Robust Registration of Optical and SAR Images Using Multi-Orientation Relative Total Variation Structural Representation

Jianwei Fan, Qing Xiong, Jian Li, and Yuanxin Ye

***Abstract*—Accurate registration of optical and synthetic aperture radar (SAR) images remains a challenging task because of the potential large modality differences across individual images. To improve the registration performance, this article proposes a robust registration method for optical and SAR images based on a novel multi-orientation relative total variation (MORTV) structural representation. The MORTV model is designed by integrating multiple orientation strategy into the original RTV to extract the structural maps, which can capture more structural features while removing image noises and textures. Then, a novel feature descriptor called layerwise multiscale histogram of oriented gradient (LMHOG) is constructed on the multiscale structural maps that are generated using the MORTV model with different parameters. The LMHOG can fully characterize structural features at different scales in a multilayer manner, further enhancing the robustness and distinctiveness of the descriptor without increasing its dimension. Comprehensive experiments on two large-scale optical and SAR image datasets validate that the proposed method obtains superior registration performance over several state-of-the-art methods.**

***Index Terms*—Image registration, multimodal remote sensing images, multi-orientation, relative total variation (RTV), structural feature extraction.**

I. Introduction

With the advancement of imaging techniques, huge amounts of multimodal remote sensing data, including multispectral, light detection and ranging, and synthetic aperture radar (SAR) images, are available. Comprehensive analysis of these data is beneficial to Earth observation since multimodal data contain richer complementary information of the same scene. Currently, optical and SAR images are the two main sources of multimodal remote sensing data. The integration of optical and SAR images has been widely used in many applications, such as image fusion [1], change detection [2], and object detection [3]. To achieve these tasks, optical and SAR image registration is a crucial preliminary step, whose goal is to identify correct correspondences and spatially align the images of the same scene acquired by different imaging modalities [4].

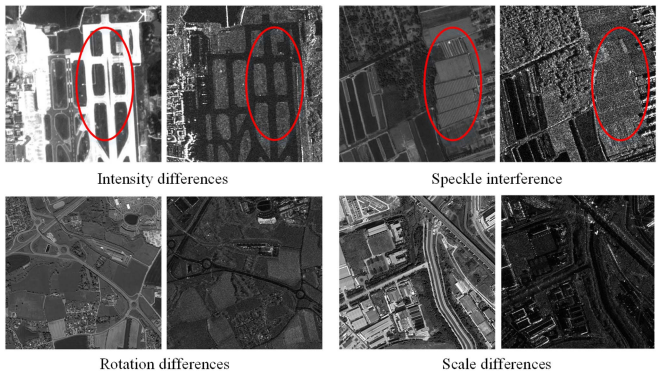

Fig. 1. Illustration of challenges of optical and SAR image registration.

It is well known that optical images acquired by passive sensors reflect the radiometric attributes of ground features and provide good interpretability. In contrast, SAR images captured by active sensors reflect the electromagnetic properties of ground objects, offering the ability of surface penetration in all weather conditions. Because of their different imaging mechanisms, there exists complex nonlinear intensity and geometric differences between optical and SAR image pairs. These challenges bring great difficulty to achieve alignment between them. Furthermore, strong speckle noise, which often appears in SAR images, makes the matching task even more challenging. Different challenges of optical and SAR image registration are shown in Fig. 1.

To address the abovementioned difficulties, optical and SAR image registration has been extensively studied over the last decades. Existing registration methods found in the literature can be roughly divided into intensity-based methods, feature-based methods, and learning-based methods [5]. Methods in the intensity-based category align the optical and SAR images based on the image similarity measures of the intensity information. Commonly used similarity measures include normalization cross correlation (NCC) [6], [7] and mutual information (MI) [8]. However, the NCC is sensitive to significant intensity variations, while the MI shows poor performance under severe noises and geometric differences. The performances of these approaches

Manuscript received 25 July 2023; revised 29 August 2023; accepted 27 September 2023. Date of publication 2 October 2023; date of current version 16 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002307 and Grant 42271446), and in part by Xinyang Normal University through Nanhu Scholars Program for Young Scholars. *(Corresponding author: Yuanxin Ye.)*

Jianwei Fan, Qing Xiong, and Jian Li are with the School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China (e-mail: fanjw@xynu.edu.cn; xqing7709@163.com; lijcit@xynu.edu.cn).

Yuanxin Ye is with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China (e-mail: yeyuanxin@home.swjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3321387

are still unsatisfactory due to an inherent limitation of similarity measures based on image intensity information [9], [10].

Compared with intensity-based methods, feature-based methods are more robust for handling nonlinear intensity and geometric differences. Methods in the feature-based category are composed mainly of two essential stages: feature extraction and feature description. As surveyed in [4], the point features have been popularly applied to optical and SAR image registration. Some well-known point feature detectors are Harris [11], oriented fast and rotated brief [12], difference of Gaussian [13], and optical-to-SAR scale invariant feature transform (OS-SIFT) [14]. Although these detectors show promising performance to some extent, they cannot effectively handle severe intensity variations and image noises that exist in the optical and SAR images, and thus, their feature repeatability is still poor [9]. In recent work, considering that the phase congruency (PC) [15] is robust to nonlinear intensity changes, some improved detectors, such as minimum moment of the PC with the Laplacian of Gaussian [16], modified uniform nonlinear diffusion-based Harris [17], radiation-variation insensitive feature transform (RIFT) [18], and feature intersection-based detector [19], have been proposed. These detectors have been shown to be robust for nonlinear intensity differences, but they provide limited performance due to strong speckles in SAR images [20] and huge complexity of the PC [21].

Describing the local characteristics for given point features is a key element for image registration tasks. Traditional registration methods generally explore image intensity or gradient information to produce the feature descriptors. Among them, scale invariant feature transform (SIFT)-based methods [14], [22], [23] are well known for feature description. However, because of obvious gradient reverses and noise interference (see Fig. 1), conventional descriptors using image gradients often fail to capture reliable feature attributions, leading to an inherent matching ambiguity [24]. For providing the robustness of the descriptors, a variety of descriptors are also developed for optical and SAR image registration. Among them, PC-based methods and local self-similarity (LSS) [25] based methods are two popular groups for feature description in the last decade. The PC is used to explore the local structure features, which are proven to be robust to nonlinear intensity changes [26]. Recently, Ye et al. [6], [16] developed two structure descriptors, called histogram of orientated PC (HOPC) and local HOPC, to encode structural information more robustly. Fan et al. [27] designed a robust structural descriptor based on multiscale PC features with an adaptive binning strategy. Li et al. [18] presented a maximum index map that was generated via log-Gabor convolution sequence for constructing the RIFT descriptor. Yu et al. [28] proposed the amplitudes of log-Gabor orientation histogram descriptor by using an extended PC model and an improved log-polar spatial structure. In addition, as a pioneering work, the LSS is first designed to capture the local structural features by measuring the correlations between the central patch and its neighbors, making the descriptor more robust against modality variations. Several LSS-based descriptors have been successfully applied to multimodal image registration, including dense LSS [29], dense rank-based LSS [30], max-index-based LSS [31], histogram of oriented self-similarity (OSS) [32], pyramid features of orientated self-similarity [33], OSS [34], and adjacent self-similarity (ASS) [21]. Although these PC-based and LSS-based descriptors perform well, they still have some limitations. On the one hand, the PC-based descriptors are vulnerable to strong speckle noise in SAR images [20] and suffer from large computational costs [21]. On the other hand, the LSS-based descriptors have low discriminative ability [35], leading to poor matching performance.

Unlike these traditional methods, the learning-based methods employ deep convolutional networks to extract more feature representations from images, which have made dramatic progress on multimodal image registration [36], [37]. Ma et al. [38] proposed a robust two-step registration method for multimodal images, in which the convolutional neural network features were combined with local features for feature matching. Quan et al. [39] introduced a self-distillation feature learning network for optical and SAR image registration that brought an obvious improvement in matching accuracy. Furthermore, Ye et al. [40] presented a multiscale registration framework with unsupervised learning for multimodal images to resist severe nonlinear intensity and geometric differences. Xiang et al. [41] designed a novel feature decoupling network based on a residual denoising network and a pseudo-Siamese fully convolutional network to achieve the registration of optical and SAR images. These learning frames have shown promising registration performance for multimodal images, but they still face two challenges. The first difficulty is a lack of sufficient and diverse real datasets that are used for training to obtain a satisfactory matching model. Another challenge is that their matching processes are also sensitive to large geometric differences and image noises [42].

Based on the observation that optical and SAR images present similar structural features but have quite different texture features and intensity information, this article proposes a novel and robust registration method using multi-orientation relative total variation (MORTV) structural representation, aiming to improve the registration performance of optical and SAR images. The MORTV, inspired by relative total variation (RTV) [43], is designed to extract the structural maps with multiple orientations from optical and SAR images, which can capture more image structural information and lessen the modality variations by blending the original image structure and texture. Intuitively, if the extracted structural maps of optical and SAR images have high similarity, then the registration task of these two images can be turned into a conventional image registration problem. Subsequently, we perform a feature detector and a novel feature descriptor on the multiscale structural maps constructed by the MORTV model with different parameters. Specifically, a multiscale block-based Shi–Tomasi (MBST) detector is introduced by integrating the block and nonmaximum suppression strategy into the Shi–Tomasi detector [44] to extract reliable and well-distributed feature points. Then, we design a novel layerwise multiscale histogram of oriented gradient (LMHOG) descriptor to enhance the distinctiveness and robustness of the structural descriptor. Generally, the major contributions of this study can be summarized as follows.

