. Registration Performance

In this section, we apply MSTM to image registration and fusion. After image matching using MSTM and calculating the transformation matrix, the two images were mapped to the same reference frame to generate corrected images. The tests were performed on six datasets, and the visualization results are shown in Fig. 18. All the images were well aligned and fused without displacement. The ghosting and blurring cases further demonstrate that the MSTM algorithm achieves matching with high accuracy and good distribution.

F. Running Time Analysis

Table III lists the average running time of each algorithm on all the datasets. CFOG, Ms-HLMO, HAPCG, and RIFT were computed using MATLAB 2021a, while SuperGlue and MSTM were computed using Python. The graph shows that MSTM runs slightly slower than CFOG and SuperGlue and is one-fifth of the running time of HAPCG and RIFT. According to the analysis of the algorithm principle, the HAPCG algorithm consumes considerable time when constructing the nonlinear diffusion scale space and generates high-dimensional feature vectors, which requires a significant amount of time for subsequent

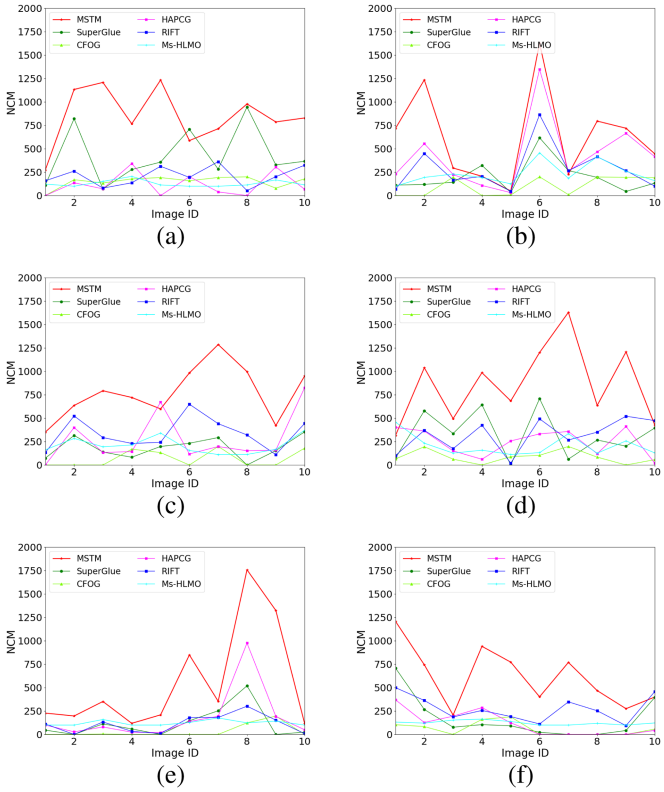

Fig. 16. (a)–(f) Comparisons on the NCM metrics.

TABLE II
COMPARISON RESULTS OF THE MATCHING METHODS

Metric	Method	Optical-depth	Optical-infrared	Optical-map	Optical–optical	Optical-SAR	Day–night
NCM	CFOG	148.2	99.7	68	85.4	32.5	59.4
	Ms-HLMO	128.5	231.5	211.3	206.7	123.1	125.1
	HAPCG	116.4	430.1	280	247.4	180.3	114.7
	SuperGlue	426.4	198.3	184.1	326.8	116.2	171.1
	RIFT	208.3	284.3	338.8	319.7	111.4	276.3
	MSTM	850.6	631.6	773.8	861.7	550.6	618.6
RMSE	CFOG	1.15	1.31	2.06	1.4	2.35	1.08
	Ms-HLMO	1.1	1.32	1.22	1.19	1.23	1.18
	HAPCG	2.21	1.85	1.94	1.88	1.94	1.87
	SuperGlue	1.26	1.25	1.46	1.43	1.8	1.17
	RIFT	1.31	1.29	1.33	1.24	1.41	1.33
	MSTM	0.94	0.89	0.94	0.94	1.19	1.07
Time(s)	CFOG	2.03	1.85	1.99	2.04	1.96	2.4
	Ms-HLMO	14.51	16.77	16.53	17.98	21.75	16.22
	HAPCG	7.56	6.81	7.51	10.59	13.93	8.61
	SuperGlue	2.95	2.86	2.85	2.94	2.88	2.85
	RIFT	18.21	6.26	8.04	8.66	7.07	8.38
	MSTM	3.7	3.28	3.69	3.97	3.84	3.8

