Self-Similarity Features for Multimodal Remote Sensing Image Matching

Xin Xiong, Guowang Jin, Qing Xu, and Hongmin Zhang

***Abstract*—Multimodal remote sensing image matching is a challenging task because of the existence of significant radiometric differences. To address the problem, we develop a novel multimodal remote sensing image matching method based on self-similarity features. The offset mean filtering method is proposed first to calculate the self-similarity features fast based on the symmetry of the self-similarity. The self-similarity features are presented through a multichannel self-similarity map (SSM) and a corresponding multichannel symmetric SSM. On this basis, we develop the image matching method, including a feature detector named improved maximal self-dissimilarities (IMSD) and a feature descriptor named oriented self-similarity (OSS). The IMSD detector is designed by introducing the two multichannel SSMs into the maximal self-dissimilarities (MSD) detector for feature point detection. The OSS descriptor is proposed based on the orientations of the self-similarities extracted from the multichannel SSMs. We conduct experiments with a variety of optical, synthetic aperture radar, and light detection, and ranging data. Our results demonstrate the advantages of our proposed IMSD detector and OSS descriptor in comparison with state-of-the-art detectors and descriptors, respectively. The image registration results further confirm the effectiveness of the proposed method.**

***Index Terms*—Image matching, improved maximal self-dissimilarities (IMSD), multimodal remote sensing images, offset mean filtering (OFM), oriented self-similarity (OSS).**

I. Introduction

INCREASINGLY, remote sensing technologies require using multiple sensors to observe specific and different characteristics of the earth's surface. These measures can be active, such as synthetic aperture radar (SAR) and light detection and ranging (LiDAR), or passive, such as optical, infrared, multispectral, and hyperspectral. The data acquired by them can provide information about the structure (optical, SAR), elevation (LiDAR), and material content (multispectral and hyperspectral) of the objects in the image [1]. For many applications of remote sensing (e.g., image fusion [1], image segmentation [2], and image classification [3], etc.), observations from heterogeneous sources need to be coupled and jointly analyzed, because their complementarity helps to achieve a more comprehensive description of a scene unachievable using single modality data [4]. In such applications, image matching is a fundamental step, of which the goal is to extract reliable corresponding features from two or more images of the same scene [5]. Because of the different imaging mechanisms of various sensors, the characteristics of the same ground scene typically vary in different images. Therefore, multimodal image matching is a challenging task, as the radiometric differences are extremely significant [6], [7].

Image matching has always been given attention, and remarkable advancements have been made in past decades. Generally, existing image matching methods can be classified as area-based and feature-based [8]. Area-based methods compare predefined templates in images through similarity metrics to search for optimal correspondences. Among various similarity metrics, two basic and widely used techniques are normalized cross correlation (NCC) [9] and mutual information (MI) [10]. The methods of this type can avoid the complicated process of feature extraction and, generally, can achieve high matching accuracy; however, they are sensitive to geometric differences and have the disadvantage of high computational complexity [11], [12].

Compared with area-based methods, feature-based methods are more robust to geometric differences [13]. These methods generally consist of three main steps—namely, feature detection, feature description, and feature matching. In these methods, salient image features such as points, lines, regions, and edges are usually first extracted and the similarity of the feature descriptors is subsequently compared to obtain correspondences, of which point feature is the simplest and most common. Over the past decades, numerous feature-based methods, such as scale-invariant feature transform (SIFT) [14], speeded-up robust features (SURF) [15], and oriented FAST and rotated BRIEF (ORB) [16]–[18] have been developed for point matching. As a classic algorithm, SIFT first constructs the difference of the Gaussian (DoG) scale space to extract feature points, and subsequently uses the gradient histogram to describe the features. With invariance to scale, rotation, and brightness, SIFT is effective for matching visible images. However, difficulties occur when working with multimodal images because of its sensitivity to nonlinear radiometric differences. Some variants of SIFT, such as principal component analysis SIFT (PCA-SIFT) [19], affine SIFT (ASIFT) [20], adaptive binning SIFT (AB-SIFT) [21], and SAR-SIFT [22] have been also proposed. As these are designed to manage various specific problems, such as large geometric differences and severe image speckles, they are also vulnerable

Manuscript received September 21, 2021; revised November 8, 2021 and November 21, 2021; accepted November 24, 2021. Date of publication November 30, 2021; date of current version December 15, 2021. This work was supported by the National Natural Science Foundation of China under Grants 41071296, 41474010, and 61401509. *(Corresponding author: Guowang Jin.)*

Xin Xiong, Guowang Jin, and Qing Xu are with the Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China (e-mail: xiongxinhbhh@163.com; guowang_jin@163.com; 13937169139@139.com).

Hongmin Zhang is with the Institute of Navigation and Aerospace Target Engineering, Information Engineering University, Zhengzhou 450001, China (e-mail: zhmin1206@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3131489

to complex radiometric differences. Therefore, these SIFT-based methods have limited capability for matching multimodal images.

Recently, several methods have been proposed for multimodal image matching. These methods are robust to nonlinear radiometric differences by capturing the structures and shape properties of the image. Among these, two common types are based on phase congruency (PC) and local self-similarity (LSS).

Morrone and Owens [23] revealed that the image features usually occur at the points of maximum PC, whereas, much later, Kovesi [24] improved the calculation model of PC. Ye *et al.* [25] built the orientation representation of the PC model and, on this basis, proposed the histogram of oriented phase congruency (HOPC) descriptor for multimodal image template matching. Subsequently, Ye *et al.* [26] introduced the minimum moment of phase congruency-Laplace (MMPC-Lap) detector and local histogram of oriented phase congruency (LHOPC) descriptor for matching optical images with radiometric differences. Fan *et al.* [27] presented the phase congruency structural descriptor (PCSD) by grouping PC maps to match SAR and optical images. Fu *et al.* [28] developed a dense descriptor named histograms of oriented magnitude and phase congruency (HOMPC) based on oriented PC maps for multisensor image matching. Li *et al.* [29] proposed a multimodal image matching method named radiation-invariant feature transform (RIFT). In RIFT, the maximum index map (MIM) is built based on the PC model for feature description. PC-based methods have been demonstrated to be robust to nonlinear radiometric differences. However, the computational complexity of the PC model is relatively high, particularly for large-size images [30].

Shechtman and Irani [31] proposed the LSS descriptor for object detection, retrieval, and action detection. This descriptor has been applied successfully also to remote sensing image matching [32]–[34]. Tombari *et al.* [35] designed the maximal self-dissimilarities (MSD) detector, which was inspired by LSS and could achieve stable detection results under complex radiometric differences. Ye *et al.* [36] introduced a feature descriptor named the dense LSS (DLSS) by integrating multiple LSS descriptors for optical-to-SAR image template matching. Our previous work [37] improved the DLSS descriptor by designing a novel descriptor named the rank-based local self-similarity (RLSS), which used rank values instead of correlation values. Sedaghat *et al.* [30], [38] extended the LSS descriptor to distinctive-order-based self-similarity (DOBSS) descriptor and histogram of oriented self-similarity (HOSS) descriptor, respectively, to match multisensor optical images. Chen *et al.* [39] designed the center-symmetric local-ternary-pattern (CSTLP) descriptor based on the self-similarity descriptor. By capturing the shape properties of images, LSS-based methods are less sensitive to complex radiometric differences. However, descriptors derived from LSS have relatively low discriminative capability [33]. In addition, owing to the numerous sum of squared differences (SSD) operations, their computational efficiency needs to be improved [40].

Other recently popular methods for multimodal image matching are learning based. One of the ideas is to use the Siamese convolutional neural network (CNN) and its variants to achieve patch matching of multimodal images. Merkle *et al.* [41] trained a Siamese CNN on optical and SAR image patches. Hughes *et al.* [42] identified corresponding patches with a pseudo-Siamese CNN for SAR and optical images. Baruch and Keller [43] combined the Siamese and pseudo-Siamese network to register visible and near-infrared images. Zhang *et al.* [44] designed a Siamese fully CNN to learn descriptors for multimodal image patch matching. Another idea is to add preprocessing steps for matching based on deep networks. After the preprocessing, multimodal images can be effectively matched by hand-crafted methods. Merkle *et al.* [45] trained a conditional generative adversarial network (cGAN) to generate artificial SAR-like image patches from optical images. Zhang *et al.* [46] applied the image transfer algorithm based on VGG-19. Ma *et al.* [47] used VGG-16 to calculate the approximate spatial relationship of multimodal image pairs. The learned features perform better than hand-crafted methods on some specific tasks, but they face some difficulties. On the one hand, it is very challenging to design a suitable network [46]. On the other hand, large and diverse datasets are needed for training to achieve excellent matching performance [44].

Our study focuses on solving the limitations in the LSS-based method. First, to improve computational efficiency, we propose the offset mean filtering (OMF) method to calculate self-similarity features fast. Using the OMF method, the obtained self-similarity features are presented through a multichannel self-similarity map (SSM) and a corresponding multichannel symmetric SSM. Second, to enhance the discriminative capability, we propose a novel feature descriptor named oriented self-similarity (OSS) based on the extracted multichannel SSMs. The main contributions of this study are the following.