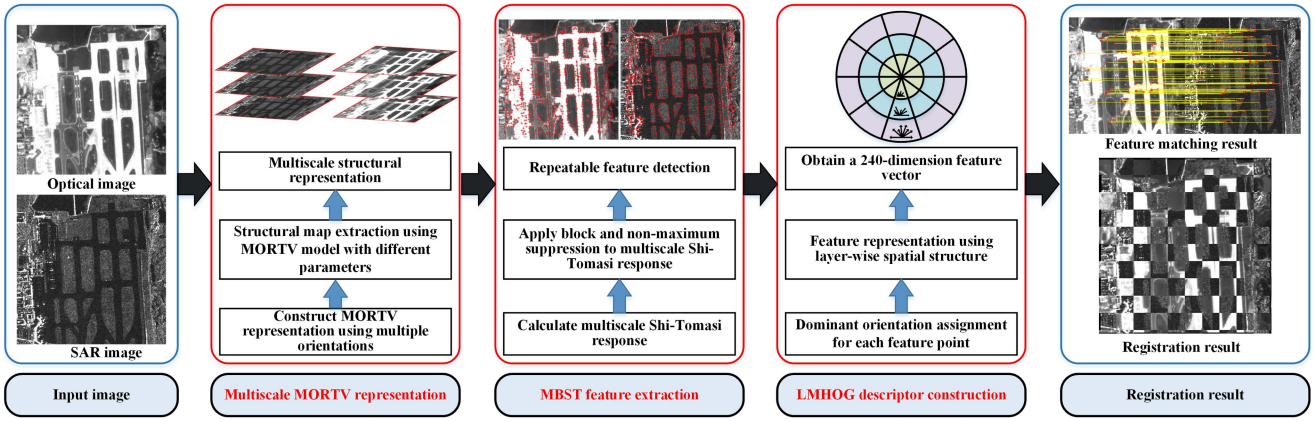

Fig. 2. Flowchart of the proposed method.

1) A novel MORTV model is designed to extract the multiscale structural maps for feature extraction and description. Different from the RTV, the MORTV identifies spatial structures in a multiple orientation manner. It captures more structural information while removing textures and noises, reducing the modality variations and further improving the similarity of optical and SAR images.
2) A novel LMHOG descriptor is constructed on multiscale MORTV representation with a multilayer manner. The LMHOG can encode multiscale structural information and enhance the robustness and distinctiveness of the descriptor without increasing its feature dimension, which is fundamentally different from the conventional multiscale descriptors.
3) A robust registration method is proposed based on the MORTV model for optical and SAR images, including two main components: the MBST detector and the LMHOG descriptor. The proposed method can significantly improve the registration performance compared with the state-of-the-art methods.

The rest of this article is organized as follows. Section II presents the whole process of the proposed registration method, including the MORTV model, MBST detector, and LMHOG descriptor. Section III illustrates the registration performance. Section IV provides a brief discussion. Finally, Section V concludes this article.

II. Proposed Registration Method

The flowchart of our proposed method is given in Fig. 2. The MORTV model is first constructed for extracting the structural maps from images. Then, based on multiscale MORTV representation, an MBST detector is introduced to extract feature points, and a novel LMHOG descriptor is designed for distinctively depicting the attributes of these detected points in a multilayer manner.

A. MORTV Structural Representation

The challenge of optical and SAR image registration lies in the inconsistency of modalities between these two images. From previous studies, it can be observed that structural features keep relative saliency under modality differences and are effective to tackle the registration of multimodal remote sensing images. Recently, the total variation (TV)-based structure extraction methods have widely applied to image processing filed, including image classification [45], texture removal [46], and image fusion [47], due to its outstanding performance in the structure–texture decomposition tasks. As an extension of the original TV model, the RTV can effectively preserve image structural information and remove noise and texture information simultaneously. According to [43], given an input image I, the RTV can be expressed as follows:

$$\mathrm{RTV}\,(p) = \frac{\mathscr{D}_x\,(p)}{\mathscr{L}_x\,(p)+\varepsilon} + \frac{\mathscr{D}_y\,(p)}{\mathscr{L}_y\,(p)+\varepsilon} \tag{1}$$

where $\mathscr{D}_x(p)$ and $\mathscr{D}_y(p)$, respectively, denote windows TVs in the x and y directions for pixel p, which measures the absolute spatial difference within the window$R(p)$. $\mathscr{L}_x(p)$ and $\mathscr{L}_y(p)$ are windows inherent variations (IV) in the x and y directions, respectively, which captures the overall spatial variation. ε is a small positive number to avoid division by zero.

We note that although the original RTV provides salient performance on structure preservation, they generally only extract the TVs within local windows of two different directions (horizontal and vertical). However, it is insufficient for these methods to use only the two orientations to capture all the structural features of images, which largely limits their applications for complex scenes. To address this issue, we propose a novel structural representation, termed MORTV, by integrating multiple orientation strategy into the original RTV model, since more orientations can enhance the local structural description ability [48]. Essentially, the MORTV aims to convert the optical and SAR images into structural maps such that more discriminative and robust structural representation can be achieved, further improving the consistency between images. For this purpose, we extend the original RTV model (1) by considering multiple orientations information and then formulate the MORTV as

$$\mathrm{MORTV}\,(p) = \sum_{o=1}^{O} \frac{\mathscr{D}_o\,(p)}{\mathscr{L}_o\,(p)+\varepsilon} \tag{2}$$

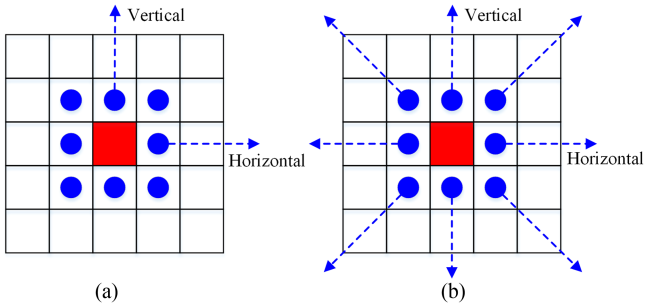

Fig. 3. Comparison of different orientation strategy. (a) Two orientations in RTV. (b) Eight orientations in MORTV.

where O is the number of orientations. $\mathscr{D}_o(p)$ denotes window TV at orientation o for pixel p

$$\mathscr{D}_o(p) = \sum_{q \in R(p)} g_{p,q} \left| (\partial_o S)_q \right| \tag{3}$$

where q belongs to $R(p)$, the local window centered at the pixel p. $g_{p,q}$ is a spatial function with standard deviation σ. ∂_o represents the partial derivative at orientation o. S is the resulting structural map; $\mathscr{L}_o(p)$ denotes window IV at orientation o for pixel p

$$\mathscr{L}_o(p) = \left| \sum_{q \in R(p)} g_{p,q} (\partial_o S)_q \right|. \tag{4}$$

In particular, we set $O = 8$ in this study since eight orientations have been shown to represent spatial structural information better [48]. The different orientation strategies used in RTV and MORTV are illustrated in Fig. 3. As seen, the MORTV produces the structural map based on eight different orientations (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°). Such an operation facilitates the preservation of structural features sufficiently.

After constructing the MORTV regularizer, a new objective function can be defined to capture the structural map from images by

$$\arg\min_S \sum_p \left\{ (S_p - I_p)^2 + \lambda \cdot \mathrm{MORTV}(p) \right\} \tag{5}$$

where the term $(S_p - I_p)$ is to make the input image and the extracted structural map similar. λ is a weight. The term MORTV(p) is introduced to preserve the main structures while removing textures. Finally, we adopt the same optimization method presented in [43] to solve (5) for obtaining the desired structural map.

To illustrate the advantage of MORTV, the MORTV and RTV are, respectively, implemented with the following same parameter settings to an optical and SAR dataset with 200 image pairs: $\{\lambda, \sigma\} = \{0.01, 3\}$. A simple comparison is presented in Fig. 4. It is clearly observed that MORTV effectively extracts more significant structural features while removing the image noises and textures, which could be helpful for the following feature detection and description. Conversely, RTV blurs some of the structure edges to a certain extent [see red rectangular regions in Fig. 4(b)]. In this way, the obtained structural maps should be more similar than the original images, namely, the modality differences between optical and SAR images can be weakened as much as possible. Furthermore, we employ the MI and structural similarity (SSIM) [49] metrics to measure the similarity between the original images (structural maps obtained by MORTV and RTV). Table I compares the average MI and SSIM of different images. Large MI and SSIM imply a higher image similarity. We can see that RTV achieves better similarity values compared with original images but remains inferior to those of MORTV, possibly because RTV only utilizes two different orientations for capturing image structure. In contrast, our MORTV provides higher average MI and SSIM than those of RTV and original images, clearly demonstrating the outstanding performance of our MORTV.

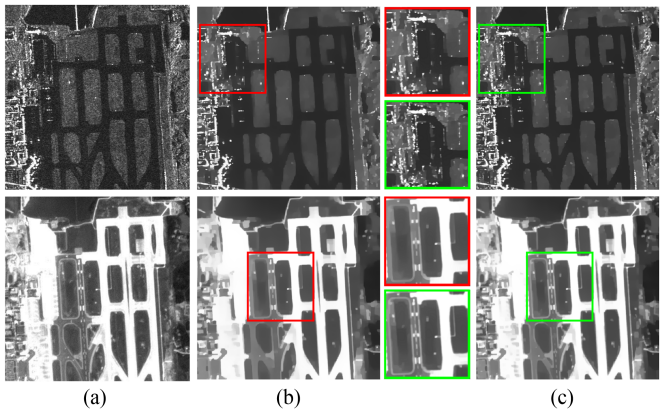

Fig. 4. Comparison of structural map extraction methods. (a) Original images. (b) RTV. (c) MORTV.

TABLE I
AVERAGE MI AND SSIM VALUES OF DIFFERENT IMAGES

Criterion	Original	RTV	MRTV
MI	0.21	0.39	0.46
SSIM	0.12	0.38	0.41

B. Multiscale MORTV Representation

Structure preservation and noise suppression are the main concern in the design of image multiscale representation. Commonly used multiscale representations are constructed with Gaussian smoothing (GS), nonlinear diffusion filter (NDF) [24], [27], co-occurrence filter (COF) [50], and rolling guidance filter (RGF) [51]. Nevertheless, due to the modality variations between optical and SAR images, the GS and NDF cannot provide consistent structural information while maintaining the robustness to noises [50]. For COF and RGF, their main deficiency is the computation overhead [50], [51]. To address that, here the MORTV is employed with different parameter settings to construct multiscale representation for feature extraction and description.