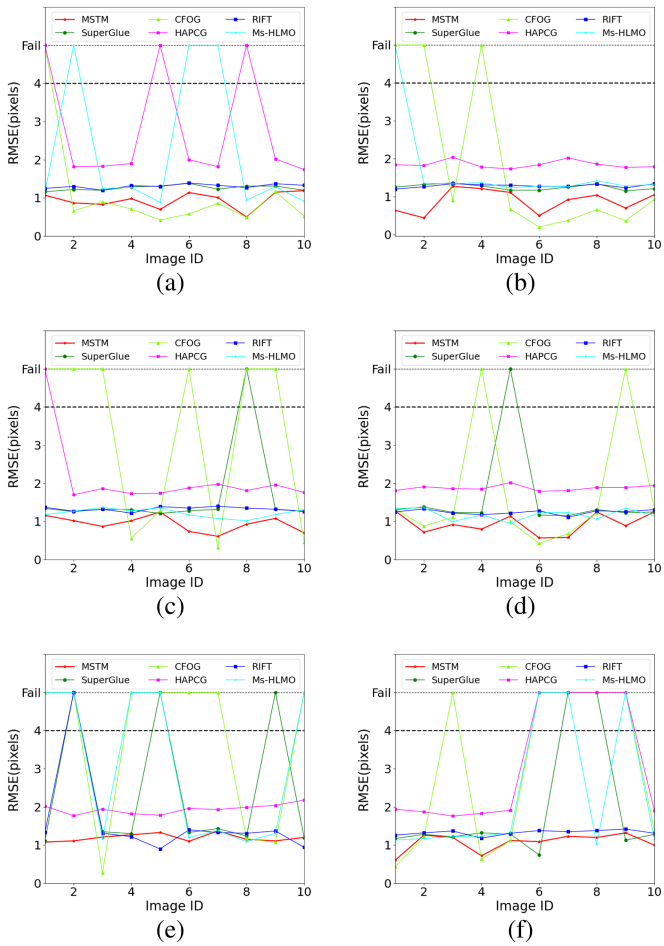

Fig. 17. (a)–(f) Comparisons of the RMSE metrics. Note that the matching failure cases are drawn onto the failed line.

TABLE III
COMPARISONS OF THE TIME METRIC

Method	CFOG (Matlab)	Ms-HLMO (Matlab)	HAPCG (Matlab)	SuperGlue (Python)	RIFT (Matlab)	MSTM (Python)
Time/(s)	2.05	17.29	9.17	2.89	9.44	3.71

matching. The RIFT algorithm mainly consumes running time when constructing the MIM graph and achieves the rotation invariance of the MIM graph. Ms-HLMO consumes the most time because the MSTM algorithm uses a multiscale matching strategy, which makes its running time much longer than that of the other algorithms. However, because the MSTM algorithm reduces the dimensionality of the multidimensional feature description vector when constructing the feature description map, it significantly reduces the search time of the template matching window. Then, the matching result information searched in the upper pyramid can be shared with the lower pyramid, and the mapping relationship can be approximately obtained, which also saves considerable time. The MSTM with rotation invariance requires more time to generate the omnidirectional aggregated feature vector than CFOG and SuperGlue.