1) The OMF method is proposed to calculate the multichannel SSM and the multichannel symmetric SSM fast, which expresses the self-similarity features of the image. Based on the symmetry of the self-similarity, OMF can avoid redundant calculations, thereby significantly reducing the computational cost.
2) The improved MSD (IMSD) detector is designed by introducing the extracted two multichannel SSMs into the MSD detector for robust feature point detection. Since the multichannel SSMs can be easily and directly embedded in the MSD detector, the IMSD detector can achieve enhanced computational efficiency.
3) The OSS descriptor is proposed by extracting the orientation information from the multichannel SSMs to enhance the distinctiveness and robustness against significant radiometric differences. The main reason for the poor discriminative capability of LSS-based descriptors is that they are sensitive to the position errors of the feature points. This is because the feature point is used as the central pixel to calculate self-similarity with all of the surrounding pixels in the local region of feature description. On the two multichannel SSMs, each pixel is used as the central pixel to calculate its own self-similarity. The OSS descriptor transforms the multichannel SSMs into an index map of orientations with minimum self-similarity values, and extracts the histogram of the index map to describe the

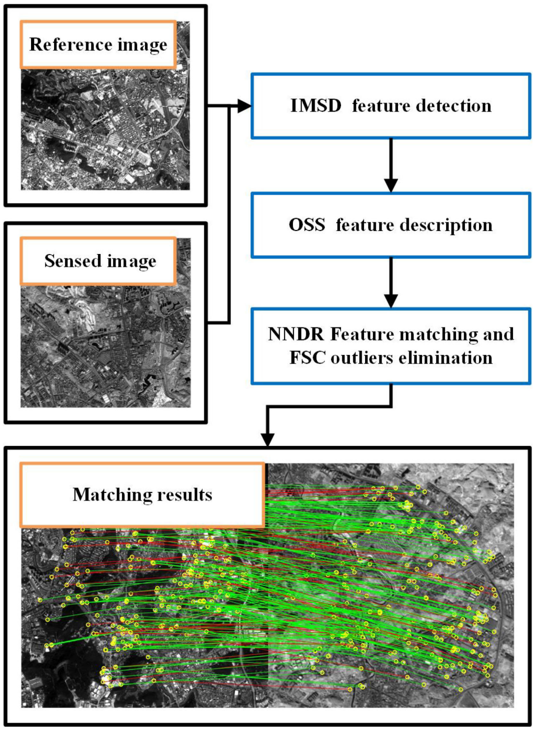


Fig. 1. Flowchart of the proposed multimodal remote sensing image matching method.

point features. Therefore, it has enhanced discriminative capability.

The rest of this article is organized as follows. The proposed multimodal image matching method is introduced in Section II. The experimental results of the proposed method are presented in Section III. Finally, Section IV concludes this article.

II. Methodology

This section introduces the OMF method and subsequently details the proposed multimodal remote sensing image matching method. The proposed matching method includes three steps. First, the IMSD detector is designed to extract reliable and sufficient feature points. Second, the OSS descriptor is proposed to distinctively describe these features. Finally, the nearest-neighbor distance ratio (NNDR) matching strategy followed by the fast sample consensus (FSC) [48] algorithm is performed to identify the effective matches. The flowchart of the proposed method is shown in Fig. 1.

A. Offset Mean Filtering

Two basic studies closely related to self-similarity are the LSS descriptor [31] and the MSD detector [35]. The self-similarity concept was first used to develop the LSS descriptor, which leverages on self-similarities between the central pixel and its surrounding pixels to provide a shape representation within a local feature region. The LSS descriptor is robust to the radiometric variations by capturing the shape structures of the image. In LSS, the similarity between pixels is measured by the SSD of the image patches centered on them. The self-similarities of a pixel are defined as the similarities between the pixel and its surrounding pixels. The self-similarity concept was developed further in the MSD detector for feature detection. MSD highlights the pixels that are most dissimilar from nearby ones within their surroundings. In MSD, the self-similarity is extended from local to global, as the feature detection process needs to be performed on the entire image. In addition, a particular form of box filtering is designed to calculate the self-similarities of all pixels with a reduced computational complexity.

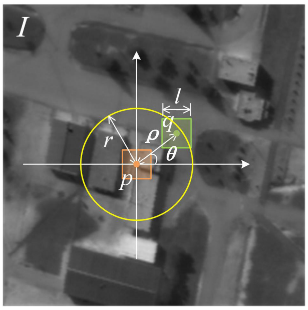


Fig. 2. Schematic of calculating the self-similarity of pixel p relative to pixel q. The red and green squares represent image patches centered on p and q, respectively.

Fig. 2 shows a schematic of calculating the self-similarity of pixel p relative to pixel q. In the figure, pixel p is on the image I. Pixel q is in the neighborhood (radius is r) of p. The pixel distance and angle between p and q are ρ and θ, respectively. l represents the side length of the square image patches for calculating similarities.

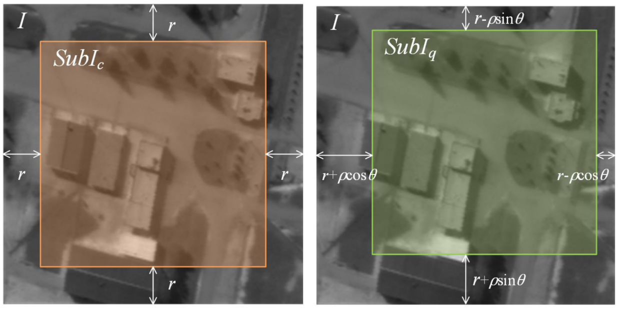
Although optimized in MSD, the computational efficiency of the aforementioned work is still limited, because they extract the self-similarity features in a pixel-by-pixel manner, resulting in numerous redundant calculations. We propose the OMF method to further improve the computational efficiency of the self-similarity in a channel-by-channel manner. The method includes mainly two steps, namely subimage construction and mean filtering.

1) Subimage Construction: The subimage is constructed by cropping the input image. For an image I with width W and height H, the central subimage $SubI_c$ and the offset subimage $SubI_q$ can be obtained separately by cropping, as shown in Fig. 3. Considering the lower left corners of images as the references, the horizontal and vertical offsets of $SubI_c$ are both r, and those of $SubI_q$ are $r + \rho\cos\theta$ and $r + \rho\sin\theta$, respectively.

2) Mean Filtering: After obtaining the subimage, the SSM S^q corresponding to q can be calculated as follows:

$$S^q = \text{meanFilter}\left(|SubI_c - SubI_q|\right) \tag{1}$$

where $\text{meanFilter}(\cdot)$ represents the mean filtering operation. Essentially, (1) uses the sum of absolute differences (SAD) instead of the SSD to calculate the self-similarity values. The operation has been proven to improve computational efficiency [40]. The window size of the mean filter is equal to the size of image patches. Herein, the window is circular (2 pixels in radius) rather than square ($l \times l$ pixels) to enhance rotational invariance [30].



(a) (b)

Fig. 3. Example of subimages construction. (a) Central subimage (red square). (b) Offset subimage (green square).

In addition, the size of S^q is equal to the size of the subimage, and the borders need to be padded to ensure that S^q is the same size as I.

For each pixel in the neighborhood of p, a corresponding SSM can be obtained. Note that the self-similarity is symmetric, i.e., the self-similarity of p relative to q is equal to the self-similarity of q relative to p. Therefore, the qth SSM and the $(N+1-q)$th SSM are almost the same with only pixel displacements, $q = 1, 2, \ldots, N$. That is

$$S^{N+1-q}(x, y) = S^q\left(x - r\cos\theta, y - r\sin\theta\right) \tag{2}$$

where N is equal to the number of pixels in the neighborhood. Here, the neighborhood is circular, as the circular neighborhood is robust under the rotation changes. Therefore, N is expressed as follows:

$$N \approx \mathrm{round}\left(\pi r^2\right) \tag{3}$$

where $\mathrm{round}(\cdot)$ represents the rounding operation.

Taking into account the symmetry of the self-similarity, we only need $N2$ OMF operations to obtain a multichannel SSM $\{S^q\}_C^1$ to express the self-similarity features of the entire image. C is the number of channels, $C = N2$. On $\{S^q\}_C^1$, every pixel p contains C values $\{s_p^q | q = 1, 2, \ldots, C\}$. According to (2), each SSM in $\{S^q\}_C^1$ corresponds to a symmetric SSM. Therefore, C symmetric SSMs can be obtained, and they form a multichannel symmetric SSM $\{S^q\}_N^{C+1}$.

Fig. 4 shows the process of extracting the self-similarity values of pixel p. After obtaining $\{S^q\}_C^1$ and $\{S^q\}_N^{C+1}$, the channel values (pixels) at point p, including effective values (pixels) and symmetric values (pixels) corresponding to $\{S^q\}_C^1$ and $\{S^q\}_N^{C+1}$, respectively, are extracted, and these pixels are arranged by channel indices to form the self-similarity features of p in its circular neighborhood.

We recommend that the radius of the neighborhood r is 4, which will be discussed in Section III-D. The corresponding number of channels C is 24. Fig. 5 details the neighborhoods of a pixel, with the orange square representing the pixel. The green and gray squares represent effective and symmetric pixels in the neighborhood, respectively. The numbers in the squares in (b) represent the channel indices.

Compared with the pixel-by-pixel calculation method used in MSD [35] (a particular form of box filtering), OMF calculates the self-similarity features in a channel-by-channel manner and reduces the calculation burden by half based on the symmetry of the self-similarity. The computational complexity of obtaining self-similarity features can be reduced from $O(W \cdot H \cdot N \cdot l^2)$ to $O(W \cdot H \cdot N)$ through a particular form of box filtering. Using OMF, the computational complexity will be reduced further to $O(W \cdot H \cdot N2)$. Therefore, the calculation efficiency is improved.

Fig. 6 shows the calculation times for the particular form of box filtering method and the proposed OMF method versus the image size. The two methods use neighborhoods with the same size and shape when calculating self-similarity features. The image size varies from 400 × 400 to 1200 × 1200. One can observe that the OMF method takes about half less time than the particular form of box filtering method, which proves the effectiveness of OMF in improving efficiency.

B. IMSD Detector

In this section, we present a novel feature detector named IMSD to extract reliable and sufficient feature points in the multimodal images. The IMSD detector is an improved version of the MSD detector, which calculates the self-similarity features using the proposed OMF method. Evidently, the IMSD detector can achieve high computational efficiency, which is attributed to two factors. First, the smaller circular neighborhood with a radius of 4 pixels replaces the square neighborhood of 11 × 11 pixels used in MSD. Second, the OMF method is used considering the symmetry of the self-similarity.