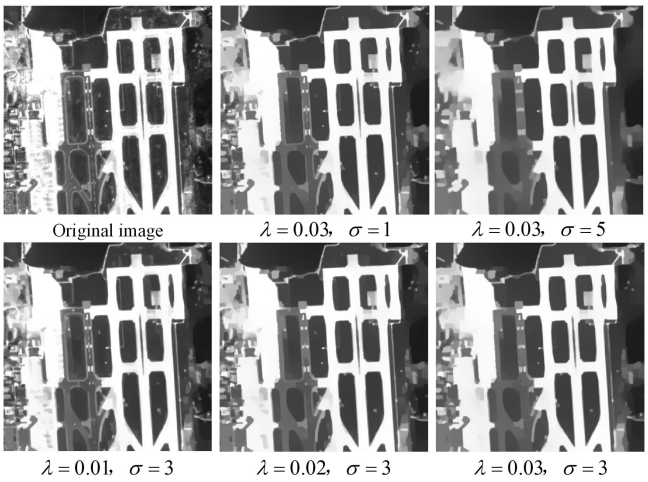

Fig. 5. Effect of different parameters on the structural maps.

Similar to RTV, the performance of MORTV relies primarily on two parameters: the weight λ and the standard deviation σ. Specifically, λ controls the degree of smoothness of the structural map. Increasing its value blurs image structures without considering texture removal. σ determines the spatial scales and altering its value maintains salient structures while separating textures. The influence of these two parameters on the structural maps is shown in Fig. 5. As shown, MORTV produces different structural maps with different parameter settings. The smaller the values of λ andσ, the more fine structural features it can provide. In contrast, the larger the values of these two parameters, the more coarse structural features they can present. Motivated by this observation, the multiscale structural representation is designed by performing MORTV with different values of λ and σ as follows:

$$S_l = \mathrm{MORTV}\left(\mathrm{I}, \lambda_l, \sigma_l\right), \mathrm{l} = 0, 1, \ldots, \mathrm{L}-1 \tag{6}$$

where S_l represents the lth layer structural map. L denotes the layer number and is suggested to $L \leq 8$. λ_l and σ_l are the lth parameters employed in the MORTV. For each layer l, λ_l, and σ_l are calculated as

$$\begin{cases} \lambda_l = \lambda_0 \cdot \sqrt[3]{2^l} \\ \sigma_l = \sigma_0 \cdot \sqrt[3]{2^l} \end{cases}, l = 0, 1, \ldots, L-1 \tag{7}$$

where λ_0 and σ_0 denote the weight and spatial scale of the first layer, respectively. To achieve better multiscale representation, we set $\lambda_0 = 0.005$ and $\sigma_0 = 1.2$ for the default values.

C. MBST Feature Detection

After constructing the multiscale MORTV representation, we perform an MBST detector to obtain repeatable and well-distributed feature points for optical and SAR images. As an improved Harris detector, the Shi–Tomasi detector has been shown to extract sufficient and reliable feature points for multimodal images [50]. However, as with many conventional feature detectors, the Shi–Tomasi detector also suffers from some problems in the distribution of feature points when directly applied to optical and SAR images. To detect reliable feature points with uniform distribution, we incorporate the block and nonmaximum suppression strategy with the Shi–Tomasi detector. Specifically, we first build a multiscale image pyramid by stacking the first layer of multiscale MORTV maps and the downsampled versions (a factor of 2) of the other layers. Then, we partition each layer into $m \times m$ nonoverlapping blocks, and perform the nonmaximum suppression on feature points obtained in each block according to the Shi–Tomas responses. Finally, we select the first h points with the highest Shi–Tomas values as the feature points.

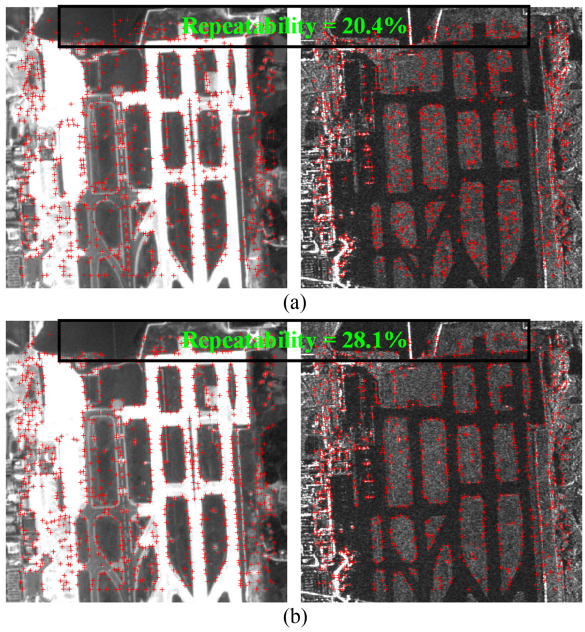

Fig. 6. Comparison of the feature detection between GSST and MBST with the same feature selection strategy. (a) GSST (repeatability = 20.4%). (b) MBST (repeatability = 28.1%).

To illustrate the performance of our MBST, the traditional multiscale Shi–Tomasi detector using the GS (denoted as GSST) is introduced for comparison. We perform these two detectors on a pair of optical and SAR images to obtain 1000 feature points, as illustrated in Fig. 6. As seen, MBST performs better than GSST with the same feature selection strategy (e.g., block and nonmaximum suppression). The repeatability [23] of MBST has about 8% improvement compared with GSST. Therefore, with multiscale MORTV representation, MBST is capable of extracting repeatable and well-distributed feature points between optical and SAR images.

D. LMHOG Descriptor Construction

For optical and SAR image registration, although multiscale feature descriptors are capable of capturing image information at different scales and enable them to encode image features more distinctively, they frequently suffer from a considerably higher feature dimension [27], leading to an increase in computational complexity. To address that, here we design a novel LMHOG descriptor on the multiscale MORTV representation with a multilayer manner to describe each feature point. The proposed

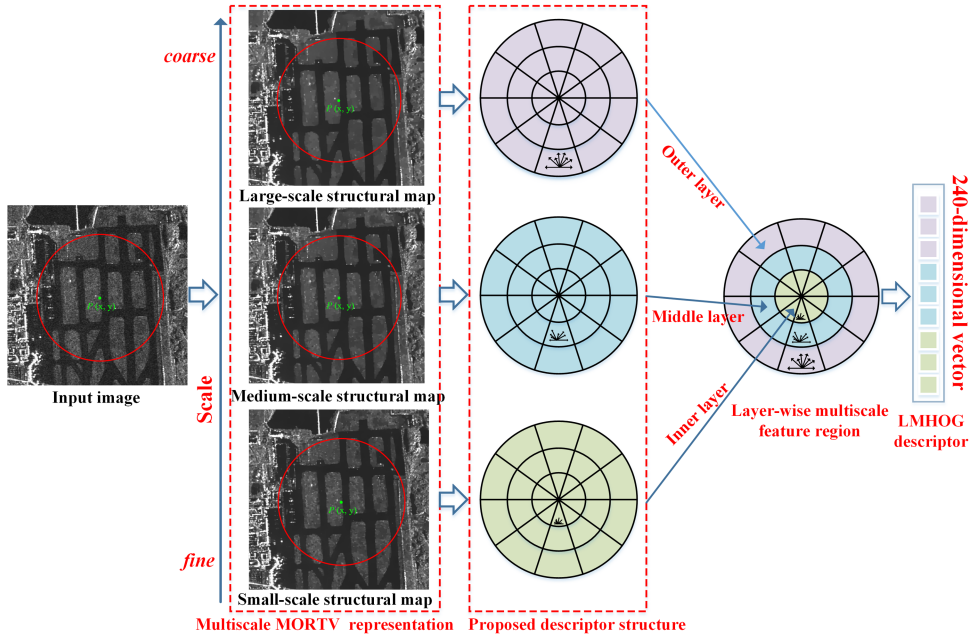

Fig. 7. Schematic of the LMHOG descriptor.

LMHOG descriptor consists mainly of two steps: orientation assignment and descriptor construction.

1) Orientation Assignment: Similar to the SIFT-like descriptors, we also use the gradient orientations to assign a dominant orientation for each feature point to maintain rotation invariance. However, our orientation assignment differs from conventional gradient-based methods. Specifically, we estimate the dominant orientation on the MORTV structural maps, while traditional assignments directly estimate on the original images. Since the MORTV largely filter out textures and noises, we can obtain a more robust estimate of the dominant orientation. Moreover, to handle the intensity inversion between structural maps, we transfer the gradient orientation from $[0°, 360°)$ to $[0°, 180°)$.

2) Descriptor Construction: Unlike these conventional multiscale descriptors, our LMHOG descriptor encodes the structural features at different scales in a multilayer manner. Specifically, the feature region (a radius of $\beta\sigma_l$) is divided into three nested nonoverlapping layers, spreading from the feature point to the edge of the feature region. Instead of using fixed-scale feature region, we treat each layer within the feature region from different scales. In other word, we capture the fine-scale structural information for the layer closest to the center of the feature region. Then, we gradually utilize larger and larger scales to capture the coarse-scale structural information with increasing distance to the region center. Intuitively, the coarsest structural information is captured for the layer closest to the boundary of the feature region. In this way, a single feature region can convey image structural information from multiple scales. More importantly, such a procedure does not increase the dimension of the descriptor, which is the main difference between our LMHOG and traditional multiscale descriptors.

