LNIFT: Locally Normalized Image for Rotation Invariant Multimodal Feature Matching

Article *in* IEEE Transactions on Geoscience and Remote Sensing · January 2022

DOI: 10.1109/TGRS.2022.3165940

5 authors, including:

Jiayuan Li
Wuhan University
53 PUBLICATIONS **1,189** CITATIONS
SEE PROFILE

Pengcheng Shi
Wuhan University
12 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Yongjun Zhang
Wuhan University
93 PUBLICATIONS **1,156** CITATIONS
SEE PROFILE

LNIFT: Locally Normalized Image for Rotation Invariant Multimodal Feature Matching

Jiayuan Li, Wangyi Xu, Pengcheng Shi, Yongjun Zhang, and Qingwu Hu

***Abstract*—Severe nonlinear radiation distortion (NRD) is the bottleneck problem of multimodal image matching. Although many efforts have been made in the past few years, such as the radiation-variation insensitive feature transform (RIFT) and histogram of orientated phase congruency (HOPC), almost all these methods are based on frequency domain information that suffers from high computational overhead and memory footprint. In this paper, we propose a simple but very effective multimodal feature matching algorithm in spatial domain, called locally normalized image feature transform (LNIFT). We first propose a local normalization filter to convert original images into normalized images for feature detection and description, which largely reduce the NRD between multimodal images. We demonstrate that normalized matching pairs have much larger correlation coefficient than the original ones. We then detect oriented FAST and rotated brief (ORB) keypoints on the normalized images and use an adaptive non-maximal suppression (ANMS) strategy to improve the distribution of keypoints. We also describe keypoints on the normalized images based on a histogram of oriented gradient (HOG) like descriptor. Our LNIFT achieves rotation invariance the same as ORB without any additional computational overhead. Thus, LNIFT can be performed in near real-time on images with 1024×1024 pixels (only costs 0.32s with 2500 keypoints). Four multimodal image datasets with a total of 4000 matching pairs are used for comprehensive evaluations, including synthetic aperture radar (SAR)-optical, infrared-optical, and depth-optical datasets. Experimental results show that LNIFT is far superior than RIFT in terms of efficiency (0.49s vs 47.8s on a 1024×1024 image), success rate (99.9% vs 79.85%), and number of correct matches (309 vs 119). The source code and datasets will be publicly available in https://ljy-rs.github.io/web.**

***Index Terms*—Multimodal image matching, Feature matching, Local descriptor, SAR-optical, Infrared-optical, Depth-optical.**

I. Introduction

IMAGE matching refers to the process of detecting reliable correspondences from the same scene images collected at different times, by different sensors or from different perspectives. Related applications based on image matching include not only aerial triangulation in remote sensing, but also visual navigation in positioning and navigation, simultaneous localization and mapping in robotics, and target tracking in intelligent transportation, etc. The development of image matching can effectively promote the progress of these applications.

Image matching technology has been extensively studied in the past few decades. However, most of current methods (e.g., scale-invariant feature transform (SIFT) [1], speeded-up robust features (SURF) [2], shape context [3], etc.) are only suitable for same-source images with linear radiation distortions [4], [5]. For example, Moghimi et al. [6], [7] proposed a relative radiometric normalization technique for multitemporal and multisensor remote sensing images based on these matching methods. However, they stated that these methods are only enough for linear radiation distortions. They perform very poor on multimodal images. Multimodal images refer to the images acquired by sensors with different imaging mechanisms, which show large apparent differences on the same ground objects, and usually suffer from serious nonlinear radiation distortion (NRD), such as synthetic aperture radar (SAR)-optical, infrared-optical, depth-optical, map-optical, etc [8]. Multimodal image matching is a current research hotspot and also a challenging task of feature matching [5].

According to [8], [9], multimodal image matching methods can also be categorized into two groups: area-based methods and feature-based ones. Area-based matching methods, such as mutual information (MI) [10], [11] and histogram of orientated phase congruency (HOPC) [12], [13], are also known as the template matching approaches, which are very suitable for image pairs with only translation changes. However, they are difficult to be applied in cases with rotation, scale, and perspective changes. Compared with template matching, feature matching is more robust to geometric distortions and has a wider range of applications. It generally consists of three steps, i.e., feature detection, feature description, and keypoint matching. However, many current multimodal feature matching methods are based on frequency domain information that relies on Fourier transform. For example, the radiation-variation insensitive feature transform (RIFT) [8] uses Fourier transform to convert original images into log-Gabor sequences and phase congruency maps for feature detection and description. Although fast Fourier transform (FFT) reduces the complexity from $O(N^2)$ to $O(Nlog(N))$, it is still very slow for large scale problems (e.g., $Nlog(N) = 1.38 \times 10^7$ for an image with a size of 1000×1000). Therefore, current multimodal feature matching methods generally suffer from high computational overhead and memory footprint, which greatly limit their potentials in practical applications.