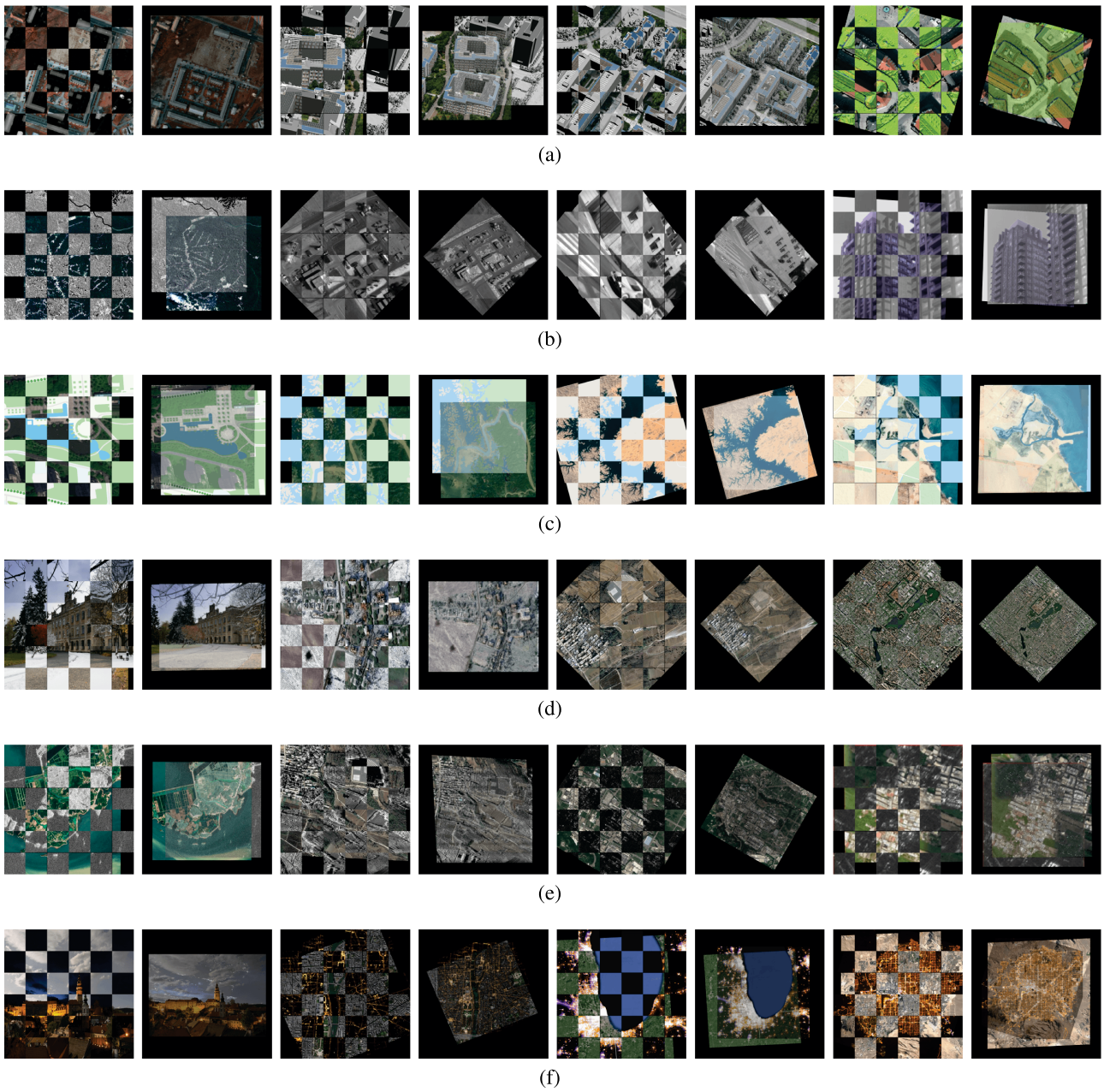

Fig. 18. Image registration and fusion results of MSTM on the multimodal image pairs. (a) Registration results of optical-depth datasets. (b) Registration results of optical-infrared datasets. (c) Registration results of optical-map datasets. (d) Registration results of optical–optical datasets. (e) Registration results of optical-SAR datasets. (f) Registration results of day–night datasets.

V. Conclusion

In this study, a stencil matching method we call MSTM with better robustness to NRD was proposed. The method adapted to multimodal image data with displacement and scale transformation better. The initial motivation of MSTM originated from the advantages of stencil matching in multimodal matching. After analyzing and summarizing the advantages and disadvantages of the existing mainstream methods, the MSTM algorithm was described in detail. We developed a novel FDCM with better robustness to NRD feature description and rotation-invariant and multiscale matching strategies for stencil matching, which significantly improved the adaptability of stencil matching to multimodal image data. Qualitative and quantitative experiments were conducted to verify the reliability and superiority of the MSTM algorithm.

At the same time, the MSTM algorithm is not particularly effective on images with both the rotation and zoom states, and in the future, we will look at how to make MSTM work better in this regard. We intend to develop a direction index map based on the FDCM to achieve rotation invariance, which can reduce the time required for matching. In addition, template matching will lead to the situation that some areas at the edges cannot participate in matching, and we will try other ways to solve this

problem and make the edge areas match better. The algorithm will play an important role in multimodal image fusion and land classification.

REFERENCES

[1] Y. Wu, J. Li, Y. Yuan, A. Qin, Q.-G. Miao, and M.-G. Gong, "Commonality autoencoder: Learning common features for change detection from heterogeneous images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4257–4270, Sep. 2022.

[2] S. Song, K. Jin, B. Zuo, and J. Yang, "A novel change detection method combined with registration for SAR images," *Remote Sens. Lett.*, vol. 10, no. 7, pp. 669–678, 2019.

[3] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[4] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.