Fig. 7 illustrates the processing of how to integrate the multiscale feature description into a single feature region. First, we produce three structural maps with different scales by using MORTV with different parameter settings (i.e., λ_l and σ_l) on the original image. Then, for each scale layer, we divide the feature region into a log-polar grid with three radial bins and ten angular bins, resulting in 30 grids. Subsequently, we calculate an eight-bin (covering 180°) gradient orientation histogram based on weighted gradient magnitudes for each grid. Finally, we separately replace the original histogram values with the new values in the smoothed structural maps at the same locations. Specifically, we replace the grid values in the middle layer with the histogram values obtained in the medium-scale structural map. Similarly, for each grid in the outer layer, we utilize the histogram values from the large-scale structural map. All the orientation histograms of the 30 grids from three different layers are concatenated into one augmented feature vector which contains all structural information from different scales. Letting $V_k(k = 1, 2, \ldots, K)$ represents the$k$th layer structural vector, the LMHOG descriptor is thus defined as follows:

$$\begin{aligned} \text{LMHOG} &= \{V_1, V_2, \ldots, V_K\} \\ V_k &= \text{GO\_HIST}(R_k) \end{aligned} \tag{8}$$

where GO_HIST$(\cdot)$ denotes the histogram of orientation gradient. R_k denotes the kth layer feature region around a given feature point.K represents the number of layers of the multiscale feature region, which is fixed as 3. As a result, the dimension of the LMHOG descriptor is $3 \times 10 \times 8 = 240$. Herein, the LMHOG feature vector is required to normalize for gaining invariance to intensity variation.

In summary, compared with the conventional multiscale descriptors, the proposed LMHOG descriptor offers three main advantages. First, LMHOG has better robustness to image noises since it is built on the multiscale MORTV representation, which

can effectively remove the image noises and textures. Second, and most importantly, LMHOG has lower feature dimension by using a multiple layer strategy, making our descriptor suitable for practical applications. Moreover, considering that multiscale MORTV representation has the advantages of complementary structural extraction and noise robustness, LMHOG has better feature distinctiveness. As a result, LMHOG has an ability to identify more reliable matches than the conventional descriptors.

E. Feature Matching

Herein, the initial matches between optical and SAR images are established by performing the nearest neighbor distance ratio matching strategy. To improve the matching reliability, the fast sample consensus [52] method is utilized to remove the matching outliers.

III. Experimental Results and Discussion

In this section, we evaluate the proposed method on two publicly available optical and SAR image datasets through both quantitative and qualitative analysis. The performance of the proposed method is compared with some state-of-the-art methods for registration, including OS-SIFT [14], ASS [21], histogram of the orientation of weighted phase (HOWP) [24], locally normalized image feature transform (LNIFT) [42], and multiscale histogram of local main orientation (MS-HLMO) [53]. OS-SIFT, MS-HLMO, and LNIFT are three recent advanced gradient-based methods, whereas HOWP is a robust PC-based method. Moreover, ASS is an advanced extension of the LSS.

A. Experimental Data

Two real optical and SAR image datasets, namely, high-resolution dataset and medium-resolution dataset, are employed to evaluate the performance of the proposed method. All image pairs have significant nonlinear intensity differences. The specific description of experiment data is presented as follows.

1) *High-Resolution Optical-SAR (HROS) Dataset*: The image pairs in the HROS dataset are derived from the optical-to-SAR (OS) dataset [54], which consists of 2673 registered optical and SAR image pairs of 512 × 512 pixels with a high resolution of 1 m. The optical images are obtained from Google Earth and the SAR images are acquired with the GaoFen (GF)-3 satellite. These pairs cover several cities around the world, such as Shanghai, Dengfeng, Renne, Tucson, and Agra. In our experiment, the HROS dataset contains three groups of experimental data for different evaluation purposes. In the first group, 200 image pairs with no rotation and scale differences compose the experimental data, which is randomly selected from the OS dataset. In the second group, 400 image pairs are randomly selected. Each pair is rotated clockwise or counterclockwise with a random angle, which is limited in a range of $[0°, 90°]$. Another 400 image pairs are also randomly selected as the third group. Each pair is scaled with a random scale ratio, which is limited in a range of $[0.5, 2]$. For all three groups, we can establish the ground truth transformation according to the rotation angles and scale factors.
2) *Medium-Resolution Optical-SAR (MROS) Dataset*: The image pairs in the MROS dataset are derived from the medium resolution OS dataset [40], which has 5800 registered optical and SAR image pairs with a high resolution of 10 m. These image pairs with the size of 512 × 512 are acquired by Sentinel-1 and Sentinel-2 in May 2021. These images cover different scenes, including rivers, forests, farmland, and urban. Similar to the HROS dataset, the MROS dataset consists mainly of two groups of experiment data. The first group contains 300 image pairs, which are randomly selected and have no geometric differences. For the second group, we randomly choose 500 image pairs and then simultaneously perform random rotation and scale transforms on them. The ranges of rotation and scale parameters are the same as the HROS dataset. Based on the parameters above, we can also establish its ground truth transformation for both groups.

B. Evaluation Criteria and Parameter Settings

Four quantitative evaluation criteria are employed for a comprehensive evaluation, including the success rate (SR) [42], the number of correct matches (NCM), the correct matching ratio (CMR), and the root mean square error (RMSE). SR is used to measure the matching SR, which can be defined as follows:

$$\mathrm{SR} = \frac{1}{T}\sum_{n} F\left(I_n\right) \times 100\% \tag{9}$$

$$F\left(I_n\right) = \begin{cases} 1, \mathrm{NCM}(I_n) \geq 5 \\ 0, \mathrm{else} \end{cases} \tag{10}$$

where $F(I_n)$ denotes a logical value, 1 represents a correct matching trail while 0 denotes a failed matching trail. $\mathrm{NCM}(I_n)$ is the NCM of the *n*th image pair. T is the total number of image pairs for each test group. NCM, CMR, and RMSE are used to measure registration performance. Here, CMR is expressed as CMR = NCM/TC, where TC denotes the total number of matches. The matching point pair with a location error less than three pixels is considered as the NCM. Higher values of SR, NCM, and CMR, and lower value of RMSE indicate a better registration performance. Note that RMSE is set to ten pixels if one image pair fails to register.

Our proposed method is composed of three main stages, i.e., MORTV representation, MBST feature extraction, and LMHOG descriptor construction. In the first stage, the number of orientations is set to eight since eight orientations can better represent spatial structural information, while other parameters are drawn from the original RTV paper [43]. We also follow the parameter settings in a multiscale representation study to set $\lambda_0 = 0.005$ and $\sigma_0 = 1.2$. During the MBST feature extraction stage, the parameters are the same as those in the uniform nonlinear diffusion-based Harris detector [26]. For the LMHOG descriptor construction stage, four key parameters directly influence the performance of the descriptor, i.e., $\beta\sigma_l$,K, N_r, and N_a.

TABLE II
ILLUSTRATION OF PARAMETERS SETTINGS

Parameters	Variable	Fixed parameters
β	$\beta=[12,13,14,15,16]$	$K=3, N_r=3, N_a=10$
K, N_r	$K, N_r=[2,3,4,5]$	$\beta=15, N_a=10$
N_a	$N_a=[6,8,10,12,14]$	$\beta=15, K=3, N_r=3$

TABLE III
AVERAGE NCM AND SR AS β VARIES FROM 12 TO 16

Criterion	$\beta, K=3, N_r=3, N_a=10$				
	12	13	14	15	16
NCM	194.2	205.6	220.4	232.6	228.7
SR(%)	92.5	95	99	100	100

TABLE IV
AVERAGE NCM AND SR AS N_r, K VARIES FROM 2 TO 5

Criterion	$N_r, K, \beta=15, N_a=10$			
	2	3	4	5
NCM	201.8	232.6	230.3	228.5
SR(%)	90	100	100	100

TABLE V
AVERAGE NCM AND SR AS N_a VARIES FROM 6 TO 14

Criterion	$N_a, \beta=15, K=3, N_r=3$				
	6	8	10	12	14
NCM	210.2	221.8	232.6	233.1	230.6
SR(%)	92	98	100	100	100

$\beta\sigma_l$ is the radius of the local region for feature description. A too small β makes descriptor retain insufficient information, leading to poor distinctiveness. On the contrary, when β is too large, the descriptor will be less distinctive. K is the number of layers of multiscale feature regions. N_r and N_a determine the spatial structure of LMHOG. The larger the N_r or N_a is, the higher the dimension of LMHOG, and the more image information that it contains, which improves the performance of the LMHOG but reduces its efficiency. Conversely, a small N_r or N_a can improve the efficiency of the LMHOG but at the cost of degraded performance and distinctiveness. Herein, we respectively select 100 image pairs from the HROS and MROS datasets to compose a test dataset and then implement a series of independent experiments on it to discuss the influence of these parameters. Note that we set the other parameters to fixed values when a particular parameter is analyzed, as given in Table II. The results are depicted in Tables III–V, where NCM and SR are used as the evaluation criteria.

We have three main observations from Tables III–V. First, the proposed method produces the highest SR and average NCM when β is 15. When $\beta > 15$, the average NCM decreases. Second, larger $N_r(K)$ obtains a better SR and average NCM. It is noted that the performance gradually decreases when $N_r(K) > 3$. Hence, we set $N_r(K) = 3$. Third, the SR and average NCM gradually increase with an increasing N_a. Nevertheless, N_a greater than 10 ($N_a > 10$) does not bring a significant increase of the average NCM and even decreases the results. Based on the above observations, $\beta = 15$, $K = 3$, $N_r = 3$, and $N_a = 10$ are set to be the default parameters in this work.

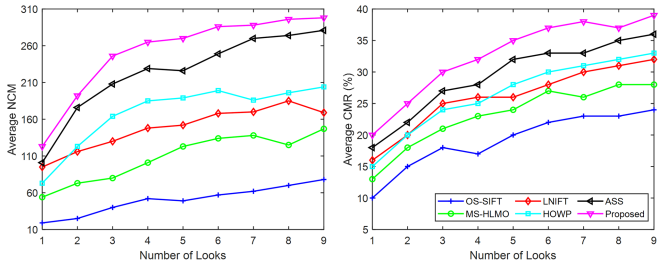

Fig. 8. Average NCMs and CMRs under different numbers of looks for six comparative methods.