After obtaining $\{S^q\}_C^1$ and $\{S^q\}_N^{C+1}$, the self-similarity values of each pixel can be extracted, as shown in Fig. 4. The feature response λ of point p is expressed as follows [35]:

$$\lambda(p) = \frac{1}{k}\sum_{i=1}^{k} s_p^i \tag{4}$$

where $s_p^1, \ldots, s_p^k$ are the smallest k self-similarity values, $k = 4$. By performing the local nonmaximal suppression on λ, the feature points can be obtained.

To detect multiscale feature points, a multiscale Gaussian pyramid is constructed for the input image. Compared with the pyramid established by direct downsampling in MSD [35], the Gaussian pyramid enhances the robustness of the detector to noise. In addition, it is helpful to improve the discriminative capability of the proposed OSS descriptor, which will be discussed in Section III-C. The image is first Gaussian smoothed and subsequently downsampled.

The number of pyramid layers L is closely related to the input image size as follows [49]:

$$L = \mathrm{floor}\left(\log_2\left(\min(W, H)\right) - 2\right) \tag{5}$$

where $\mathrm{floor}(\cdot)$ represents the rounding operation toward negative infinity.

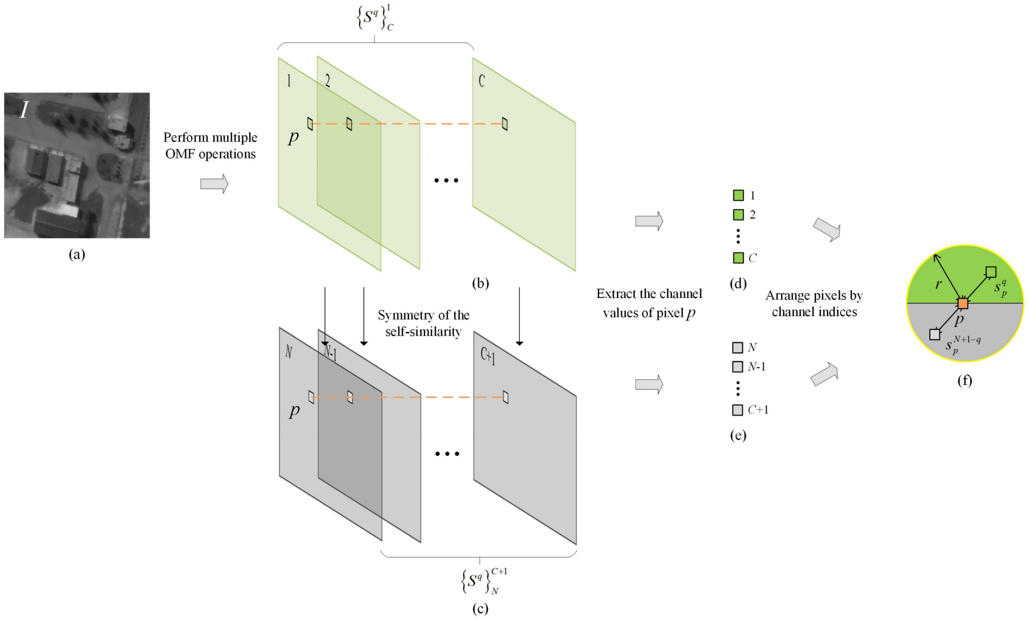
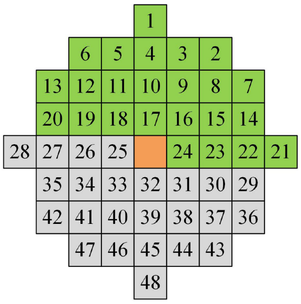

$\{S^q\}_C^1$
I
(a)
Perform multiple OMF operations
1
2
C
p
(b)
Symmetry of the self-similarity
N
N-1
C+1
p
$\{S^q\}_N^{C+1}$
(c)
Extract the channel values of pixel p
1
2
C
(d)
N
N-1
C+1
(e)
Arrange pixels by channel indices
r
s_p^q
p
s_p^{N+1-q}
(f)

Fig. 4. Process of extracting the self-similarity features of a pixel. (a) Input image. (b) Multichannel SSM. (c) Multichannel symmetric SSM. (d) Effective values (pixels). (d) Symmetric values (pixels). (e) Self-similarity values in the neighborhood of pixel p.



				1				
		6	5	4	3	2		
	13	12	11	10	9	8	7	
	20	19	18	17	16	15	14	
28	27	26	25	🟧	24	23	22	21
	35	34	33	32	31	30	29	
	42	41	40	39	38	37	36	
		47	46	45	44	43		
				48				

Fig. 5. Example neighborhoods of a pixel. The radius of the neighborhood r is 4. The orange square represents the central pixel. The green and gray squares represent effective and redundant pixels in the neighborhood, respectively.

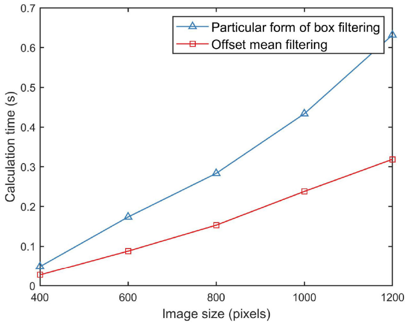


Fig. 6. Calculation time taken from the particular form of box filtering method and the proposed OMF method versus the image size.

For the ith layer, the standard deviation of the Gaussian function is expressed as follows:

$$\sigma_i = \sigma_0 f^{i-1}, \quad 1 \leq i \leq L \tag{6}$$

where σ_0 is the initial standard deviation, and we set $\sigma_0 = 1.2$ based on the discussion in Section III-D. f is the scale factor, $f = 2^{13}$.

C. OSS Descriptor

When the feature points are detected, a novel feature descriptor called OSS is designed for them. The proposed descriptor first extracts the orientation information from the two multichannel SSMs to form an index map and subsequently transforms the index map into a GLOH-like grid to generalize the descriptor values. The proposed descriptor involves mainly two steps: orientation assignment and grid representation.

1) Orientation Assignment: To render the descriptor invariant to rotation, orientation assignment should be conducted for each feature point. Similar to the method in MSD [35], a histogram generation method based on the self-similarity values is used to assign dominant orientations. Specifically, for the feature point p, 36 points (with an interval of 10° and covering 360°) are sampled uniformly on a circle (neighborhood edge of p) with p as the center and r as the radius, as shown in Fig. 7(a). The self-similarity values of p relative to these points are $s_p^1, \ldots, s_p^{36}$, respectively. These values can be extracted from $\{S^q\}_C^1$ and $\{S^q\}_N^{C+1}$ as shown in Fig. 4, and are stretched linearly to the range of [0, 1], as follows:

$$\hat{s}_p^i = \frac{\max\left\{s_p^1, \ldots, s_p^{36}\right\} - s_p^i}{\max\left\{s_p^1, \ldots, s_p^{36}\right\} - \min\left\{s_p^1, \ldots, s_p^{36}\right\}}, i = 1, \ldots, 36. \tag{7}$$

Subsequently, an orientation histogram with 36 bins (covering 360°) is generated, as shown in Fig. 7(b). The stretched values are added into the histogram based on its angle to p. The local peak(s) of the histogram (within 80% of the highest peak) is/are

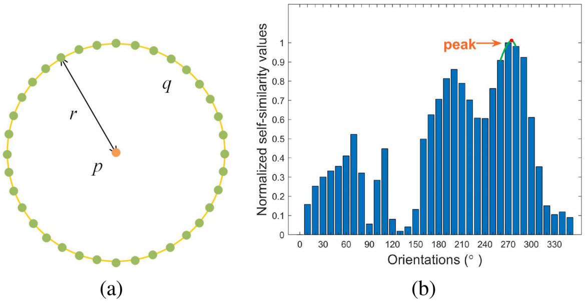


Fig. 7. Process of the orientation assignment. (a) 36 uniformly sampled points on a circle with p as the center and r as the radius. (b) Orientation histogram generation and dominant orientation assignment. The red point is the interpolated peak.

selected as dominant orientation(s) of p. Finally, a parabola is fitted to the three histogram values closest to each peak to interpolate the peak location for improved accuracy.

2) Grid Representation: To construct the descriptor, grid representation should be implemented to divide the feature-centric local region into multiple grid bins. Specifically, for feature point p, multiple circular local regions (radius R) are extracted from both $\{S^q\}_C^1$ and $\{S^q\}_N^{C+1}$. $R = 36$ is recommended based on the discussion in Section III-D. Subsequently, the extracted local regions are transformed into an index map of orientations with minimum self-similarity values. Finally, the index map is divided into multiple bins based on a descriptor grid, and a specific distribution histogram with N_o bins is built for each grid bin to generate descriptor values. The process of grid representation is shown in Fig. 8.

A key issue in the above process is to generate the index map. For each point in the local region, the neighborhood of the point is divided into $2N_o$ orientation bins. The start orientation of the division is the dominant orientation of the feature (central) point. The bin index numbers are marked counterclockwise $1, 2, \ldots, 2N_o$. The mean value of the self-similarity values is calculated in each of first N_o bins, and the bin index number corresponding to the bin with the smallest mean value is regarded as the index value of the point. The self-similarity features in the last N_o bins are ignored because they are redundant information according to the symmetry of the self-similarity. Therefore, the index value can be $1, 2, \ldots, N_o$. Fig. 8(b)–(e) shows examples of generating the index values of three points. In Fig. 8(b), the neighborhood is divided into eight orientation bins, that is, $N_o = 4$. As shown by the green sector in Fig. 8(c)–(e), the third bin of v_1,the fourth bin of v_2, and the first bin of v_3 have the smallest mean values, respectively. Therefore, the index values of v_1, v_2, and v_3 are 3, 4, and 1, respectively. Each point in the local region can generate an index value, and a local index map can therefore be obtained. We recommend $N_o = 8$ based on the discussion in Section III-D.