The difficulty of multimodal image matching mainly lies in the inconsistency of modalities between the reference image and the target image, which results in large NRD and lack of sufficiently similar image structures. If the images of two different modalities can be converted to the same intermediate modal, so that it contains the information shared between the two modalities, then, the multimodal image matching problem can be turned into a single-modal matching problem. At this time, traditional methods or conventional deep learning methods can also achieve good performance. Based on this

This work was supported by National Natural Science Foundation of China (No. 42030102 and 41901398).

basic idea, we develop a simple but very effective multimodal feature matching method in spatial domain, called locally normalized image feature transform (LNIFT). We propose a local normalization filter to convert multimodal images into an intermediate modal (normalized image). As demonstrated, gradients of the intermediate modal images have much larger normalized cross-correlation coefficient (NCC) [14] than the ones of original images. For efficiency, we use the integral image technique in the filter. Then, we perform an improved oriented fast and rotated brief (ORB) [15] detector and a histogram of oriented gradient (HOG) [16] like descriptor on the intermediate modal images to extract and describe features. The rotation angle of a keypoint is calculated the same as ORB without any additional computational overhead. We use four large-scale datasets with a total of 4000 matching pairs for comprehensive evaluations, including 2000 synthetic aperture radar (SAR)-optical pairs, 1000 infrared-optical pairs, and 1000 depth-optical pairs. Experimental results show that our LNIFT is much better than RIFT. LNIFT can run in near real-time on a 1024×1024 image with 5000 keypoints.

The contributions of this paper are as follows:

- A multimodal feature matching algorithm, called LNIFT, is developed, which is far superior than RIFT in terms of efficiency, success rate, and number of correct matches.
- A local normalization filter is proposed. With normalization, multimodal images become much more similar and have much higher MI and NCC scores. The integral image technique is also used for speeding up the filter.
- An improved ORB keypoint detector is presented, which uses an adaptive non-maximal suppression (ANMS) strategy to improve the distribution of keypoints.
- Different from current literatures which only use several or dozens multimodal image pairs for comparison, we collect four real large-scale datasets with a total of 4000 image pairs for evaluation.

II. Related Work

According to the central ideas, multimodal image matching methods can be divided into two categories: area-based algorithms and feature based ones.

A. Area-based Matching

The core idea of area-based matching is to compare the pixel similarity between the target image and the reference one by sliding window strategy, which finds the window pair with the highest similarity as the correct correspondence. This process is also known as template matching.

Mutual information (MI) methods: MI is usually used to describe the correlation measure between two data sets or events, which has been widely used in multi-source image registration in the fields of medicine and remote sensing [10], [17]–[20]. MI-based methods need to search for the best similarity in the entire search space, which is difficult to achieve the global optimum. Moreover, they suffer from high computational overhead.

Fourier transform methods: These methods generally describe images in frequency domain based on the Fourier transform [21]. However, if the spectral components between image pairs differ greatly, this type of methods will become very unreliable. Recently, Ye et al. [13] proposed the HOPC and the channel features of orientated gradients (CFOG) [4], which use the FFT for frequency domain feature representation. Xiang et al. [22] proposed an improved phase congruency model. These methods show good robustness to NRDs. However, they generally rely on the prior of geographic information. In addition, the FFT suffers from high computational overhead and memory footprint as abovementioned.

Learning-based methods: Using Deep learning (DL) technique for multimodal image registration is also a research hotspot. For instance, Siamese convolutional neural networks [23]–[27] show the potentials for multimodal template matching. Merkle et al. [26] used a SAR-optical matching Siamese network to improve the geo-localization accuracy of optical satellite images. Hughes et al. [28] used a hard-negative mining strategy to improve the performance of deep networks. However, as data-driven approaches, deep-learning based methods heavily depend on the variety of training datasets (multimodal image matching problem contains many different image modalities, such as SAR, optical, depth, night-time, infrared, map, LIDAR intensity, etc.) and require higher computing resources [29]. Moreover, these patch-matching methods are also sensitive to large geometric transforms (e.g., large rotations) [5], [30].

B. Feature-based Matching

Features generally refer to salient keypoints such as corners in the image. Feature-based matching contains three major stages: feature detection (e.g., features from accelerated segment test (FAST) [31], Harris [32], Difference-of-Gaussian (DoG) [1], etc.), feature description (e.g., SIFT [1], SURF [2], RIFT [8], etc.), and keypoint matching (e.g., random sample consensus (RANSAC) family [9], [33] and robust estimators [34]–[38]).

Feature matching for same-source images has been well-studied in the past several decades. Many well-known methods have been proposed, such as SIFT [1], SURF [2], affine SIFT (ASIFT) [39], ORB [15], learned invariant feature transform (LIFT) [40], and SuperPoint [41], etc. However, these methods are generally not suitable for multimodal feature matching since they do not consider the NRDs [5]. Several methods are also developed for heterogeneous feature matching, such as the local self-similarity descriptor (LSS) [5], [42], [43], partial intensity invariant feature descriptor (PIIFD) [44], position-scale-orientation SIFT (PSO-SIFT) [45], optical-SAR SIFT (OS-SIFT) [46] and its variant [47], DL-based matching [48], etc. These methods achieve good results on specific image types. However, they are generally not suitable for other types of images, which largely limits their applications.

Recently, Li et al. [8] proposed the RIFT to cope with severe NRDs, which is suitable for different types of images, such as infrared-optical, SAR-optical, depth-optical, and map-optical, etc. Cui et al. [49] extended the RIFT to achieve scale invariance. However, these methods are frequency domain methods, which suffer from high computational complexity. Moreover, they are sensitive to severe speckle noise.