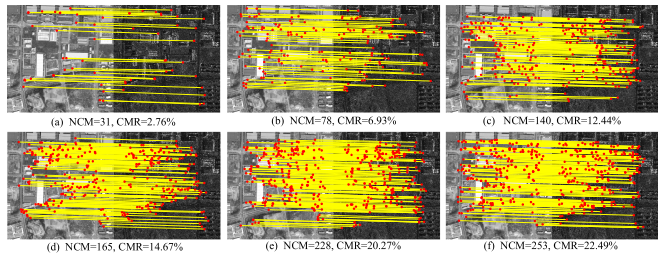

Fig. 9. Feature matching results between the original optical image and the simulated SAR image with three-look speckles. (a) OS-SIFT. (b) MS-HLMO. (c) LNIFT. (d) HOWP. (e) ASS. (f) Proposed.

C. Analysis of Noise Robustness

To evaluate the noise robustness of the proposed method, we implement a series of experiments on ten groups of simulated image pairs. For each image pair, a real optical image is registered with simulated SAR images with different noise levels, which is characterized by the number of looks, ranging from one-look to nine-look. In general, a small look refers to the high noise level in the simulated SAR image. The average NCMs and CMRs obtained by six methods under various noise conditions are presented in Fig. 8. It can be observed that, as the look number becomes larger, all the methods suffer from obvious performance degradation, while the proposed method still has higher performance under increasing noise, i.e., average NCM = 123 and average CMR = 20%, when the original SAR images are corrupted by the speckle noise with one-look. This demonstrates that the proposed method exhibits much better robustness to speckle noise than other compared methods. The reason is that the proposed method is built on the MORTV, which makes the feature detection and description process more robust against speckle noise. For better illustration, Fig. 9 presents the feature matching results of different methods on an original optical

TABLE VI
COMPARATIVE RESULTS OF SIX METHODS ON THE HROS DATASET

HROS dataset	Criterion	Method					
		OS-SIFT	MS-HLMO	LNIFT	HOWP	ASS	Proposed
Group 1	NCM	32.2	49.5	120.7	108.7	131.6	140.8
	CMR (%)	11.6	17.5	23.4	22.9	28.6	31.3
	SR (%)	92.5	100	100	100	100	100
	RMSE (pixels)	6.16	2.81	2.40	2.15	2.08	1.94
Group 2	NCM	16.4	35.7	48.9	50.2	110.5	115.3
	CMR (%)	9.8	16.7	20.6	20.8	28.2	29.9
	SR (%)	52.5	87.5	93.8	88.8	97.5	99.5
	RMSE (pixels)	8.53	3.43	2.57	3.12	2.25	2.06
Group 3	NCM	16.2	37.8	42.6	44.7	78.7	85.6
	CMR (%)	9.7	17.2	18.5	20.6	28.2	28.6
	SR (%)	58.5	88.0	59.3	94.8	99.0	100
	RMSE (pixels)	8.44	3.05	5.13	2.56	2.10	1.95
Average value	NCM	21.6	41.0	70.7	67.9	106.9	113.9
	CMR (%)	10.4	17.1	20.8	21.4	28.3	29.9
	SR (%)	67.8	91.8	84.4	94.5	98.8	99.8
	RMSE (pixels)	7.71	3.10	3.37	2.61	2.14	1.98

image and a simulated SAR image with a three-look. As seen, the proposed method still obtains the best performance in terms of NCM and CMR.

D. Registration Performance

In this section, we evaluate the proposed method on the HROS and MROS image datasets, in comparison with five other state-of-the-art methods including OS-SIFT, MS-HLMO, LNIFT, HOWP, and ASS. For these two datasets, we regard the optical image as the reference image, and generate the sensed image by imposing the simulated transformation on SAR images. The implementations of all the comparison method are downloaded from the personal website of their authors. For each of these methods, its best performance is presented according to the optimal parameters provided by the authors. In addition, we fix the number of extracted feature points to 4000 for all methods during our experimental stage. The registration results on two datasets are as follows.

1) *Comparison results on the HROS Dataset*: In this dataset, we conduct three groups of experiments to evaluate the performance of the proposed method. The quantitative results for the six comparison methods are given in Table VI. In the first group, except for OS-SIFT, each of these methods can achieve an average SR of 100%. OS-SIFT obtains the lowest SR of 92.5% since it is a gradient-based method that is vulnerable to intensity differences and speckle noise. In the second group, the proposed method performs better than the other five methods and the average SR is 99.5%, which is closely followed by ASS and LNIFT. The remaining methods can only produce a lower average SR than 90%. In the third group, the best performance is obtained by the proposed method, which achieves the average SR as 100%. By contrast, the average SR of LNIFT is only 59.3%, which indicates that LNIFT can only be employed to register the images with small scale differences.

As given in Table VI, the proposed method yields better average performance for the HROS dataset than the other five comparison methods. The obtained average SR over the entire HROS dataset is 99.8% by the proposed method, 98.8% by ASS, 94.5% by HOWP, 84.4% by LNIFT, 91.8% by MS-HLMO, and 67.8% by OS-SIFT. Our proposed method has a maximum improvement of 32 percentage points. In addition, the average NCM of the proposed method is 113.9, whereas the average NCM of HOWP is 67.9. Our method identifies about two times as many correct matches as HOWP. We can also observe that the proposed method obtains better average CMR and RMSE (29.9%, 1.98 pixels) than ASS (28.3%, 2.14 pixels), HOWP (21.4%, 2.61 pixels), LNIFT (20.8%, 3.37 pixels), MS-HLMO (17.1%, 3.10 pixels), and OS-SIFT (10.4%, 7.71 pixels). All the results indicate that the proposed method is effective to handle the registration task of optical and SAR images.

To visually investigate the performance of different methods, Figs. 10 and 11 illustrate the comparative matching results on eight image pairs that are randomly selected from three subsets. The corresponding registration results of the proposed method are presented in Fig. 12. As seen, the proposed method achieves satisfactory matching (see Fig. 11) and registration results (see Fig. 12), thereby demonstrating that the proposed method is more robust than the other methods in addressing image noises, intensity changes, and geometric differences. OS-SIFT, as shown in Fig. 10(a), performs the worst for all the images.

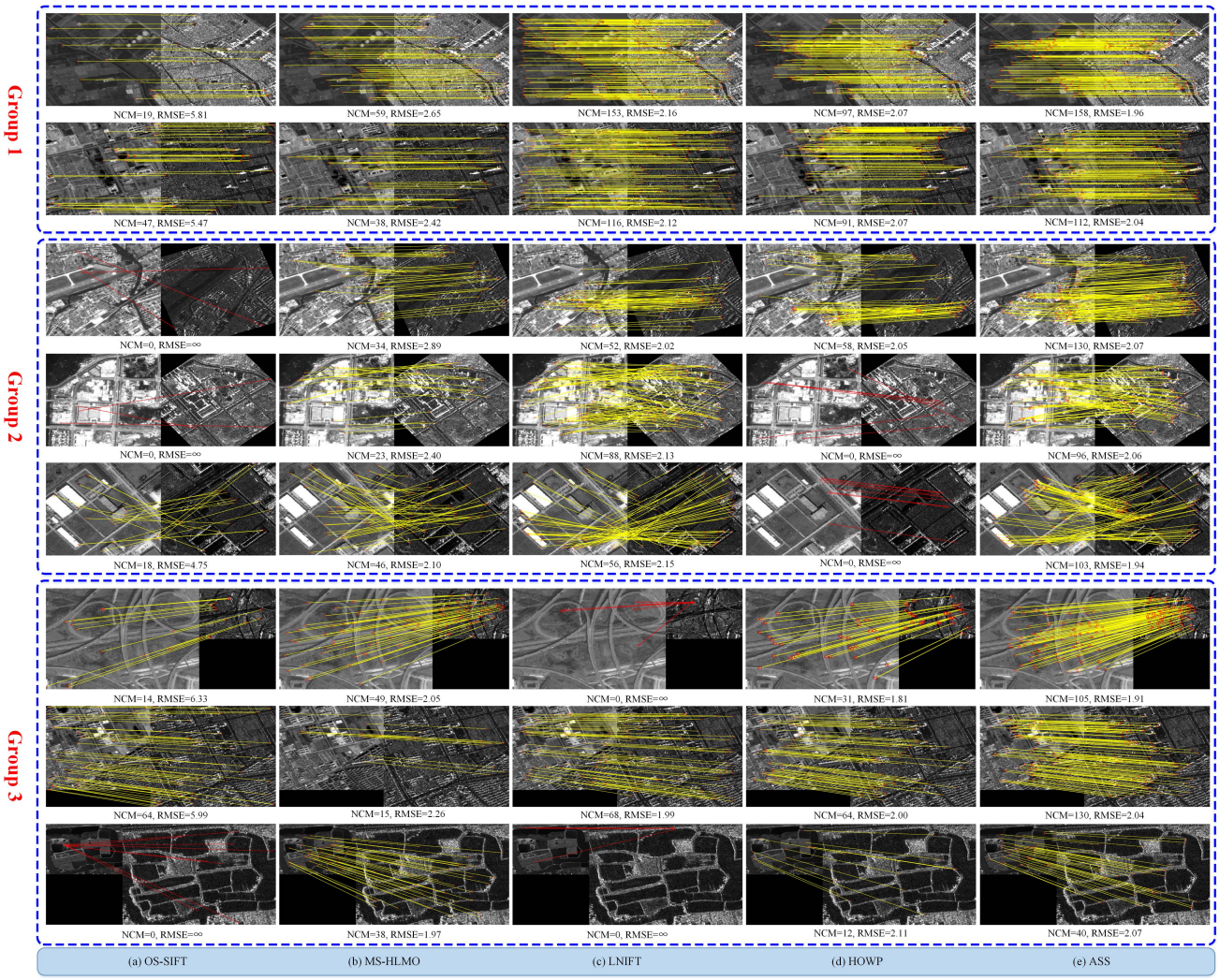

Fig. 10. Feature matching of the comparison methods on the HROS dataset. The yellow lines and red lines denote correct matches and false matches, respectively. RMSE $= \infty$ indicates the method fails to register this image pair.