Another key issue is to choose the descriptor grid. Among the well-known descriptor grids, the GLOH grid (as used in SAR-SIFT) is more adaptable to geometric distortion than the square grid (as used in SIFT) [21]. As the proposed OSS descriptor requires relatively large local region to ensure the distinctiveness (discussed in Section III-C), we recommend a denser GLOH-like grid, as shown in Fig. 9. The proposed GLOH-like grid involves two parameters, i.e., the number of radial bins N_r and the number of angular bins N_a. The larger the N_r or N_a the denser will be the grid. The density of the grid should be appropriate to ensure the distinctiveness of the descriptor and, more importantly, excessive density is not considered as it can lead to an over-dimensional descriptor, resulting in low computational efficiency. $N_r = 4$ and $N_a = 10$ are recommended based on multiple experiments to balance the distinctiveness and the computational efficiency, as shown in Fig. 9(b). As a result, the dimension of the OSS descriptor is $(N_r N_a - N_a + 1)N_o = (4 \times 10 - 10 + 1) \times 8 = 248$.

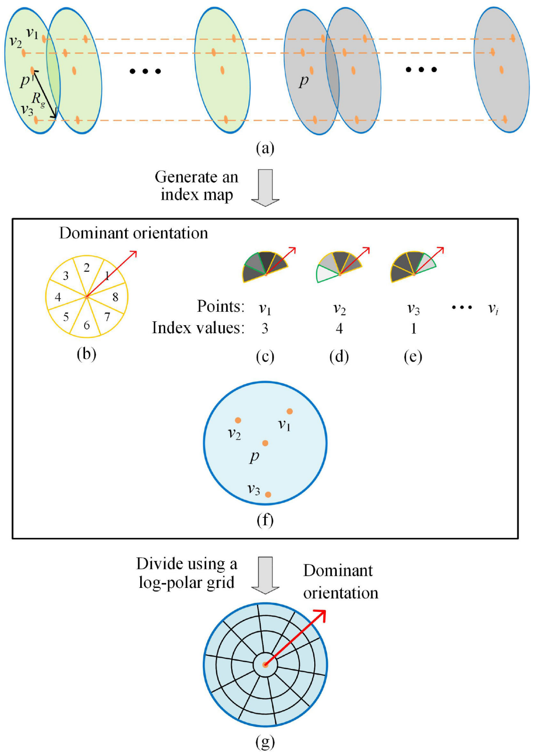


Fig. 8. Process of grid representation. (a) Extracted circular local regions. (b) Divided and numbered neighborhood. (c)–(e) Index values of points v_1, v_2, and v_3, corresponding to the bins (green sectors) with the smallest mean values of the self-similarity values. (f) Index map. (g) Divided index map.

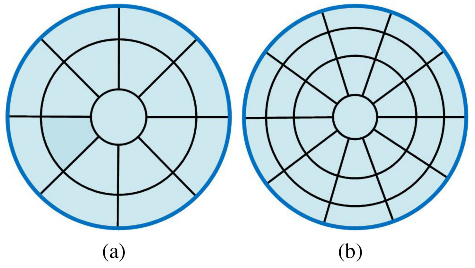


Fig. 9. Descriptor grid. (a) Regular GLOH grid. (b) Proposed denser GLOH-like grid.

TABLE I
DETAILS OF EXPERIMENTAL DATASETS

Category	NO.	Image source	Spectrum/ Polarisation	Size (pixels)	GSD (m)	Date	Location or/and description
V-I	1	Landsat TM	Band 2	600 × 600	30	06/2009	Suburban area in
		Landsat TM	Band 5	600 × 600	30	06/2010	Zhongwei, Ningxia, China
	2	GF 1	Band 3	550 × 550	8	11/2016	Suburban area in
		GF 1	Band 4	550 × 550	8	12/2017	Shenzhen, Guangdong, China
	3	GF 1	Pan	700 × 700	2	11/2017	Urban area in
		GF 2	Band 1	400 × 400	3.24	12/2017	Luoyang, Henan, China
V-S	1	Landsat TM	Band 3	600 × 600	30	05/2007	Suburban area in
		TerraSAR-X	N/A	600 × 600	30	03/2008	Rugen, Germany
	2	Google map	N/A	520 × 520	0.95	N/A	Suburban area in
		GF 3	VV	500 × 500	1	02/2018	Newport News, Virginia, USA
	3	Google map	N/A	700 × 700	2.02	N/A	Suburban area in
		GF 3	HH	494 × 494	3	11/2016	Fuzhou, Jiangxi, China
L-V	1	LiDAR	Intensity	621 × 617	2	10/2010	Urban area
		WorldView 2	Visible	621 × 621	2	10/2011	
	2	LiDAR	Height	524 × 524	2.5	06/2012	Urban area
		Airborne	Visible	453 × 453	2.5	06/2012	
	3	Leica ALS50	Height	420 × 420	1	08/2008	Urban area in
		DMC	Visible	840 × 840	0.5	08/2008	Vaihingen, Germany

D. Matching Algorithm

After feature detection and description, the NNDR matching strategy is adopted to select the initial matches by using the distance between feature descriptors. Furthermore, the FSC algorithm is used to remove outliers from the initial matches [48]. FSC can robustly extract effective matches from a large number of outliers in few iterations. However, the small distance ratio threshold d_h of FSC is extremely sensitive to the multimodal image scene, which controls the size of the sample set. Instead of subjectively fine-tuning d_h, the top 300 matches with the smallest distance are selected as the sample set. In addition, the effective matches are identified with an affine transformation model between the reference and sensed images by considering an empirical threshold of $T_0 = 3\sqrt{2}$ pixels.

III. Experiments and Results

In this section, we present the evaluation and validation of the proposed multimodal image matching method. First, the experimental data sets are introduced. Subsequently, the IMSD detector and OSS descriptor are evaluated, and the parameters of the proposed method are discussed. Finally, the proposed method is applied to image registration.

A. Datasets

The experimental data include nine multimodal image pairs that can be divided into three categories, namely 1) visible-to-infrared (V-I); 2) visible-to-SAR (V-S); and 3) LiDAR-to-visible (L-V). Image pairs V-S 1, L-V 1, and L-V 2 are disclosed in [25]. Each category contains three image pairs with significant radiometric differences. In each category, the first pair has almost no rotation and scale differences, whereas the second pair has a significant rotation difference with little or no scale difference, and the third pair has a significant scale difference with inconspicuous rotation difference. In addition, these image pairs contain a variety of medium resolution and high resolution remote sensing images from Google map, airborne sensors, and spaceborne sensors, such as Landsat TM, GaoFen (GF) 1, GF 2, TerraSAR-X, GF 3, and Worldview 2. The images have different spectrums or polarizations and cover different scenes, including suburban and urban areas. The descriptions of the datasets are presented in Table I.

Category V-I: V-I 1 to V-I 3 are visible and infrared data. V-I 1 and V-I 2 are two pairs of medium resolution images covering suburban areas. There is a temporal difference over 1 year between the images in the two pairs. During the period, the river areas in them changed. These changes make the matching more difficult. V-I 3 is a pair of medium resolution images located in an urban area.

Category V-S: V-S 1 to V-S 3 are composed of visible and SAR data covering suburban areas. V-S 1 contains a pair of medium resolution images. The SAR image in the pair suffers from strong speckles, which increases the difficulty of the matching. V-S 2 and V-S 3 are two image pairs of high resolution images.

Category L-V: L-V 1 to L-V 3 are three pairs of high resolution LiDAR and visible data covering urban areas. They have obvious local geometric distortions caused by the relief displacement of buildings. The LiDAR image used in L-V 1 is an interpolated raster intensity map. The intensity map has significant noise, which increases the difficulty of the matching. The LiDAR images used in L-V 2 and L-V 3 are the interpolated raster height maps. These height maps have a sawtooth effect at the edge of the building, which makes the matching more challenging.

B. Detector Evaluation

To evaluate the performance of the IMSD detector, comparative experiments are conducted with three popular detectors (DOG [14], SAR-Harris [22], and MSD [35]). The evaluation criteria and experimental results are detailed in the following sections.

1) Evaluation Criteria: The repeatability rate is used as criterion to evaluate the performance of the detectors [50]. A higher repeatability rate corresponds to more stable detection under imaging conditions changes. The repeatability rate is defined as follows:

$$\text{Repeatability rate} = \frac{C}{(M_r + M_s)/2} \tag{8}$$

where C represents the number of corresponding feature points. M_r and M_s represent the number of feature points in the reference and sensed images, respectively.

The corresponding feature points should satisfy both the location and scale conditions [26]. The location condition is expressed as follows:

$$\|\mathbf{H} \cdot (x_1, y_1) - (x_2, y_2)\|_2 \leq T_1 \tag{9}$$

where (x_1, y_1) and (x_2, y_2) denote the feature points in the reference and sensed images, respectively. $\mathbf{H}$ is the projection transformation model between two images, which is computed from 40–60 manually selected and well-distributed check points. T_1 is the threshold of the location error, which is empirically set to $2\sqrt{2}$.

The scale condition is given as follows:

$$\left|1 - s^2 \frac{\min\left(\sigma_1^2, \sigma_2^2\right)}{\max\left(\sigma_1^2, \sigma_2^2\right)}\right| \leq T_2 \tag{10}$$

where σ_1 and σ_2 denote the scales of the feature points in the reference and sensed images, respectively. s is the scale ratio of the two images ($s \geq 1$). T_2 is the threshold of the scale difference, which is recommended to be set to 0.4.

The processing time (PT) is used as the criterion to evaluate the computational efficiency. A smaller PT corresponds to higher efficiency. The PT is counted using a laptop with Intel(R) Core(TM) i7-8750H 2.20 GHz CPU, 32 GB RAM, and NVIDIA Quadro P1000 graphics card, using MATLAB R2018b software.