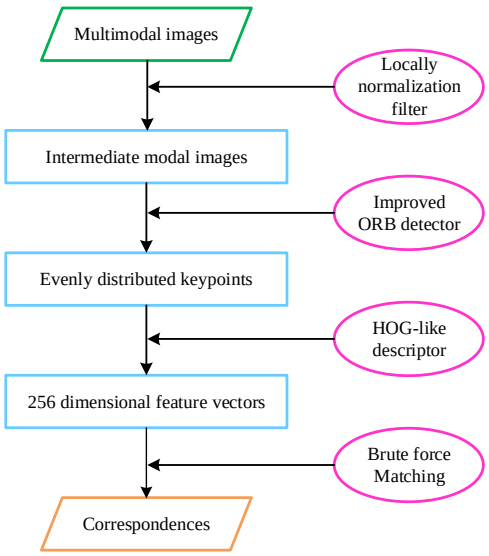
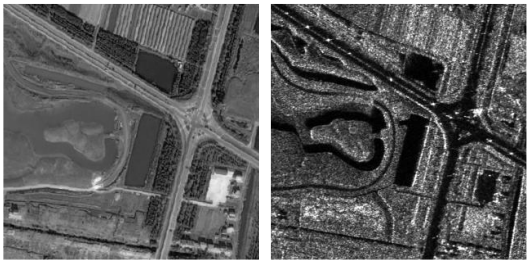
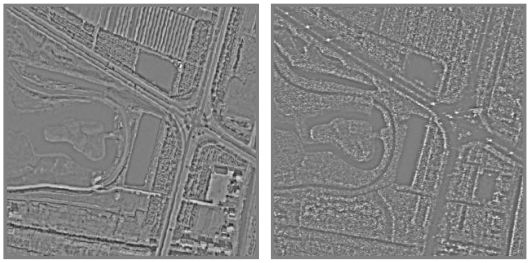


Fig. 1. The framework of our LNIFT. A given multimodal image pair is first converted into the same intermediate modal based on a locally normalization filter. Then, improved ORB detector and HOG-like descriptor are applied on the normalized images to detect and describe keypoints.



(a) Original optical and SAR images. (MI=0.165, NCC=0.018)



(b) Locally normalized images. (MI=0.315, NCC=0.407)

Fig. 2. An example pair of multimodal images. (a) A pair of SAR-optical images. (b) The corresponding normalized images of (a). After normalization, the images become much more similar, i.e., the MI and NCC become much higher.

Actually, multimodal feature matching is far from a solved problem. In this paper, we aim to develop a practical multimodal feature matching method with the properties of high efficiency, excellent robustness to NRDs, and rotation invariance. First, to achieve high efficiency, we detect and describe features in spatial domain, while most current methods are based on frequency information. We use a simple filter with only $O(1)$ per-pixel complexity to extract structure information. Thus, our LNIFT can run in near real-time on a 1024×1024 image with 5000 keypoints. Second, to achieve high robustness to NRDs, we introduce the idea that if two different modalities can be converted to the same intermediate modal, the multimodal image matching problem can be turned into a same-source matching problem. We propose a locally normalization filter which makes images with different modalities become more similar. Therefore, our LNIFT get far superior performance than RIFT on multimodal images.

III. Methodology

Figure 1 displays the framework of our LNIFT. Given a pair of multimodal images, we first transform them into the same intermediate modal images based on a locally normalization filter. Then, an improved ORB detector is applied on the normalized images to extract evenly distributed features. These keypoints are encoded into feature vectors by a HOG-like descriptor. Finally, we use a brute force searching strategy to establish correspondences. The details of each stage are described below.

A. Locally Normalization Filter

As aforementioned, if we can convert multimodal images into the same intermediate modal which contains the common information between the two modalities, then, the multimodal image matching problem becomes a conventional image matching problem. Hence, typical hand-crafted and learning-based feature matching methods can be adapted. In this section, we propose a locally normalization filter to achieve this goal, whose mathematical definition is,

$$I_N(x,y) = I(x,y) - \frac{1}{|W(x,y,s)|} \sum_{W(x,y,s)} I(x,y), \qquad (1)$$

where I_N and I represent the normalized image and its original image, respectively; (x,y) represents a 2D image coordinate; $W(x,y,s)$ is a local window centered at (x,y) with size $(2*s+1) \times (2*s+1)$.

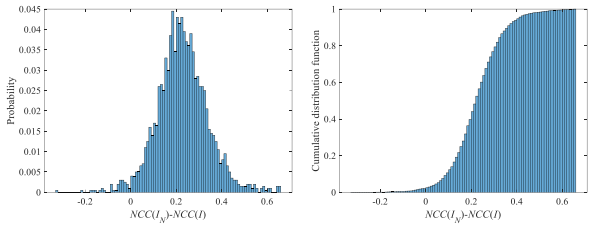
Essentially, our locally normalization filter is equivalent to the image I minus its average filtering result. The average filter is a smooth filter, which removes the details from the image I. Thus, our filter only preserves the detail component; namely, the normalized image contains most of the structure information in I, which is very important for multimodal image matching. This is the theoretical reason why our method works well on multimodal images. Actually, there are many filters can be used to separate the detail component. Using these filters instead of the average filter in Eq. (1) can also get good results. The reason that we choose the average filter is because of its efficiency. After calculating the integral image I_Σ of I,

$$I_\Sigma(x,y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i,j), \qquad (2)$$

our locally normalization filter can be implemented in $O(1)$ time. Eq. (1) becomes,

$$\begin{aligned} I_N(x,y) = I(x,y) + I_\Sigma(x+s,y+s) + I_\Sigma(x-s,y-s) \\ -I_\Sigma(x+s,y-s) - I_\Sigma(x-s,y+s) \end{aligned}. \qquad (3)$$

Theoretically, locally normalization filter can largely reduce the NRD between multimodal images. Then, the overlapping areas of normalized image pairs should be more similar than



(a) Probability distribution (b) Cumulative distribution

Fig. 3. The probability distribution of the increment $\Delta_{NCC} = NCC(I_N) - NCC(I)$.