It fails to match the image pairs with obvious intensity differences and large geometric differences. Fig. 10(b) and (c) provides the matching results of MS-HLMO and LNIFT. The former is based on the histogram of the local main orientation feature, and the latter is based on the local normalized filter. Both operations make these two methods robust to nonlinear intensity differences. However, MS-HLMO and LNIFT are still sensitive to speckle noise since they both utilize gradient information for feature description. In addition, without a multiscale strategy, LNIFT is not invariant to scale differences. From Fig. 10(d), it can be seen that HOWP produces favorable results in coping with intensity, scale differences, and small rotation differences, but it still cannot effectively address large rotation differences. Fig. 10(e) shows that the performance of ASS far outperforms the other methods but remains inferior to those of the proposed method. The results imply that the proposed method exhibits superior adaptability to modality differences and speckle noise compared with the other methods.

2) *Comparison results on the MROS Dataset:* In this dataset, we conduct two groups of experiments to evaluate the performance of the proposed method. The comparative results for the six methods are given in Table VII. In the first group, both OS-SIFT and MS-HLMO have a smaller average NCM (35.3 for OS-SIFT and 53.6 for MS-HLMO) than other methods, possibly because these two methods perform feature detection on the original images, resulting in a low-feature repeatability. In comparison, the proposed method has an average NCM of 142.5, which is more than four times as many as OS-SIFT and almost three times as many as MS-HLMO. In the second group, all these methods decrease their average NCM by different degrees due to complex geometric differences. It can be seen that the gradient-based methods, such as OS-SIFT and LNIFT, are not reasonably effective, thus they can only obtain the

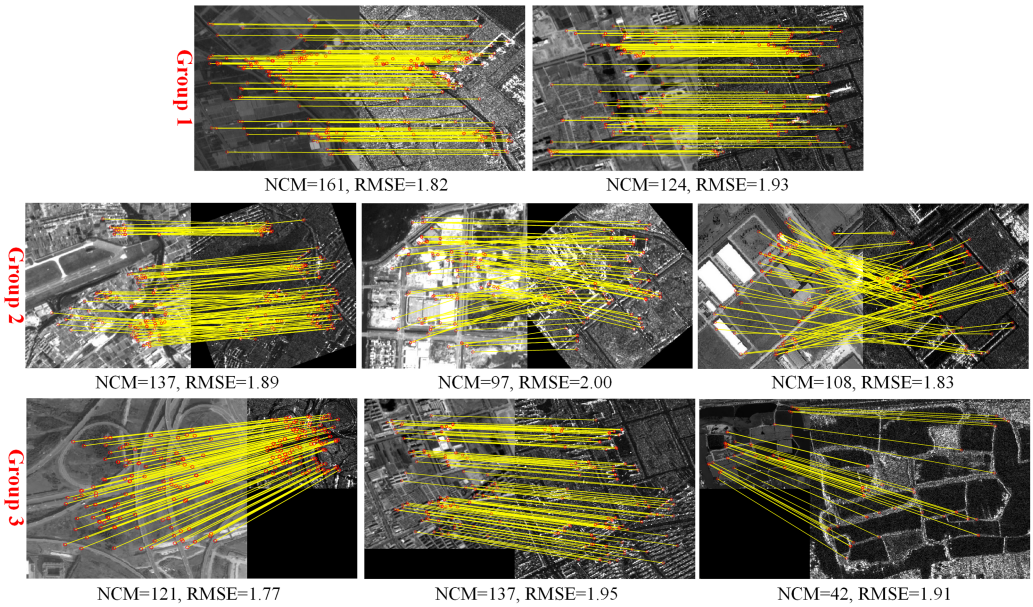

Fig. 11. Feature matching of the proposed method on the HROS dataset.

TABLE VII
COMPARATIVE RESULTS OF SIX METHODS ON THE MROS DATASET

MROS dataset	Criterion	Method					
		OS-SIFT	MS-HLMO	LNIFT	HOWP	ASS	Proposed
Group 1	NCM	35.3	53.6	109.5	110.4	125.7	142.5
	CMR	13.5	20.1	23.6	24.8	30.4	35.2
	SR	95.3	100	100	100	100	100
	RMSE	5.83	2.73	2.10	2.06	2.02	1.91
Group 2	NCM	12.7	34.6	31.7	47.5	73.4	89.8
	CMR	10.2	19.0	20.5	23.1	28.3	31.3
	SR	38.5	48.8	43.6	62.4	97.2	99.6
	RMSE	7.96	6.38	6.52	5.01	2.18	2.02
Average value	NCM	24.0	44.1	70.6	79.0	99.6	116.2
	CMR	11.9	19.6	22.1	24.0	29.4	33.3
	SR	66.9	74.4	71.8	81.2	98.6	99.8
	RMSE	6.90	4.56	4.31	3.54	2.10	1.97

average NCM of 12.7 and 21.7, respectively. By contrast, the proposed method still achieves the best performance and has an average NCM of 89.8.

In Table VII, we further illustrate the average registration accuracy on the MROS dataset. It is observed that the proposed method produces the best average registration accuracy, with an average NCM of 116.2, average CMR of 33.3%, average SR of 99.8%, and average RMSE of 1.97 pixels. It is worth noting that ASS obtains comparable performance on these two subsets and has an average NCM of 99.6, with a mean CMR of 29.4%, a mean SR of 98.6%, and a mean RMSE of 2.10 pixels. HOWP performs feature detection and description using the PC information, it is, thus, more robust to intensity and geometric differences and obtains a mean NCM of 79.0, with a mean CMR of 24.0%, a mean SR of 81.2%, and a mean RMSE of 3.54 pixels. Since LNIFT only considers the rotation differences in the feature extraction stage, it performs a bad performance on the image pairs with both rotation and scale variations, and only gets the average SR of 71.8%. In addition, OS-SIFT and MS-HLMO are also not applicable to the MROS dataset. The average SRs are only 66.9% and 74.4%, respectively.

The feature matching results are also provided to validate the effectiveness of the proposed method, as shown in Figs. 13 and 14. The corresponding registration results of the proposed method are presented in Fig. 15. As shown in Group 1, all the methods can establish sufficient NCM and correctly match these two image pairs only with intensity variations. For Group 2, it is more challenging than Group 1. As seen, OS-SIFT, MS-HLMO,

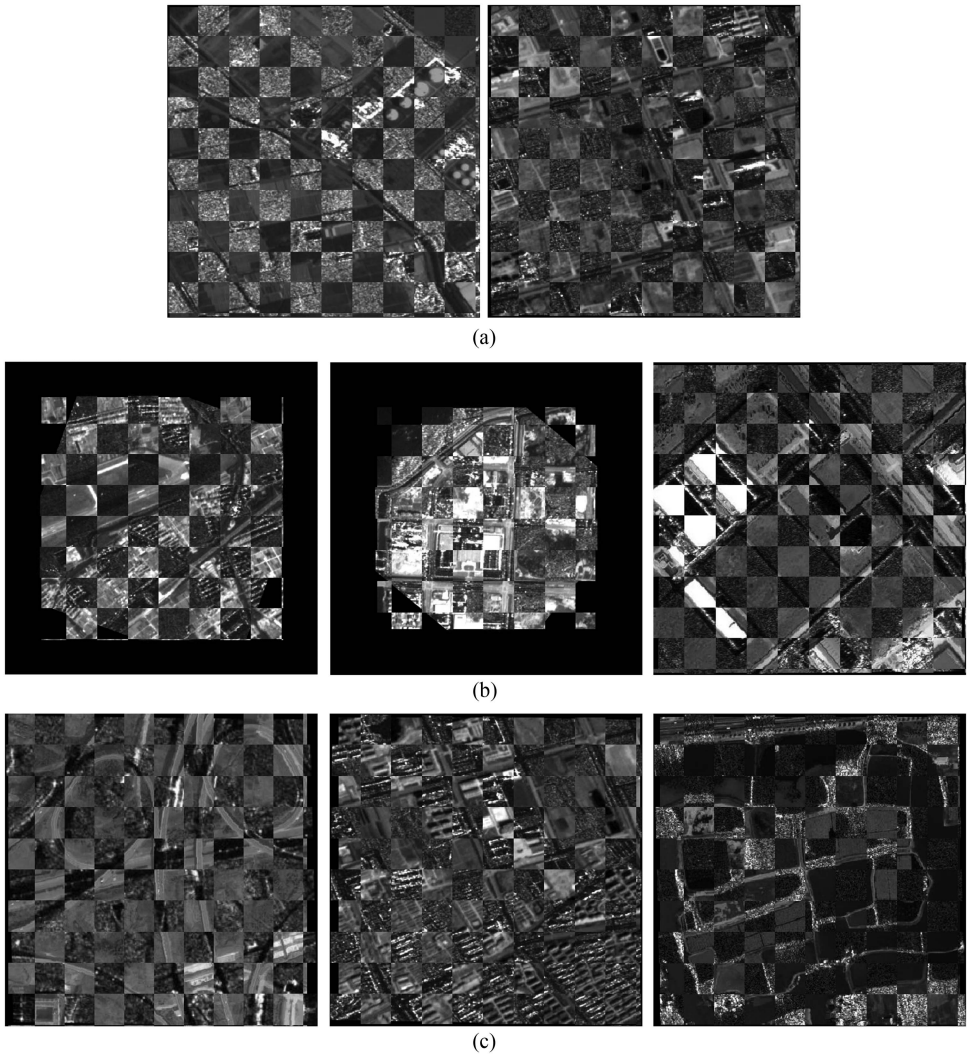

Fig. 12. Registration results of the proposed method on the HROS dataset. (a) Group 1. (b) Group 2. (c) Group 3.

and LNIFT fail to register the image pairs with both larger rotation and scale changes, which indicates that gradient-based methods cannot capture distinctive common features on this challenging dataset. HOWP still has limited performance in handling large rotation differences despite its robust to nonlinear intensity and scale differences, possibly because HOWP is vulnerable to speckle noise in the SAR image, leading to inaccurate PC orientations. Both ASS and the proposed method achieve good performance on all the image pairs. One can clearly see that the proposed method achieves a larger NCM and a lower RMSE compared with ASS.