2) Experimental Results: The proposed IMSD detector, and the DOG, SAR-Harris, and MSD detectors are compared by conducting feature detection experiments on nine image pairs (see Table I). DOG and SAR-Harris are feature detectors based on gradient information. MSD and IMSD are feature detectors that use self-similarity information. Compared with the MSD detector, the IMSD detector calculates self-similarity features using the OMF method and establishes the Gaussian pyramid instead of the direct downsampling pyramid. In the experiments, four detectors acquire the same number of feature points by adjusting the contrast parameters.

Fig. 10 shows the repeatability rates of nine image pairs for four detectors. As seen, the performances of the self-similarity-based detectors (MSD and IMSD) are generally better than those of the gradient-based detectors (DOG and SAR-Harris). This result confirms that the former can adapt better to the complex radiometric differences between images. The performance of IMSD is superior to that of MSD. The reason for this finding is as follows. First, the OMF method uses a circular filter window with a radius of 2 pixels rather than a square filter window at a size of 7 × 7 pixels (used in MSD), which increases the position precision of the detected feature points. Second, the Gaussian pyramid is used to replace the pyramid built by direct downsampling, which enhances the robustness of the detector to noise. Both these factors lead to an improved repeatability rate for the proposed IMSD detector.

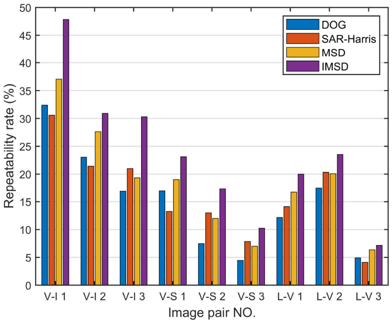


Fig. 10. Repeatability rates of nine image pairs for four detectors.

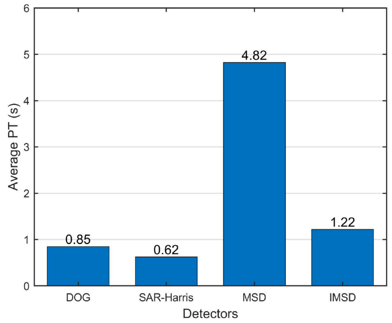


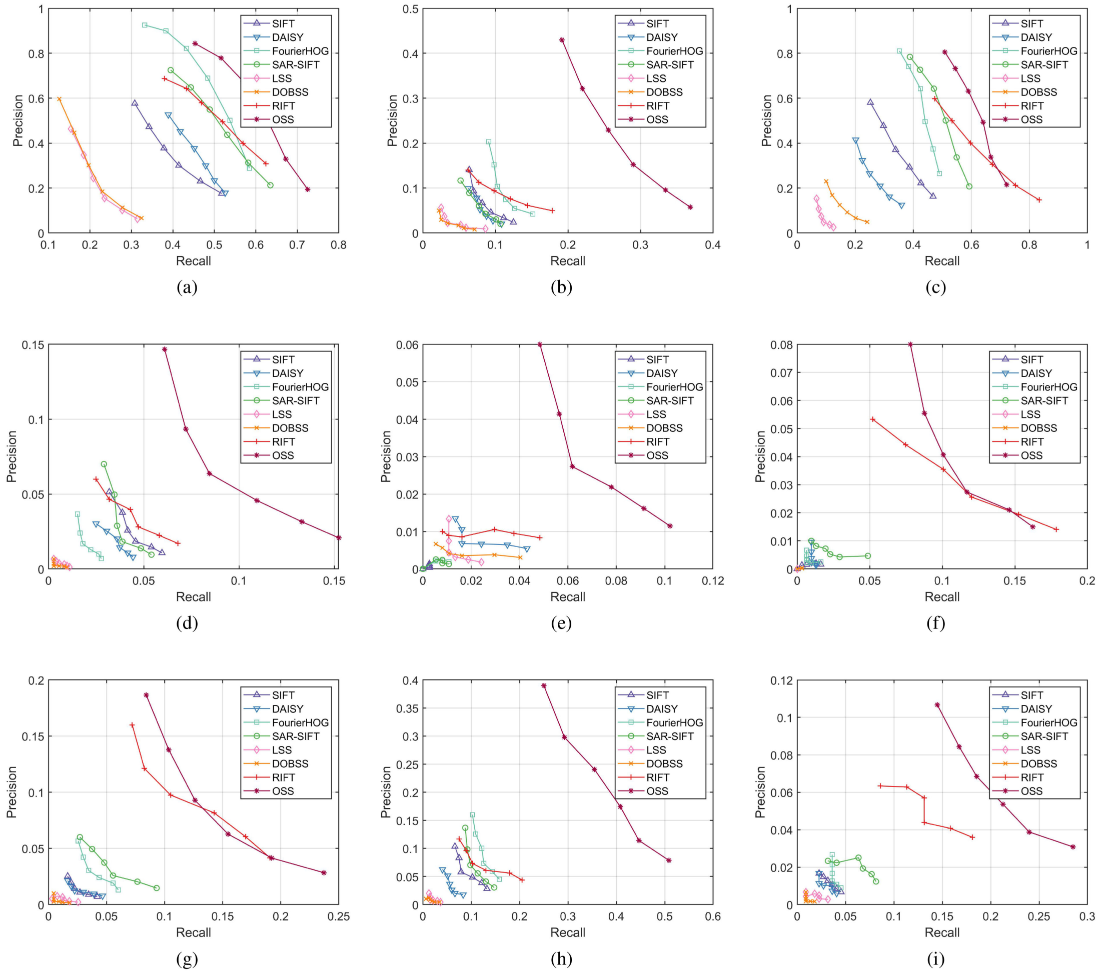
Fig. 11. Average PT of nine image pairs for four detectors.

Fig. 11 shows the average PT of nine image pairs for four detectors. The gradient-based detectors (DOG and SAR-Harris) show superior results compared with the self-similarity-based detectors (MSD and IMSD) in terms of the average PT. This is because the former only compares pixels (average image blocks) with their nearest pixels (average image blocks), whereas the latter compares image blocks in a larger neighborhood. Further, the average PT of IMSD is much smaller than that of MSD, revealing that the computational efficiency of IMSD is evidently higher than that of MSD. Two factors explain this finding. On the one hand, the two detectors use neighborhoods with different sizes and shapes when calculating self-similarity features. MSD uses a square neighborhood at a size of 11 × 11 pixels (containing 120 pixels), whereas IMSD uses a relatively small circular neighborhood with a radius of 4 pixels (containing 48 pixels). On the other hand, the OMF method used in IMSD computes the self-similarity features based on the symmetry of the self-similarity. Both these two factors result in a greatly reduced computational cost.

C. Descriptor Evaluation

To evaluate the performance of the OSS descriptor, comparative experiments are conducted with seven state-of-the-art descriptors (SIFT [14], DAISY [51], FourierHOG [52], SAR-SIFT [22], LSS [31], DOBSS [38], and RIFT [29]). The evaluation criteria and experimental results are presented in the following sections.

1) Evaluation Criteria: In our experiments, the performance of the proposed descriptor is evaluated mainly by the precision



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 12. RPC of nine image pairs for eight descriptors. (a) V-I 1. (b) V-I 2. (c) V-I 3. (d) V-S 1. (e) V-S 2. (f) V-S 3. (g) L-V 1. (h) L-V 2. (i) L-V 3.

versus recall curve (PRC) [53]. The recall and precision values are estimated as follows:

$$recall = \frac{CM}{C}, \quad precision = \frac{CM}{CM + FM} \tag{11}$$

where CM and FM are the number of correct matches and the number of false matches in the initial matches, respectively. CM can be obtained by the method described in Section III-B, and FM is equal to the total number of the initial matches minus CM. By tuning the threshold of NNDR, the different recall and precision values can be obtained in the matching process. Herein, the threshold spans between 0.9 and 1, with a step of 0.02. Subsequently, the PRC can be obtained. A higher recall or precision corresponds to the descriptor being more robust to the image scene. In other words, the farther the PRC is from the coordinate origin the superior is the descriptor performance.

2) Experimental Results: This section evaluates the performance of the proposed OSS descriptor with experiments on nine image pairs (see Table I). Furthermore, various advanced descriptors, including SIFT, DAISY, FourierHOG, SAR-SIFT, LSS, DOBSS, and RIFT, are used for comparisons. SIFT, DAISY, FourierHOG, and SAR-SIFT are feature descriptors based on gradient information. LSS, DOBSS, and OSS are feature descriptors using self-similarity information. Compared with LSS and DOBSS, OSS calculates the orientations of the self-similarity values instead of the self-similarity intensities. RIFT is a feature descriptor based on PC information, and it is not sensitive to radiometric differences. In the experiments, the IMSD detector is used to extract feature points for all the descriptors.

Fig. 12 presents the RPC of nine image pairs for eight descriptors. As seen, the proposed OSS descriptor significantly outperforms the other descriptors in all nine image pairs, with the main reason being that the descriptor extracts the orientations of the self-similarity values for feature description instead of simple gradients or self-similarity intensities. An index map is used to capture the orientations of the self-similarities and is extremely robust against significant radiometric differences.

After the OSS descriptor, better results are achieved by the RIFT descriptor, but the performance of the RIFT is unstable, and the results of some image pairs are degraded. Specifically, for image pairs V-I 1, V-I 3, V-S 3, L-V 1, and L-V 3, the results of the RIFT descriptor are relatively good, and are almost superior to those of other descriptors except the OSS descriptor. But for other image pairs, the results of the RIFT descriptor are poor and the advantage is not obvious. There are two reasons for

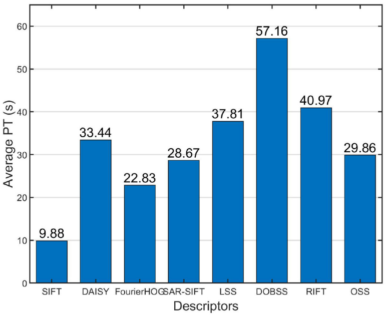


Fig. 13. Average PT of nine image pairs for eight descriptors.

this. One is that the rotation invariance of the RIFT descriptor is unstable. The RIFT descriptor does not perform the orientation assignment, but extracts six descriptors in different orientations (between 0° and 180°, with an angular interval of 30°). The image pairs V-I 2, V-S 2, and L-V 2 have obvious rotation differences, and the rotation angles are about 40°, 9°, and 10°, respectively. The rotation angles deviate from the orientations of the descriptors, so the results of the three image pairs are degraded. The second is that the RIFT descriptor is sensitive to noise [54]. The SAR image in image pair V-S 1 is affected by strong speckles. Therefore, the superiority of RIFT is not obvious for the image pair.