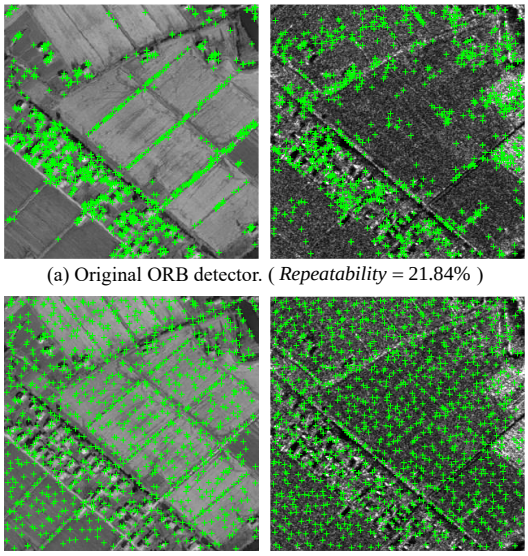
TABLE I
IMAGE SIMILARITY COMPARISON

Metric	Original images	Normalized images
MI↑	0.145	0.328
NCC↑	0.102	0.330

the ones of original multimodal images. Thus, the normalized images should have higher similarity scores than multimodal images. To verify this conclusion, we conduct an experiment on a SAR-optical dataset with 2000 image pairs. Each image pair is pre-registered. For each pair I^1 and I^2, we calculate their normalized images I_N^1 and I_N^2. Since the gradients will be used for feature description, we also compute the gradients of these images, obtaining gI^1, gI^2, gI_N^1 and gI_N^2. Then, we use the MI metric to measure the similarity between I^1 and I^2 (I_N^1 and I_N^2), and adopt the NCC metric to measure the similarity between gI^1 and gI^2 (gI_N^1 and gI_N^2). Figure 2 provides an example. As shown, the NRD between SAR and optical images is serious. For instance, the man-made objects (e.g., roads and cement grounds) are black in the SAR image while white or gray in the optical image. Fortunately, they become much similar after normalization, i.e., the MI and NCC are largely increased. We analysis the probability distribution of the increment $\Delta_{NCC} = NCC(I_N) - NCC(I)$ in Figure 3. The normalized images have higher NCC scores than the original ones in 98% of matching pairs. The average MI and NCC results are reported in Table I. As shown, the MI of normalized images is two times of the one of original images, and the NCC is three times of original images (Both MI and NCC are the larger the better).

B. Improved ORB Detector

We propose an improved ORB detector to extract keypoints on the normalized images, since ORB detector has a very high efficiency compared with SIFT [1], and SURF [2] detectors. ORB improves the FAST [31] detector to achieve rotation invariance. It uses FAST-9 (circular radius is 9) to detect features and sorts these keypoints based on a Harris cornerness measure. It then picks the best N keypoints with the highest Harris scores. For remote sensing images, the scale differences can be easily eliminated by the prior of ground sampling distance (GSD). Although the ORB detector achieves scale invariance based on a scale pyramid, it largely increases the computational complexity. Therefore, we remove the scale pyramid construction from the classical ORB detector and only detect features in the first level (original resolution).



(a) Original ORB detector. (*Repeatability* = 21.84%)

(b) Our ORB with ANMS detector. (*Repeatability* = 29.04%)

Fig. 4. Comparison of the distributions of keypoints between the original ORB and our improved ORB.

TABLE II
EVALUATING THE EFFECT OF LOCAL NORMALIZATION FILTER ON FEATURE DETECTION AND DESCRIPTION. (LNIFT-DETECT: IROB_{ori} + IHOG_{lni}; LNIFT-DESCRIBE: IROB_{lni} + IHOG_{ori}; LNIFT: IROB_{lni} + IHOG_{lni}.)

Metric	LNIFT-detect	LNIFT-describe	LNIFT
n ↑	173	434	577
CMR (%) ↑	3.45	8.68	11.53

Similar to the ORB, we also use the intensity centroid to calculate a dominant angle for each keypoint to achieve rotation invariance. Differently, we normalize the angle into $[0°, 180°)$ instead of $[0°, 360°)$. As mentioned in [46], [50], intensities often have a reversal in multimodal images such as SAR-optical and infrared-optical images. Thus, we regard the orientations θ and $\theta + 180°$ as the same. The calculation of intensity centroid is based on the image moment m_{pq}, whose definition is as follows:

$$m_{pq} = \sum_{(x,y)\in\Omega(k_i)} x^p y^q I_N(x,y), \tag{4}$$

where $\Omega(k_i)$ is a local image patch centered at keypoint k_i. Then, the intensity centroid of patch $\Omega(k_i)$ is,

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right). \tag{5}$$

The vector $\overrightarrow{k_i C}$ is the dominant orientation of patch $\Omega(k_i)$ and the angle is $\tilde{\theta}_{k_i} = \arctan 2\left(\frac{m_{01}}{m_{10}}\right)$, where $\arctan 2$ represents the quadrant-aware version of $\arctan$. Considering the intensity reversal problem, the orientation θ_{k_i} of keypoint k_i is,

$$\theta_{k_i} = \begin{cases} \tilde{\theta}_{k_i} - \pi & \tilde{\theta}_{k_i} > \pi \\ \tilde{\theta}_{k_i} & otherwise \end{cases} \tag{6}$$

One problem of the modified ORB detector is the cluster phenomenon of keypoints (See Figure 4(a)), which decreases

TABLE III
DETAILED SETTINGS OF EACH COMPARED ALGORITHMS (MNOF REPRESENTS MAXIMUM NUMBER OF FEATURES)