These experimental results presented in the HROS and MROS datasets demonstrate that the proposed method achieves state-of-the-art performance for optical and SAR image registration. There are two factors contributing to the improved performance. First, we convert the optical and SAR images into their corresponding structural maps based on the MORTV, which captures more structural information while removing noises and textures, and further decreases the modality differences. Second, based on the multiscale MORTV representation, MBST has an advantage in extracting sufficient feature points with high repeatability and uniform distribution, and LMHOG can encode multiscale structural information in a multilayer manner, which significantly improves the robustness and discrimination of the final feature representation without increasing the dimension of the descriptor.

IV. Discussion

A. Adaptability to Complex Image Scenes

To further evaluate the effectiveness of the proposed method, we implement more experiments on three pairs of optical and SAR images with complex scenes. Pair 1 contains two images of GF-1 and GF-3 in Fuzhou, China. This image pair covers a river and several mountains, and has an obvious scale difference. Pair 2 also covers a mountainous area, where optical image is from the Google Earth map and SAR image is from GF-3. The image pair has significant rotation difference. Pair 3 covers a desert area, which is obtained from the website of NASA. This pair also has rotation difference and the optical

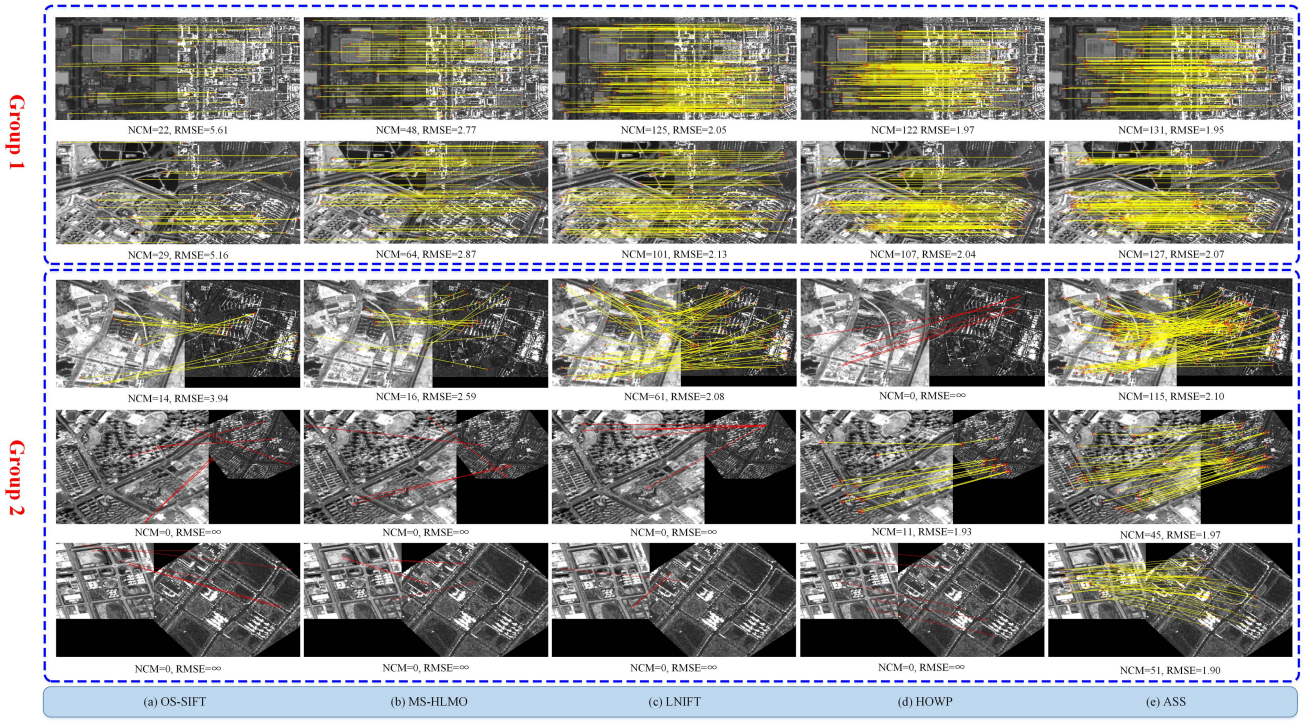

Fig. 13. Feature matching of the comparison methods on the MROS dataset. The yellow lines and red lines denote correct matches and false matches, respectively. RMSE $= \infty$ indicates the method fails to register this image pair.

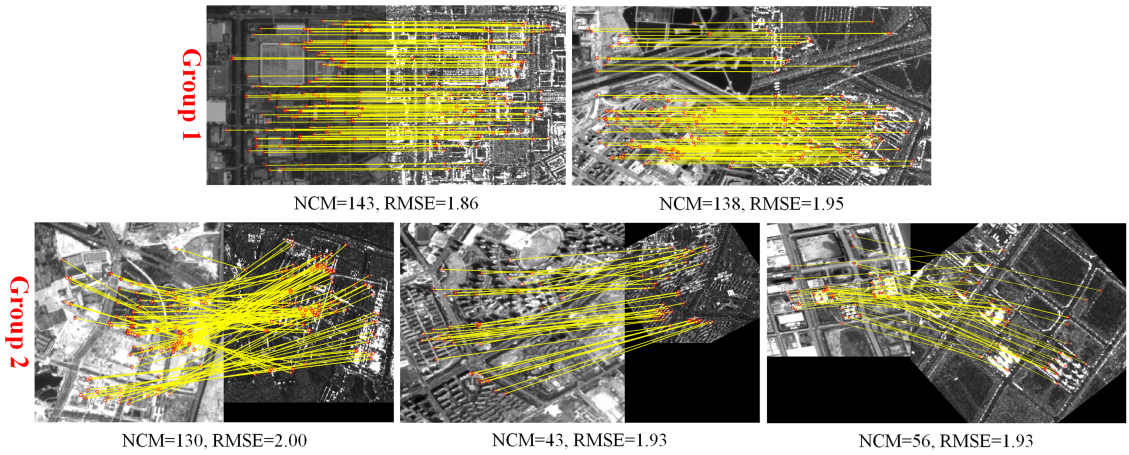

Fig. 14. Feature matching of the proposed method on the HROS dataset.

image in this pair is textureless. Fig. 16 shows the feature matching, evaluation, and registration results on the three image pairs. As seen, the proposed method successfully registers the three image pairs, which indicates that our method has a good adaptability for dealing with optical and SAR image registration.

B. *Effectiveness of MORTV*

In this study, MORTV is used to convert optical and SAR images into structural maps. Such a representation significantly filters out textures and noises, and decreases the modality differences between optical and SAR images. To verify the effectiveness of MORTV, four other edge-preserving filters, including the NDF, COF, RGF, and RTV, are utilized to construct image multiscale representation, instead of using MORTV. We then separately perform the MBST feature detection and the LMHOG description on them. For comparison, the GS is also included. A series of experiments are implemented on the MROS dataset and the detailed comparison results are given in Table VIII. Compared with these approaches, MORTV can bring a significant improvement in terms of four evaluation criteria. This is because that MORTV can well characterize significant structural features in images using a multi-orientation strategy

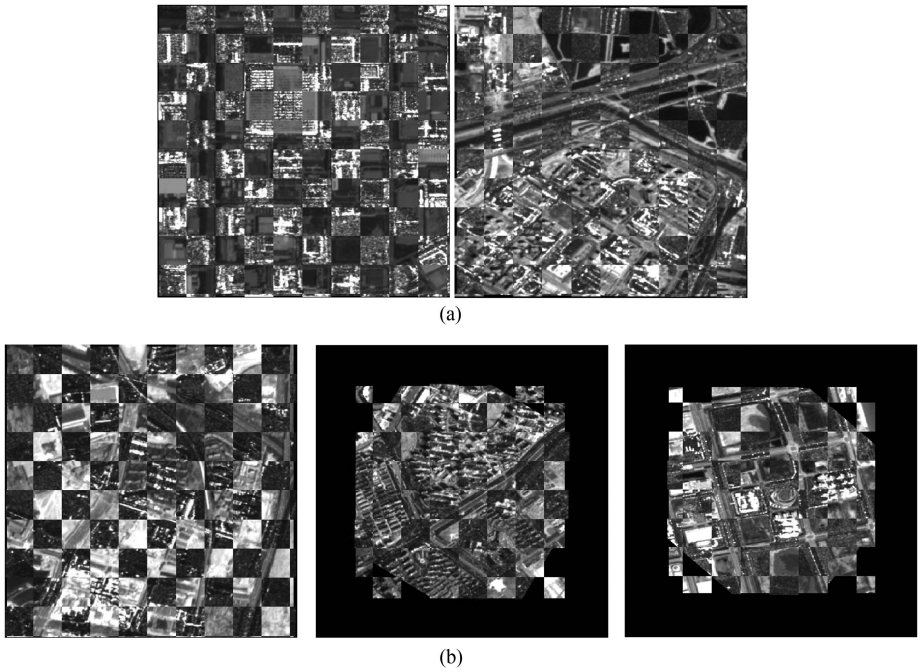

Fig. 15. Registration results of the proposed method on the MROS dataset. (a) Group 1. (b) Group 2.

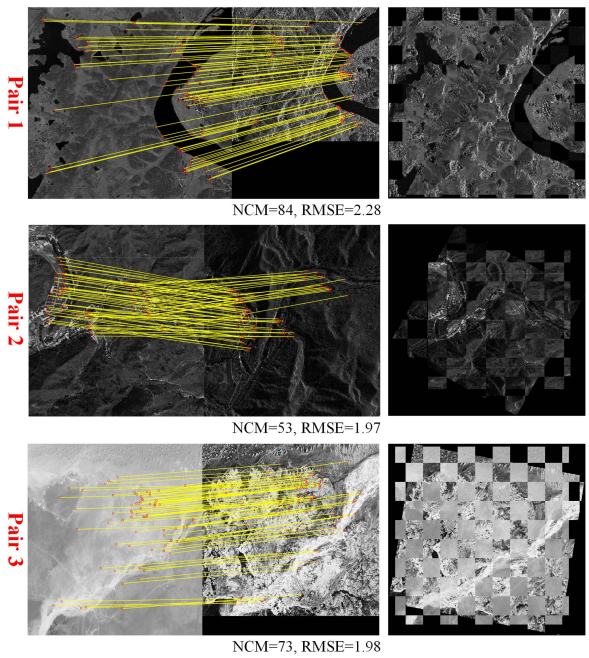

Fig. 16. Registration results of the proposed method on three complex scenes.

and further enhance the performance of feature detection and description.