The performances of the other six descriptors (SIFT, DAISY, FourierHOG, SAR-SIFT, LSS, and DOBSS) are poor, especially for image pairs in Category V-S and Category L-V, which shows that they are limited in matching multimodal image pairs.

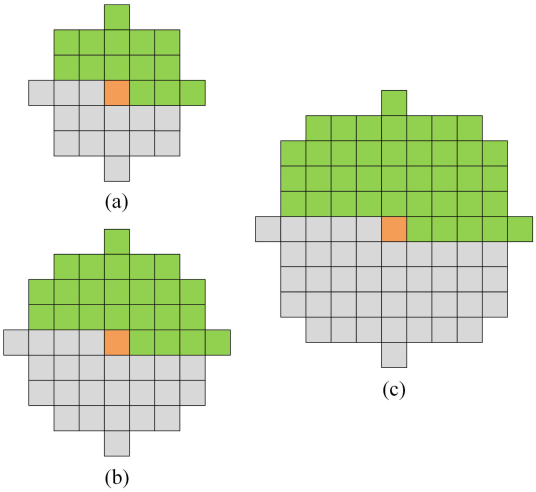
The three categories of image pairs obtains different matching results because of the differing characteristics of the images. The results of image pairs in Category V-I are better than those of image pairs in Category V-S and Category L-V. This is because the difference in the imaging mechanism between images in Category V-I is smaller than those of the other categories, and the matching is relatively easy. The results of image pairs in Category V-S are inferior to those of image pairs in Category L-V. This is because the SAR images in Category V-S are affected by speckles, and the multiplicative noise makes matching difficult.

Fig. 13 shows the average PT of nine image pairs for eight descriptors. The result of the OSS descriptor is better than those of the DAISY, LSS, DOBSS, and RIFT descriptors but inferior to the results of other descriptors. This indicates the computational efficiency of the proposed OSS descriptor is moderate among the compared descriptors, and is superior to those of the compared self-similarity-based descriptors (LSS and DOBSS).

D. Parameters Discussion

In this section, we discuss the parameter tuning of the OMF method, the IMSD detector, and the OSS descriptor, respectively.

1) Parameter Tuning of OMF Method: The proposed OMF method involves one key parameter, namely the radius of the neighborhood r. From (3), it can be seen that as r increases the number of channels also increases rapidly, leading to a sharp increase in computational cost. However, r should not be too small, because it determines the amount of information in the multichannel SSMs. The amount of information should be sufficient for reliable feature detection and description.


(a)

(b)

(c)

Fig. 14. Example neighborhoods of a pixel. The radius of the neighborhood r corresponding to (a), (b), and (c) are 3, 4, and 5, respectively. The orange square represents the central pixel. The green and gray squares represent effective and redundant pixels in the neighborhood, respectively.

Fig. 14 details the neighborhoods of a pixel, with the orange square representing the pixel. The green and gray squares represent effective and symmetric pixels in the neighborhood, respectively. The radius of the neighborhood r corresponding to (a), (b), and (c) are 3, 4, and 5, respectively, and the corresponding number of channels C is 14, 24, and 40, respectively. We will discuss the influence of r on the IMSD detector and the OSS descriptor in the following subsections.

2) Parameter Tuning of IMSD Detector: The proposed IMSD detector involves two key parameters—namely, the initial standard deviation σ_0 and the radius of the neighborhood r. The initial standard deviation σ_0 determines the smoothing ability of the Gaussian filtering in the Gaussian pyramid. The parameter affects both the proposed detector and descriptor. σ_0 should not be too small or too large. The former limits the ability to denoise, while the latter blurs the boundaries of the image. These two cases not only reduce the position accuracy of the detected features, but also make the extracted orientation information of the self-similarity unreliable. The value of r should be appropriate for reliable feature detection.

To analyze the influences of σ_0 and r, independent experiments are conducted on nine image pairs (see Table I) with different σ_0 and r. Each experiment has one parameter as a variable, with the other parameter as an invariant. In the experiments, $\sigma_0 = 0$ means that the pyramid is established by direct downsampling, similar to the method used in MSD. The average repeatability rate is used as the evaluation criterion. The experimental results are shown in Figs. 15 and 16. As seen, the detector performs better in terms of average repeatability rate when $\sigma_0 = 1.2$ and $r = 4$. Herein, $\sigma_0 = 1.2$ and $r = 4$ are selected as the default value, and the same values are also

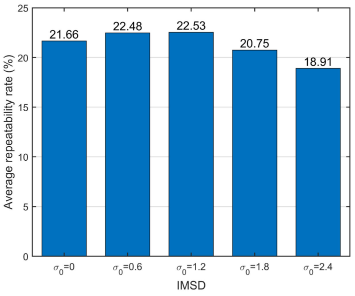


Fig. 15. Average repeatability rate with σ_0 varying from 0 to 2.4.

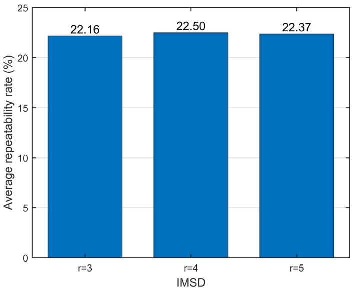


Fig. 16. Average repeatability rate with r varying from 3 to 5.

recommended in the following subsection by analyzing the influences of σ_0 and r on the proposed descriptor.

3) Parameter Tuning of OSS Descriptor: The proposed OSS descriptor involves four key parameters, namely the radius of local region R, the number of orientations N_o, the initial standard deviation σ_0, and the radius of the neighborhood r. Specifically, a larger R corresponds to a richer amount of information captured by the descriptor. If the amount of information is inadequate, it will be difficult to describe features discriminatively; however, if the amount of information is overlarge, the descriptor may be sensitive to the local geometric distortion. The larger the N_o, the more accurately the orientation information of the self-similarity be extracted. However, as the neighborhood of the point is relatively small, N_o should not be too large to avoid redundancy. The value of σ_0 and r should be appropriate to ensure the extracted orientation information of the self-similarity reliable.

To analyze the influences of these parameters, independent experiments are conducted on nine image pairs (see Table I). Each experiment has only one parameter as a variable, with the other parameters as invariants. The average PRC is used as the evaluation criterion in the experiments. The experimental results are shown in Figs. 17– 20. As seen, as R, N_o, σ_0, or r increases, the descriptor perform better in terms of average PRC until $R = 36$, $N_o = 8$, $\sigma_0 = 1.2$, and $r = 4$. Therefore, $R = 36$, $N_o = 8$, $\sigma_0 = 1.2$, and $r = 4$ are selected as the default values.

E. Application to Image Registration

In this section, we apply the proposed method to image registration and analyze the registration performance by comparing it with two state-of-the-art methods: SAR-SIFT [22] and RIFT [29].

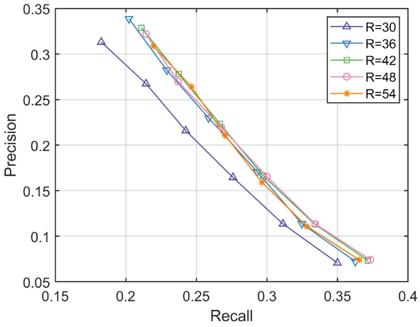


Fig. 17. Average PRC with R varying from 30 to 54.

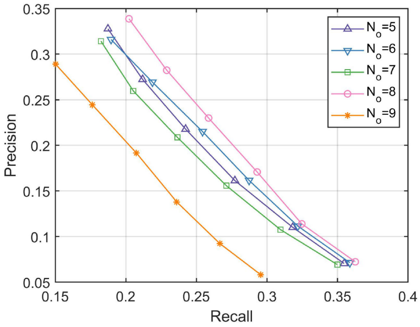


Fig. 18. Average PRC with N_o varying from 5 to 9.

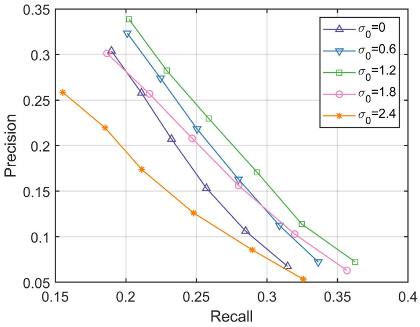


Fig. 19. Average PRC with σ_0 varying from 0 to 2.4.

1) Evaluation Criteria: In the experiments, the performance of the proposed method is evaluated by three criteria: CM, precision, and root mean square error (rmse). Different from Section III-C, CM and precision here are calculated based on the final matches, which are obtained after outlier removal using the FSC algorithm. A higher value of CM or precision indicates a better performance of the registration method; rmse is computed with the correct matches to evaluate the positional accuracy. A smaller rmse value corresponds to a higher positional accuracy.

2) Parameter Tuning of Matching Threshold: The matching threshold T_0 determines the sensitivity of the FSC algorithm to errors. The value of T_0 should be appropriate. If T_0 is too small, the correct matches will be eliminated. However, if T_0 is too large, the false matches will be retained.