Method	Parameters	Implementations
SIFT	MNoF: 5000; patch size: 24*scale; contrast threshold: 0.001; edge threshold: 31; descriptor type: regular grid; descriptor size: 128.	C++ code: https://www.vlfeat.org/overview/sift.html
PSO-SIFT	MNoF: 5000; patch size: 24*scale; contrast threshold: 0.001; edge threshold: 31; descriptor type: log polar gird; descriptor size: 136.	MATLAB code: https://github.com/ZeLianWen/Image-Registration
OS-SIFT	MNoF: 5000; patch size: 24*scale; Harris function threshold: 0.001; scale ratio: $\sqrt[3]{2}$; descriptor type: log polar gird; descriptor size: 136.	MATLAB code: https://sites.google.com/view/yumingxiang/
RIFT	MNoF: 5000; patch size: 96; FAST contrast threshold: 0.001; FAST quality threshold: 0.001; descriptor type: regular grid; descriptor size: 216.	MATLAB code: https://ljy-rs.github.io/web/
Our LNIFT	MNoF: 5000; patch size: 96; ORB edge threshold: 5; filter window size: $s = 3$; descriptor type: regular grid; descriptor size: 256.	C++ code: https://ljy-rs.github.io/web/

the accuracy of structure from motion or registration and increases redundant information. To obtain a homogeneous distribution of keypoints, we introduce an ANMS strategy to suppress clustered features. Suppose the output of ANMS is M well-distributed keypoints, we first lower the parameter of FAST to get a large size $N > 2M$ of keypoints $K = \{k_i\}_1^N$ sorted according to Harris scores. For each $k_i \in K$, we search its neighbors K_{k_i} in a range of $\sqrt{\frac{w \cdot h}{4M}}$ using KD-tree and remove them from K, i.e., $K = K \backslash K_{k_i}$, where w and h are the width and height of an image. Finally, we select the first M keypoints in the remaining K as the output. Figure 4(b) shows an example result of our ORB with ANMS detector (denoted as IORB). As can be seen, the keypoints are homogeneously distributed cover the whole image and the detector repeatability is even better than the original ORB.

To show the effect of our locally normalization filter on feature detection, we conduct an experiment on a depth-optical dataset with 1000 image pairs. We perform our IORB on each original image (denoted as IROB$_{ori}$) and normalized image (denoted as IROB$_{lni}$) to obtain 5000 keypoints, respectively. Then, our HOG-like descriptor is applied on the normalized image (denoted as IHOG$_{lni}$) to describe features for both the two detection strategies. The combination of IROB$_{lni}$ + IHOG$_{lni}$ is our LNIFT algorithm, and the combination of IROB$_{ori}$ + IHOG$_{lni}$ is denoted as LNIFT-detect. Feature vectors are matched via brute force searching without nearest neighbor distance ratio (NNDR) test. We use the number of correct matches n and the correct matching rate (CMR) as evaluation metrics, where a correct match represents a correspondence whose residual under ground truth transformation is smaller than 3 pixels, and the CMR is defined as the ratio between n and the total number of matches. The results are summarized in Table II. As reported, with local normalization for feature detection, the number of correct matches and CMR have increased by more than 3 times.

C. HOG-like Descriptor

For each feature k_i, we first rotate its local image patch according to the dominant angle to achieve rotation invariance. Then, we present a simple HOG-like descriptor to describe this local patch. As analysed in Section III-A, locally normalized images can largely reduce the NRDs compared with original

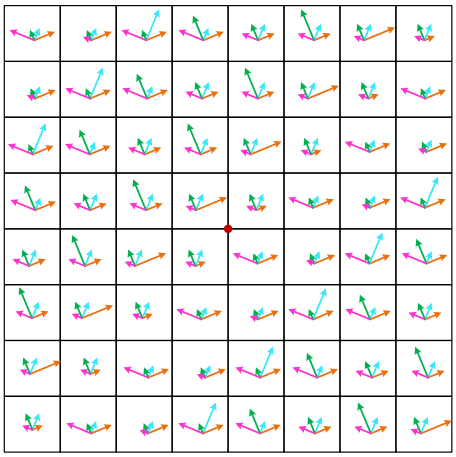


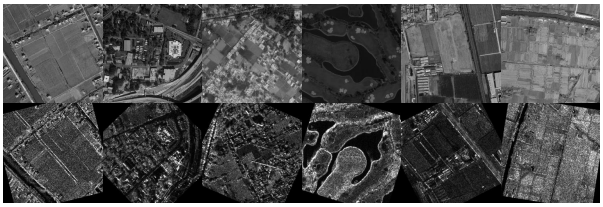
Fig. 5. Our HOG-like descriptor, where the red dot represents the keypoint k_i. The local image patch is divided into 8×8 grids. Each grid is encoded into a 4-bin histogram of oriented gradients, where the orientations are normalized into $[0°, 180°)$.

multimodal images. Therefore, we calculate gradients on the normalized images instead of original ones for description. Different from traditional gradient-based descriptors, we normalize the gradient orientations into $[0°, 180°)$, since gradient orientations often have a reversal in multimodal images as aforementioned. Suppose the size of local patch is $J \times J$ pixels, we divide the patch into 8×8 grids since the patch size is much larger than SIFT. In SIFT-like descriptors, they build a 8-bin histogram for orientations (belong to $[0°, 360°)$) in each grid. In our case, the orientations are belong to $[0°, 180°)$. Thus, we only compute a distribution histogram with 4 bins for each grid (See Figure 5). Finally, a total of 64 histograms are concatenated together to obtain the feature vector of k_i, which is then normalized into a unit vector to gain invariance to illumination changes. Hence, the length of our HOG-like descriptor is $8 \times 8 \times 4 = 256$. There are three differences between our HOG-like descriptor and the HOG: first, our descriptor is built on the locally normalized images; second, the grids in our descriptor have no overlaps between each other, which is similar to the SIFT; third, we only use a 4-bin histogram to encode $[0°, 180°)$ orientations.