C. *Effectiveness of LMHOG*

In our work, LMHOG is used to encode structural information with a layerwise multiscale feature region. To verify the effectiveness of LMHOG, we extract only feature region at a fixed scale while keeping other stages unchanged. Similarly, the quantitative results of five methods on the MROS dataset are presented in Table IX. Here, LMHOG-S1, LMHOG-S2, and LMHOG-S3 are the modified LMHOG that are, respectively,

TABLE VIII
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE MROS DATASET

Method	Criterion			
	NCM	CMR	SR	RMSE
GS	39.5	11.3	70.2	6.12
NDF	72.6	18.2	78.5	4.22
COF	94.3	25.8	89.4	2.61
RGF	112.8	31.5	98.2	2.03
RTV	96.3	29.2	97.5	2.18
MORTV	116.2	33.3	99.8	1.97

Each value is the average result.

TABLE IX
QUANTITATIVE RESULTS OF DIFFERENCE METHODS ON THE MROS DATASET

Method	Dimension	Criterion			
		NCM	CMR	SR	RMSE
LMHOG-S1	240	79.3	23.6	91.8	2.44
LMHOG-S2	240	71.2	19.4	84.6	2.93
LMHOG-S3	240	66.5	16.8	79.3	3.32
LMHOG-C	720	109.4	33.0	99.2	1.95
LMHOG	240	116.2	33.3	99.8	1.97

Each value is the average result.

constructed on small-scale, medium-scale, and large-scale structural maps, while LMHOG-C refers to the modified LMHOG that is generated by directly combining all three scale structural maps. From the results, it can be seen that LMHOG has a significant performance improvement over the best performance when using a single feature region. Moreover, LMHOG-C obtains comparable results to LMHOG, but it has a feature dimension of 720, which greatly increases the computational complexity. Such an ablation experiment demonstrates the effectiveness of our LMHOG.

V. Conclusion

In this article, we present a novel MORTV structural representation for optical and SAR image registration. The MORTV model is first designed to produce the structural maps with different orientations. Then, based on multiscale MORTV representation, the MBST detector is introduced to extract the shared features. For each extracted point, the LMHOG descriptor is designed to encode the structural features at different scales in a multilayer manner, which improves the discriminative ability of the descriptor without increasing its dimension. Extensive experimental results on two large-scale datasets demonstrate that the proposed method brings a significant improvement in registration performance and produces competitive performance in matching optical and SAR images with intensity changes, geometric differences, and image noises. Although the proposed method provides robust registration performance, it cannot handle the images that contain less structural information. Our future work will explore an image enhancement method to improve image quality.

Acknowledgment

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

Reference

[1] Y. Chen and L. Bruzzone, "Self-supervised SAR-optical data fusion of Sentinel-1/-2 images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5406011.

[2] Y. Ye, M. Wang, L. Zhou, G. Lei, J. Fan, and Y. Qin, "Adjacent-level feature cross-fusion with 3-D CNN for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Aug. 2023, Art. no. 5618214.

[3] Y. Ye et al., "An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 516.

[4] 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.

[5] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, 2021.

[6] 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.

[7] 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," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammetry Remote Sens.*, vol. 188, pp. 331–350, 2022.

[8] 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.

[9] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, and Y. Ye, "R_2FD_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, Apr. 2023, Art. no. 5606115.

[10] 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.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vis. Conf.*, 1988, pp. 147–151.

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

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

[14] 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. 8, pp. 3078–3090, Jun. 2018.

[15] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychol. Res.*, vol. 64, no. 2, pp. 136–148, Dec. 2000.

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

[17] J. Fan, Y. Ye, G. Liu, J. Li, and Y. Li, "Phase congruency order-based local structural feature for SAR and optical image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, May 2022, Art. no. 4507105.

[18] 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, Dec. 2020.

[19] D. Xiang, Y. Xu, J. Cheng, Y. Xie, and D. Guan, "Progressive keypoint detection with dense Siamese network for SAR image registration," *IEEE Trans. Aerosp. Electron. Syst.*, early access, Apr. 11, 2023, doi: 10.1109/TAES.2023.3266415.

[20] J. Huang, F. Yang, and L. Chai, "Robust registration of multimodal remote sensing images with spectrum congruency," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5103–5114, May 2023.

[21] X. Xiong, G. Jin, Q. Xu, H. Zhang, L. Wang, and K. Wu, "Robust registration algorithm for optical and SAR images based on adjacent self-similarity feature," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5233117.

[22] 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.

[23] 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.

[24] Y. Zhang et al., "Histogram of the orientation of the weighted phase descriptor for multimodal remote sensing image matching," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammetry Remote Sens.*, vol. 196, pp. 1–15, 2023.

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

[26] 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.

[27] J. Fan, Y. Ye, J. Li, G. Liu, and Y. Li, "A novel multiscale adaptive binning phase congruency feature for SAR and optical image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5235216.

[28] Q. Yu, D. Ni, Y. Jiang, 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," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammetry Remote Sens.*, vol. 171, pp. 1–17, 2021.

[29] 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.

[30] 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.

[31] Y. Hong, C. Leng, X. Zhang, J. Peng, L. Jiao, and A. Basu, "Max-index based local self-similarity descriptor for robust multi-modal image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Mar. 2022, Art. no. 4505805.

[32] A. Sedaghat and N. Mohammadi, "Illumination-robust remote sensing image matching based on oriented self-similarity," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammetry Remote Sens.*, vol. 153, pp. 21–35, Jul. 2019.
[33] J. Fan, Q. Xiong, Y. Ye, and J. Li, "Combining phase congruency and self-similarity features for multimodal remote sensing image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jan. 2023, Art. no. 4001105.
[34] X. Xiong, G. Jin, Q. Xu, and H. Zhang, "Self-similarity features for multimodal remote sensing image matching," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12440–12454, Nov. 2021.
[35] A. Sedaghat and H. Ebadi, "Distinctive order based self-similarity 1descriptor for multi-sensor remote sensing image matching," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammerty Remote Sens.*, vol. 108, pp. 62–71, Oct. 2015.
[36] H. Zhang et al., "Optical and SAR image dense registration using a robust deep optical flow framework," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1269–1294, Jan. 2023.
[37] Y. Deng and J. Ma, "ReDFeat: Recoupling detection and description for multimodal feature learning," *IEEE Trans. Image Process.*, vol. 32, pp. 591–602, Dec. 2023.
[38] 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.
[39] D. Quan et al., "Self-distillation feature learning network for optical and SAR image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 4706718.
[40] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5622215.
[41] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and SAR image registration based on feature decoupling network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5235913.
[42] J. Li, W. Xu, P. Shi, Y. Zhang, and Q. Hu, "LNIFT: Locally normalized image for rotation invariant multimodal feature matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5621314.
[43] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *Assoc. Comput. Machinery Trans. Graph.*, vol. 31, no. 6, pp. 1–10, 2012.
[44] J. Shi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 593–600.
[45] P. Duan, X. Kang, S. Li, and P. Ghamisi, "Noise-robust hyperspectral image classification via multi-scale total variation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1948–1962, Jun. 2019.
[46] M. Wei, Y. Feng, and H. Chen, "Selective guidance normal filter for geometric texture removal," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 12, pp. 4470–4483, Dec. 2021.
[47] Y. Ye, W. Liu, L. Zhou, T. Peng, and Q. Xu, "An unsupervised SAR and optical image fusion network based on structure-texture decomposition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 4028305.
[48] Y.-F. Yu, C.-X. Ren, D.-Q. Dai, and K.-K. Huang, "Kernel embedding multiorientation local pattern for image representation," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1124–1135, Apr. 2018.
[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simomcelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 31, no. 4, pp. 600–612, Apr. 2004.
[50] Y. Yao, Y. Zhang, Y. Wan, X. Liu, X. Yan, and J. Li, "Multi-modal remote sensing image matching considering co-occurrence filter," *IEEE Trans. Image Process.*, vol. 31, pp. 2854–2597, Mar. 2022.
[51] Q. Yu, S. Zhou, Y. Jiang, P. Wu, and Y. Xu, "High-performance SAR image matching using improved SIFT framework based on rolling guidance filter and ROEWA-powered feature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 920–933, Mar. 2019.
[52] 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.
[53] 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, Jul. 2022, Art. no. 5626714.
[54] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and SAR images via improved phase congruency model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5847–5861, Sep. 2020.

Jianwei Fan received the B.S. degree in electronical information science and technology from the Henan University of Science and Technology, Luoyang, China, in 2011, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2017.

He is currently a Lecturer with the School of Computer and Information Technology, Xinyang Normal University, Xinyang, China. His research interests include remote sensing image processing, image registration, and feature extraction.

Qing Xiong received the B.S. degree in data science and big data technology in 2018 from Xinyang Normal University, Xinyang, China, where she is currently working toward the M.S. degree in computer science and technology.

Her research focuses on remote sensing image matching.

Jian Li received the B.E. degree in computer science and technology from Hainan University, Haikou, China, in 2014, and the Ph.D. degree in information and communication engineering from Sun Yat-sen University, Guangzhou, China, in 2019.

He is currently a Lecturer with the School of Computer and Information Technology, Xinyang Normal University, Xinyang, China. His current research interests include neural networks, numerical computation, and robotics.

Yuanxin Ye received the B.S. degree in remote sensing science and technology from Southwest Jiaotong University, Chengdu, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

He is currently a Research Fellow with the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu. His research interests include remote sensing image processing, image registration, change detection, and object detection.

Dr. Ye was the recipient of ISPRS Prizes for Best Papers by Young Authors of the 23rd International Society for Photogrammetry and Remote Sensing Congress (Prague, 2016) and Best Youth Oral Paper Award of ISPRS Geospatial week 2017 (Wuhan, 2017), respectively.