To analyze the influences of T_0, experiments are conducted on nine image pairs (see Table I) with different T_0. In the experiments, the average CM and average precision of nine

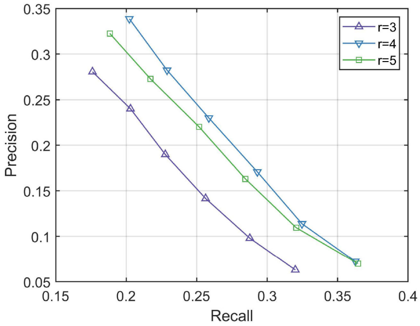


Fig. 20. Average PRC with r varying from 3 to 5.

TABLE II
AVERAGE REGISTRATION RESULTS WITH THE MATCHING THRESHOLD VARYING FROM $\sqrt{2}$ TO $5\sqrt{2}$

Criterion	Matching threshold T_0				
	$\sqrt{2}$	$2\sqrt{2}$	$3\sqrt{2}$	$4\sqrt{2}$	$5\sqrt{2}$
Average CM	117.8	167.2	178.1	177.2	177.0
Average precision	0.847	0.815	0.672	0.558	0.492

TABLE III
COMPARATIVE REGISTRATION RESULTS OF SAR-SIFT, RIFT, AND THE PROPOSED METHOD FOR NINE IMAGE PAIRS

Category	NO.	Criterion	SAR-SIFT	RIFT	Proposed
V-I	1	CM	339	770	734
		Precision	0.741	0.795	0.889
		RMSE	1.418	1.212	1.010
		PT (s)	21.155	53.689	38.772
	2	CM	N/A	171	194
		Precision	N/A	0.670	0.798
		RMSE	N/A	1.474	1.279
		PT (s)	N/A	41.812	31.584
	3	CM	209	2	382
		Precision	0.768	0.222	0.874
		RMSE	1.510	1.640	1.170
		PT (s)	20.562	29.820	35.351
V-S	1	CM	N/A	49	91
		Precision	N/A	0.644	0.777
		RMSE	N/A	1.737	1.477
		PT (s)	N/A	50.878	38.753
	2	CM	N/A	8	35
		Precision	N/A	0.242	0.636
		RMSE	N/A	1.770	1.658
		PT (s)	N/A	33.749	26.325
	3	CM	5	14	27
		Precision	0.172	0.285	0.337
		RMSE	1.921	1.960	1.844
		PT (s)	25.848	41.007	39.622
L-V	1	CM	11	120	131
		Precision	0.478	0.517	0.519
		RMSE	1.663	1.709	1.646
		PT (s)	23.316	55.401	40.208
	2	CM	6	58	161
		Precision	0.375	0.707	0.712
		RMSE	1.799	1.766	1.599
		PT (s)	11.228	29.435	25.123
	3	CM	N/A	N/A	22
		Precision	N/A	N/A	0.354
		RMSE	N/A	N/A	1.375
		PT (s)	N/A	N/A	43.746

image pairs are used as the evaluation criteria. The experimental results are shown in Table II. As seen, as T_0 increases, the average CM increases first and then remains almost unchanged, and the average precision continues to decrease. When T_0 is taken as $3\sqrt{2}$, the average CM reaches a high value. Therefore, $T_0 = 3\sqrt{2}$ is set in the article.

3) Comparative Analysis: To analyze the registration performance, comparative experiments are conducted on nine image pairs (see Table I) with three methods: SAR-SIFT, RIFT, and the proposed method. SAR-SIFT detects and describes scale-invariant features based on the gradient by ratio to improve the robustness to speckles. RIFT detects and describes radiation-invariant features based on PC information, and it is not scale-invariant. All methods use the same matching method (NNDR and FSC). For these methods, almost all parameter settings follow the recommendations of their author, except that the contrast threshold is fine-tuned to ensure that they extract approximately equal numbers of feature points.

Table III presents the comparative registration results of SAR-SIFT, RIFT, and the proposed method for nine image pairs. The proposed method is capable of robustly registering multimodal image pairs and generally outperforms SAR-SIFT and RIFT in terms of CM, precision, and rmse. This is because the proposed method uses the IMSD detector and the OSS descriptor. The IMSD detector can reliably detect a large number of feature points with a high repeatability rate; the OSS descriptor can robustly describe different features in a discriminative manner. They exhibit excellent performances under significant radiometric differences.

The RIFT method is capable of registering almost all image pairs, and even obtains a larger CM for the image pair V-I 1 than the proposed method. This is because the RIFT method is based on PC features, and it is robust to nonlinear radiometric variations. However, its overall performance is limited. Specifically, the registration performances of image pairs V-I 2, V-S 1, L-V 1, and L-V 2 are inferior to the proposed method; the registration performances of image pairs V-I 3, V-S 2, and V-S 3 are very poor; and image pair L-V 3 fail to register. There are two reasons for this. One is that the RIFT method only extracts feature points on a single scale and therefore does not have scale invariance. Image pairs V-I 3, V-S 3, and L-V 3 have obvious scale differences, so they are almost fail to register. Second, the RIFT descriptor does not have stable rotation invariance and is sensitive to noise. Image pairs V-I 2, V-S 2, and L-V 2 have rotation differences, image pairs V-S 1, V-S 2, and V-S 3 are affected by speckles, and image pair L-V 1 is affected by strong noise. Therefore, the registration results of these image pairs are degraded.

The performance of SAR-SIFT is the most vulnerable. Specifically, image pairs V-I 2, V-S 1, V-S 2, and L-V 3 fail to register, and other image pairs succeed in registering, but the performance is poor. This is because SAR-SIFT focuses on overcoming the image speckles and is relatively sensitive to complex nonlinear radiometric variations.

In terms of rmse, the proposed method outperforms SAR-SIFT and RIFT. The positional accuracy of the matching more depends on the type of feature detector than the type of feature descriptor [30]. Therefore, the reason for the results of rmse is that the IMSD detector can obtain feature points with subpixel

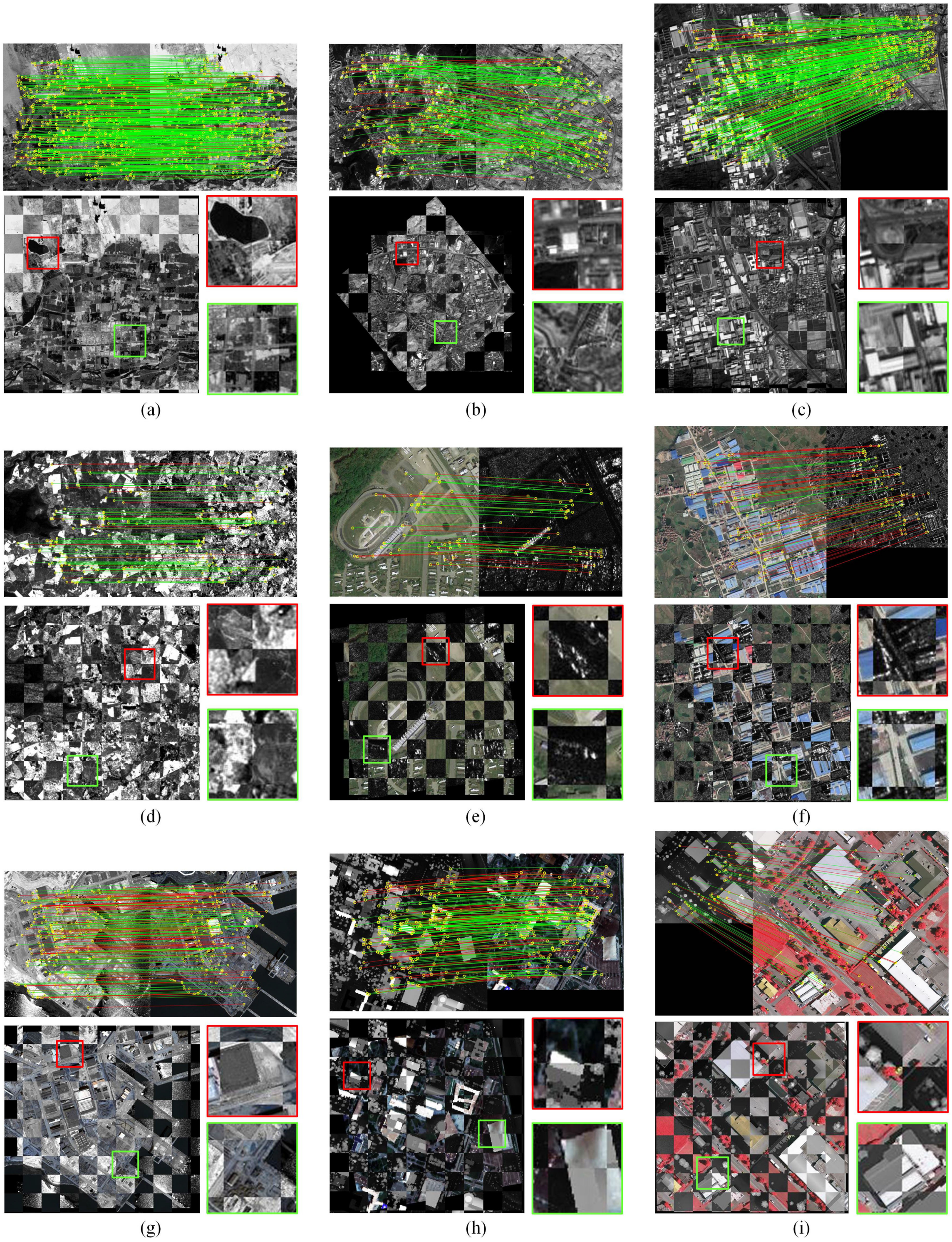


Fig. 21. Registration results of the proposed method for nine image pairs. (a) V-I 1. (b) V-I 2. (c) V-I 3. (d) V-S 1. (e) V-S 2. (f) V-S 3. (g) L-V 1. (h) L-V 2. (i) L-V 3.

precision, while SAR-Harris and FAST detectors can only obtain feature points with pixel precision.