To show the effect of our locally normalization filter on feature description, we also conduct an experiment that is

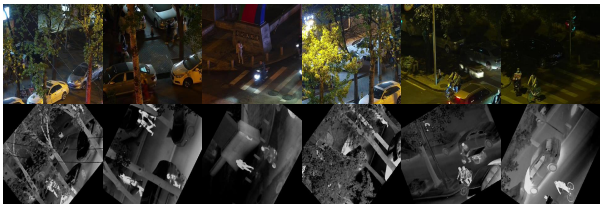
similar to the one in Section III-B. First, our IROB$_{lni}$ detector is applied to obtain 5000 keypoints. Then, we perform our IHOG descriptor on each original image (denoted as IHOG$_{ori}$) and normalized image (denoted as IHOG$_{lni}$) for description. The combination of IROB$_{lni}$ + IHOG$_{ori}$ is denoted as LNIFT-describe. The number of correct matches n and CMR results are also summarized in Table II. As reported, with local normalization for feature description, the number of correct matches and CMR have increased by more than 30%.

(a) Dataset 1: SAR-optical without rotations (1000 pairs)



(b) Dataset 2: SAR-optical with rotations (1000 pairs)

(c) Dataset 3: depth-optical without rotations (1000 pairs)



(d) Dataset 4: infrared-optical with rotations (1000 pairs)

Fig. 6. Sample data of our four multimodal image datasets.

IV. Experiments

In this section, we comprehensively evaluate our LNIFT on four real multimodal datasets. Different from traditional methods that use several or dozens of image pairs for test, we use 4000 image pairs for comparisons. The proposed LNIFT is compared with four baseline or state-of-the-art methods, including SIFT [1], PSO-SIFT [45], OS-SIFT [46], and RIFT [8]. For fair comparisons, we use the official implementations of each method provided by the authors (apart from the SIFT, which is implemented by the VLFeat toolbox), and fine-tune their parameters to achieve the best total performance on 4000 pairs. For example, we set the Harris function threshold of OS-SIFT to be very small (0.001) to extract as many feature points as possible. We also fix the scales of SIFT, PSO-SIFT, and OS-SIFT keypoints to be 4, so that their local patches for description have a size 96×96 pixels, which is the same as our LNIFT. The same matching strategy is applied for all compared methods, i.e., feature vectors are matched via brute force searching without NNDR test. The parameter settings and implementation details of each compared algorithm are summarized in Table III. We use the success rate γ, number of correct matches n, and root mean square error (RMSE) r as the quantitative evaluation metrics. The same as in the above, a correct match represents a correspondence whose residual under ground truth transformation is smaller than 3 pixels. The success rate γ is the ratio between the number of correctly matched image pairs and the total number of image pairs. If the number of correct matches n of an image pair is not smaller than 10, this image pair is regarded as correctly matched, since too small n will make robust estimation technique fail. The formula of RMSE is,

TABLE IV
The Details of Parameter Settings

Parameter	Variable	Fixed parameters
J	$J=[64, 80, 96, 112, 128]$	$n_{bin}=4$, $n_{grid}=8$, $s=2$
n_{bin}	$n_{bin}=[2, 4, 6, 8, 10]$	$J=96$, $n_{grid}=8$, $s=2$
n_{grid}	$n_{grid}=[4, 6, 8, 10, 12]$	$J=96$, $n_{bin}=4$, $s=2$
s	$s=[1, 2, 3, 4, 5]$	$J=96$, $n_{bin}=4$, $n_{grid}=8$

$$r=\sqrt{\frac{1}{C}\sum_{i=1}^{C}\left(\boldsymbol{y}_i-T(\boldsymbol{x}_i)\right)^2}, \tag{7}$$

where C is the number of correct matches, $T(\cdot)$ is the ground truth transformation, $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_1^C$ are correct matches. If one image pair is not correctly matched (i.e., $n<10$), its RMSE is set to be 20 pixels. All the experiments are performed on a PC with i9-10850K CPU at 3.6GHz and 64 GB of RAM.

A. Datasets

We collect four real multimodal datasets for evaluations, including SAR-optical, depth-optical, and infrared-optical datasets. The detailed information of each dataset is described below. Figure 6 shows several sample image pairs of our datasets.

Dataset 1: This is a SAR-optical dataset without rotations. We randomly pick 1000 image pairs with size of 256×256 pixels from the QXS-SAROPT [51] dataset as the Dataset 1. The QXS-SAROPT dataset consists of 20000 registered SAR-optical image pairs with a high resolution of 1m, where the SAR images suffered from severe speckle noise are obtained from the GaoFen-3 SAR satellite and the optical correspondences are collected from Google Earth. These pairs spread across multiple scenes, including Shanghai, Qingdao, and San Diego. Since each pair is registered, its ground truth transformation is an identity matrix.