[5] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[6] A. Brown, M. Walter, and T. Cudahy, "Hyperspectral imaging spectroscopy of a Mars analogue environment at the North Pole Dome, Pilbara Craton, Western Australia," *Aust. J. Earth Sci.*, vol. 52, no. 3, pp. 353–364, 2005, doi: 10.1080/08120090500134530.

[7] B. Zhang, H. Yang, and Z. Yin, "A region-based normalized cross correlation algorithm for the vision-based positioning of elongated IC chips," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 3, pp. 345–352, Aug. 2015.

[8] H. Zhang and R. Xu, "Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the Pearl River Delta," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 87–95, 2018.

[9] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 178, pp. 68–80, 2021.

[10] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, 2021.

[11] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2019.

[12] M. A. Treiber, *An Introduction to Object Recognition: Selected Algorithms for a Wide Variety of Applications*. New York, NY, USA: Springer, 2010.

[13] Y. Wang, Y. Y. Tang, and L. Li, "Correntropy matching pursuit with application to robust digit and face recognition," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1354–1366, Jun. 2017.

[14] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for SAR and optical images using multiscale convolutional gradient features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4017605.

[15] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[16] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, and Y. Ye, "R$_2$FD$_2$: Fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606115.

[17] L. Yu, D. Zhang, and E.-J. Holden, "A fast and fully automatic registration approach based on point features for multi-source remote-sensing images," *Comput. Geosci.*, vol. 34, no. 7, pp. 838–848, 2008.

[18] H. Li, B. Manjunath, and S. K. Mitra, "A contour-based approach to multisensor image registration," *IEEE Trans. image Process.*, vol. 4, no. 3, pp. 320–334, Mar. 1995.

[19] H. Goncalves, L. Corte-Real, and J. A. Goncalves, "Automatic image registration through image segmentation and sift," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 7, pp. 2589–2600, Jul. 2011.

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

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

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

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

[24] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[25] M. L. Uss, B. Vozel, V. V. Lukin, and K. Chehdi, "Multimodal remote sensing image registration with accuracy estimation at local and global scales," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6587–6605, Nov. 2016.

[26] J. Ma, J. C.-W. Chan, and F. Canters, "Fully automatic subpixel image registration of multiangle CHRIS/Proba data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2829–2839, Jul. 2010.

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

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

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[30] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[31] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–152.

[32] J. Shi and Tomasi, "Good features to track," in *Proc. Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 593–600.

[33] O. Bailo, F. Rameau, K. Joo, J. Park, O. Bogdan, and I. S. Kweon, "Efficient adaptive non-maximal suppression algorithms for homogeneous spatial keypoint distribution," *Pattern Recognit. Lett.*, vol. 106, pp. 53–60, 2018.

[34] Y. W. F. M. Wei Jianlong and T. Zheng, "Image multiscale registration in wavelet domain using SURF," *Comput. Eng. Appl.*, vol. 50, no. 2, pp. 200–204, 2014.

[35] S. Fischer, F. Šroubek, L. Perrinet, R. Redondo, and G. Cristóbal, "Self-invertible 2D log-Gabor wavelets," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 231–246, 2007.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[37] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004.

[38] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "MAGSAC++, A fast, reliable and accurate robust estimator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1304–1312.

[39] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *J. Appl. Statist.*, vol. 21, nos. 1/2, pp. 225–270, 1994.

[40] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.

[41] C. Gao, W. Li, R. Tao, and Q. Du, "MS-HLMO: Multiscale histogram of local main orientation for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626714.

Tian Gao is working toward the master's degree in photogrammetry and remote sensing with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China.

His research interests include remote sensing image processing and unmanned aerial vehicle navigation.

Chaozhen Lan received the B.S. and M.S. degrees in photogrammetry and remote sensing and the Ph.D. degree in surveying and mapping from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 2002, 2005, and 2009, respectively. He is currently an Associate Professor and a Master's Supervisor with the Informati on Engineering University, Zhengzhou, China.

His research interests include photogrammetry and unmanned aerial vehicle remote sensing.

Wenjun Huang is working toward the master's degree in photogrammetry and remote sensing with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China.

Her research interests include remote sensing image processing and geographic information systems.

Longhao Wang is working toward the Ph.D. degree in photogrammetry and remote sensing with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China.

His research interests include remote sensing image digital intelligent processing.

Zijun Wei is working toward the master's degree in photogrammetry and remote sensing with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China.

Her research interests include remote sensing image matching and registration.

Fushan Yao is working toward the Ph.D. degree in photogrammetry and remote sensing with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China.

His research interests include digital image processing and autonomous navigation of unmanned aerial vehicles.