Overall, the PT results of the proposed method are better than those of RIFT but inferior to the results of SAR-SIFT. This indicates the proposed method spends a moderate level of computation time among compared methods. The results are consistent with the average PT results of the descriptors. This is because in the image matching process, the PT of the descriptor construction is much greater than that of other steps.

Fig. 21 shows the registration results of the proposed method for nine image pairs. The proposed method achieves a sufficient number of uniformly distributed matches in multimodal image pairs with significant geometric (rotation and scale) and radiometric differences, confirming the effectiveness of the proposed method for multimodal image registration.

IV. Conclusion

We propose the OMF method to calculate the self-similarity features fast and, on that basis, propose a novel multimodal remote sensing image matching method, including the IMSD detector and the OSS descriptor. With IMSD, we introduce the extracted multichannel SSMs into the MSD detector to detect a sufficient number of feature points with a high repeatability rate. With OSS, we utilize the orientations of the self-similarities with a denser GLOH-like grid to describe the features discriminatively. We evaluate the proposed method using a variety of multimodal remote sensing images, including optical, SAR, and LiDAR data. The experimental results demonstrate that IMSD and OSS can outperform state-of-the-art feature detectors and descriptors, and the promising results (better than those of the compared self-similarity-based methods) of IMSD and OSS in terms of computational efficiency reveal the effectiveness of the OMF method. In addition, we apply the proposed method to image registration. The registration results demonstrate that the proposed method is robust against nonlinear radiometric differences, which further confirm the effectiveness of the proposed method.

In the future, we will test the proposed method on more multimodal remote sensing images. In addition, the effective matching results motivate us to integrate the algorithm into various remote sensing applications, such as image fusion and change detection.

References

[1] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proc. IEEE*, vol. 103, no. 9, pp. 1585–1601, Sep. 2015.

[2] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral-multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5103–5113, Jun. 2021.

[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] R. Feng, H. Shen, J. Bai, and X. Li, "Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions," *IEEE Geosci. Remote Sens. Mag.*, Early access, doi: 10.1109/MGRS.2021.3081763.

[5] A. Wong and D. A. Clausi, "ARRSI: Automatic registration of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1483–1493, May 2007.

[6] A. Moghimi, T. Celik, A. Mohammadzadeh, and H. Kusetogullari, "Comparison of keypoint detectors and descriptors for relative radiometric normalization of bitemporal remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4063–4073, Mar. 2021, doi: 10.1109/JSTARS.2021.3069919.

[7] A. Moghimi, A. Sarmadian, A. Mohammadzadeh, T. Celik, M. Amani, and H. Kusetogullari, "Distortion robust relative radiometric normalization of multitemporal and multisensor remote sensing images using image features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, Mar. 2021, doi: 10.1109/TGRS.2021.3063151.

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

[9] J. C. Yoo and T. H. Han, "Fast normalized cross-correlation," *Circuits Syst. Signal Process.*, vol. 28, no. 2, pp. 144–156, 2009.

[10] S. Suri and P. Reinartz, "Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 939–949, Feb. 2010.

[11] R. Feng, Q. Du, X. Li, and H. Shen, "Robust registration for remote sensing images by combining and localizing feature- and area-based methods," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 15–26, 2019.

[12] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.

[13] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4516–4527, Nov. 2011.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, vol. 11, pp. 2564–2571.

[17] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, Jan. 2010.

[18] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.

[19] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," *Comput. Vis. Pattern Recognit.*, vol. 4, pp. 506–513, 2004.

[20] J. M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.

[21] A. Sedaghat and H. Ebadi, "Remote sensing image matching based on adaptive binning SIFT descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5283–5293, Oct. 2015.

[22] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[23] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recogn. Lett.*, vol. 6, no. 5, pp. 303–313, 1987.

[24] P. Kovesi, "Image features from phase congruency," *Videre: J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.

[25] 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.

[26] Y. Ye, J. Shan, S. Hao, L. Bruzzone, and Y. Qin, "A local phase based invariant feature for remote sensing image matching," *ISPRS J. Photogramm. Remote. Sens.*, vol. 142, pp. 205–221, 2018.

[27] J. Fan, Y. Wu, M. Li, W. Liang, and Y. Cao, "SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5368–5379, Sep. 2018.

[28] Z. Fu, Q. Qin, B. Luo, H. Sun, and C. Wu, "HOMPC: A local feature descriptor based on the combination of magnitude and phase congruency information for multi-sensor remote sensing images," *Remote Sens.*, vol. 10, no. 8, pp. 1234, 2018.

[29] 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, vol. 29, pp. 3296–3310, Dec. 2020, doi: 10.1109/TIP.2019.2959244.

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

[31] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Minneapolis, MN, USA, 2007, vol. 2, pp. 1–8.

[32] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *ISPRS J. Photogramm. Remote. Sens.*, vol. 90, pp. 83–95, 2014.

[33] L. Yang, Z. Tian, W. Zhao, W. Yan, and J. Wen, "Description of salient features combined with local self-similarity for SAR image registration," *J. Indian Soc. Remote Sens.*, vol. 45, no. 1, pp. 131–138, 2017.

[34] X. Liu, S. Chen, L. Zhuo, J. Li, and K. Huang, "Multi-sensor image registration by combining local self-similarity matching and mutual information," *Front. Earth Sci.*, vol. 12, no. 4, pp. 779–790, 2018.

[35] F. Tombari and L. D. Stefano, “Interest points via maximal self-dissimilarities,” in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 586–600.

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

[37] X. Xiong, Q. Xu, G. Jin, H. Zhang, and X. Gao, “Rank-based local self-similarity descriptor for optical-to-SAR image matching,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1742–1746, Oct. 2020.

[38] A. Sedaghat and H. Ebadi, “Distinctive order based self-similarity descriptor for multi-sensor remote sensing image matching,” *ISPRS J. Photogramm. Remote. Sens.*, vol. 108, pp. 62–71, 2015.

[39] S. Chen, X. Li, H. Yang, and L. Zhao, “Robust local feature descriptor for multisource remote sensing image registration,” *J. Appl. Remote Sens.*, vol. 12, no. 1, pp. 1–17, 2018.

[40] J. Liu, G. Zeng, and J. Fan, “Fast local self-similarity for describing interest regions,” *Pattern Recognit. Lett.*, vol. 33, no. 9, pp. 1224–1235, 2012.

[41] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, “Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images,” *Remote Sens.*, vol. 9, no. 6, 2017, Art. no. 586.

[42] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, “Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[43] E. B. Baruch and Y. Keller, “Multimodal matching using a hybrid convolutional neural network,” 2018, *arXiv:1810.12941*.

[44] H. Zhang *et al.*, “Registration of multimodal remote sensing image based on deep fully convolutional neural network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3028–3042, Aug. 2019.

[45] N. Merkle, S. Auer, R. Müller, and P. Reinartz, “Exploring the potential of conditional adversarial networks for optical and SAR image matching,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 6, pp. 1811–1820, Jun. 2018.

[46] J. Zhang, W. Ma, Y. Wu, and L. Jiao, “Multimodal remote sensing image registration based on image transfer and local features,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1210–1214, Aug. 2019.

[47] W. Ma, J. Zhang, Y. Wu, L. Jiao, H. Zhu, and W. Zhao, “A novel two-step registration method for remote sensing images based on deep and local features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4834–4843, Jul. 2019.

[48] Y. Wu, W. Ma, M. Gong, L. Su, and L. Jiao, “A novel point-matching algorithm based on fast sample consensus for image registration,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 43–47, Jan. 2015.

[49] S. Liu and J. Jiang, “Registration algorithm based on line-intersection-line for satellite remote sensing images of urban areas,” *Remote Sens.*, vol. 11, no. 12, 2019, Art. no. 1400.

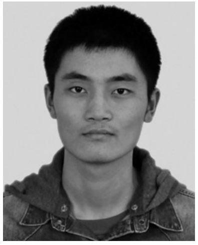
[50] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[51] E. Tola, V. Lepetit, and P. Fua, “DAISY: An efficient dense descriptor applied to wide-baseline stereo,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[52] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, “Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[53] M. Gesto-Diaz, F. Tombari, D. Gonzalez-Aguilera, L. Lopez-Fernandez, and P. Rodriguez-Gonzalvez, “Feature matching evaluation for multimodal correspondence,” *ISPRS J. Photogramm. Remote. Sens.*, vol. 129, pp. 179–188, 2017.

[54] Q. Yu, D. Ni, Y. Jiang, and Y. Yan, J. An, and T. Sun, “Universal SAR and optical image registration via a novel sift framework based on nonlinear diffusion and a polar spatial-frequency descriptor,” *ISPRS J. Photogramm. Remote. Sens.*, vol. 171, pp. 1–17, 2021.



Xin Xiong received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 2014, 2017, and 2021, respectively.

He is currently a Lecturer with Information Engineering University. His research interests include image feature extraction and matching.

Guowang Jin received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 2000, 2003, and 2007, respectively.

He is currently a Professor and a Doctoral Supervisor with Information Engineering University. His research interests include radargrammetry and synthetic aperture radar (SAR) interferometry.

Qing Xu received the B.S., M.S., and Ph.D. degrees from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 1985, 1990, and 1995, respectively.

Currently, he is a Professor and a Doctoral Supervisor with Information Engineering University. His research interests include photogrammetry and deep space remote sensing mapping.

Dr. Xu has been a member of the Professional Committee of the Photogrammetry and Remote Sensing of the Chinese Society for Geodesy, Photogrammetry and Cartography (CSGPC), since 2000. Since 2004, he has been a member of the Professional Committee of the Deep Space Exploration of the Chinese Society of Astronautics (CSA). Since 2006, he has been a member of the Expert Committee of the Scientific Application of the Exploring Project around the Moon.

Hongmin Zhang received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from the Institute of Geospatial Information, Information Engineering University, Zhengzhou, China, in 2007, 2010, and 2013, respectively.

She is currently an Associate Professor with Information Engineering University. Her research interests include radar signal processing and radargrammetry.