Dataset 2: This is a SAR-optical dataset with rotations. We randomly pick 1000 image pairs with size of 512×512 pixels from the OS-DATASET [22] to construct our Dataset 2. Then, each selected image pair is rotated around its center according to a randomly generated angle within $[0°, 90°)$. Based on the coordinates of image center and rotation angle, we can

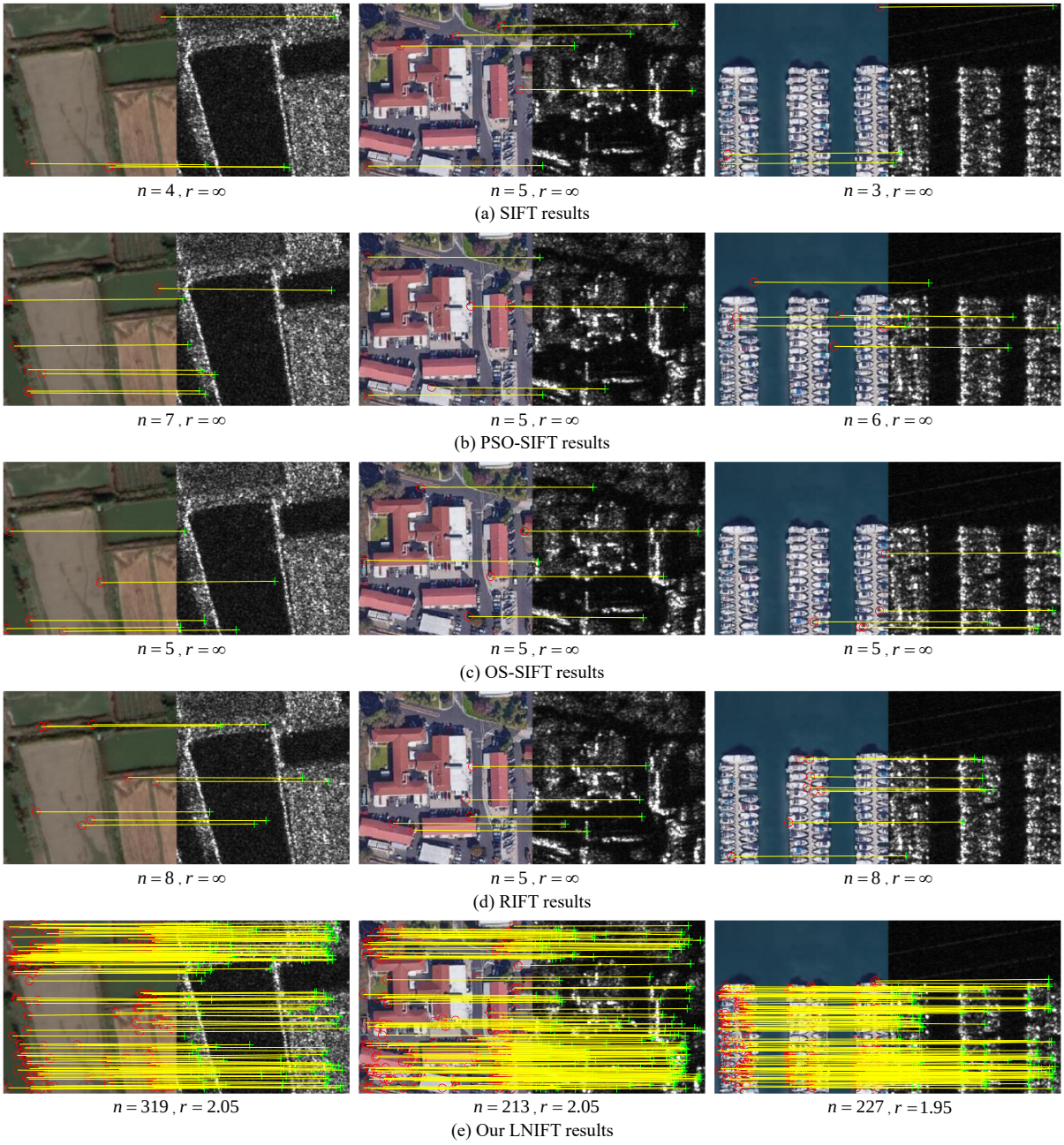


Fig. 7. Qualitative comparison results on the Dataset 1. Red circles and green crosshairs are keypoints of the reference and target images, respectively; yellow lines represent correct matches. A method with RMSE $r = \infty$ indicates that it fails to match this image pair. For better visualization, no more than 200 matches are displayed.

TABLE V
THE RESULTS OF PARAMETER J

Metric	J, $n_{bin} = 4$, $n_{grid} = 8$, $s = 2$				
	64	80	96	112	128
RMSE r (pixels)↓	2.77	2.15	2.13	2.13	2.16
success rate γ (%)↑	96.0	99.6	99.7	99.7	99.6
correct match number n ↑	145	187	208	215	213

establish a ground truth transformation for this image pair. The OS-DATASET consists of 2673 registered SAR-optical image pairs with a high resolution of 1m. The types of SAR and optical images are the same as the QXS-SAROPT dataset. These images cover over more than 15 cities around the world, including Beijing, Wuhan, Rennes, Omaha, Dwarka, etc.

Dataset 3: This is a depth-optical dataset without rotations. We randomly pick 500 registered indoor image pairs and 500 registered outdoor image pairs from the DIML/CVL RGB-D [52] dataset. The indoor images with size of 512×288 pixels are captured by the Microsoft Kinect v2 RGBD camera. These images cover over various scenes (such as dormitory, offices, exhibition center, rooms, etc) of South Korea. The outdoor optical images with size of 640×384 pixels are captured by the ZED camera. Their corresponding depth images are generated

TABLE VI
THE RESULTS OF PARAMETER n_{bin}

Metric	n_{bin}, $J = 96$, $n_{grid} = 8$, $s = 2$				
	2	4	6	8	10
RMSE r (pixels)↓	2.16	2.13	2.09	2.11	2.13
success rate γ (%)↑	99.5	99.7	99.9	99.8	99.7
correct match number n ↑	139	208	212	211	209

TABLE VII
THE RESULTS OF PARAMETER n_{grid}

Metric	n_{grid}, $J = 96$, $n_{bin} = 4$, $s = 2$				
	4	6	8	10	12
RMSE r (pixels)↓	2.34	2.15	2.13	2.12	2.10
success rate γ (%)↑	98.6	99.6	99.7	99.7	99.8
correct match number n ↑	117	174	208	228	236

TABLE VIII
THE RESULTS OF PARAMETER s

Metric	s, $J = 96$, $n_{bin} = 4$, $n_{grid} = 8$				
	1	2	3	4	5
RMSE r (pixels)↓	2.13	2.13	2.13	2.15	2.15
success rate γ (%)↑	99.7	99.7	99.7	99.6	99.6
correct match number n ↑	200	208	213	216	218

by a stereo matching algorithm. These outdoor images cover over the scenes of roads, parks, buildings, etc. The ground truth transformation of an image pair is also an identity matrix.

Dataset 4: This is a thermal infrared-optical dataset with rotations. We randomly pick 1000 registered image pairs with size of 320×256 pixels from the LLVIP [53] dataset to construct our Dataset 4. Similar to the Dataset 2, each selected image pair is rotated according to a random angle. The ground truth transformation is also established. The LLVIP dataset consists of 16836 image pairs, which are captured by a binocular camera system (a visible camera and a thermal infrared camera) under low-light conditions from 26 different locations.

B. Parameter Study

There are mainly four parameters affect the performance of our LNIFT, i.e., J, n_{bin}, n_{grid}, and s. Parameter J is the patch size for feature description, n_{bin} represents the number of bins of the oriented histogram, n_{grid} is the number of subgrids inside a local image patch, and s is the half of the window size of our locally normalization filter. Small J and n_{bin} may not contain sufficient information for keypoint description, which decreases the distinctiveness of features. Parameter n_{grid} is mainly used to encode spatial information. If n_{grid} is small, the descriptor may be very sensitive to local geometric distortions. However, large values of J, n_{bin}, and n_{grid} greatly increase the computational complexity. Smaller s preserves less image details. However, a larger s will highlight more noise. Thus, it is very important to learn suitable parameters to balance the accuracy and the efficiency. Here, we perform four independent experiments to study these four parameters on the Dataset 2. In each experiment, only one parameter is treated as a variable and the others are fixed. Table IV provides the detailed experimental settings. The results are summarized in Table V∼Table VIII.

According to the results, we can see that: If the values of J, n_{bin}, and n_{grid} are too small, the numbers of correct matches n are low. For example, when $n_{grid} = 4$, its number of correct matches $n = 117$, which is only the half of the one of $n_{grid} = 10$. However, it is not that the larger of J or n_{bin} is, the better the performance. When J or n_{bin} reaches a certain value (e.g., $J = 96$, $n_{bin} = 4$), the performance only changes slightly as their values increase. Large values of J or n_{bin} will significantly increase the computational complexity. For parameter n_{grid}, the larger values mean better performance. However, if $n_{grid} = 10$, $n_{bin} = 4$, the dimension of our descriptor is 400, which is too large and will largely decrease the efficiency. Parameter s has litter influence on the performance. After considering both the performance and the efficiency, we set $J = 96$, $n_{bin} = 4$, $n_{grid} = 8$, and $s = 3$ in the following experiments.

C. Qualitative Evaluations

Three image pairs from each multimodal dataset are selected for qualitative comparisons, as displayed in Figure 7 ∼ Figure 10. Among them, Figure 7 suffers from severe speckle noise. Figure 8 has better image quality than Figure 7, but it suffers from rotation changes. Figure 9 suffers from very large differences in imaging mechanisms. Strictly speaking, a depth map is not really an image. Figure 10 contains huge variations in lighting (RGB night-time and infrared images) and suffers from rotation changes. Thus, it is quite challenging to match these image pairs.

From the results, we can see that SIFT fails on all the 12 image pairs. As analyzed in [8], descriptors based on typical gradients are very sensitive to NRDs and not suitable for multimodal image matching. This is the fundamental reason why SIFT performs so bad. PSO-SIFT improves the gradient calculation of SIFT. However, it only matches successfully on three image pairs, whose success rate is only 25%. Moreover, although the matching is successful, the number of correct matches n is very small. OS-SIFT performs slightly better than PSO-SIFT, getting a success rate of 33.3%. OS-SIFT detects features based on a multiscale Harris function, which usually extracts fewer keypoints than others such as the ORB. Compared with above methods, RIFT achieves much better results. The success rate of RIFT is 50% on these 12 pairs. It uses the phase congruency map for feature detection and maximum index map for feature description. These two maps are specially designed to decrease the effect of NRDs. Hence, RIFT is very suitable for multimodal image matching problem. However, RIFT is sensitive to severe speckle noise. For example, it totally fails on the Dataset 1. The reason may be that the severe speckle noise causes inaccurate edge structure information in the phase congruency map. In contrast, our proposed LNIFT achieves the best results. It performs very well on all image pairs, i.e., a 100% success rate. Moreover, our number of correct matches n is much higher than others.

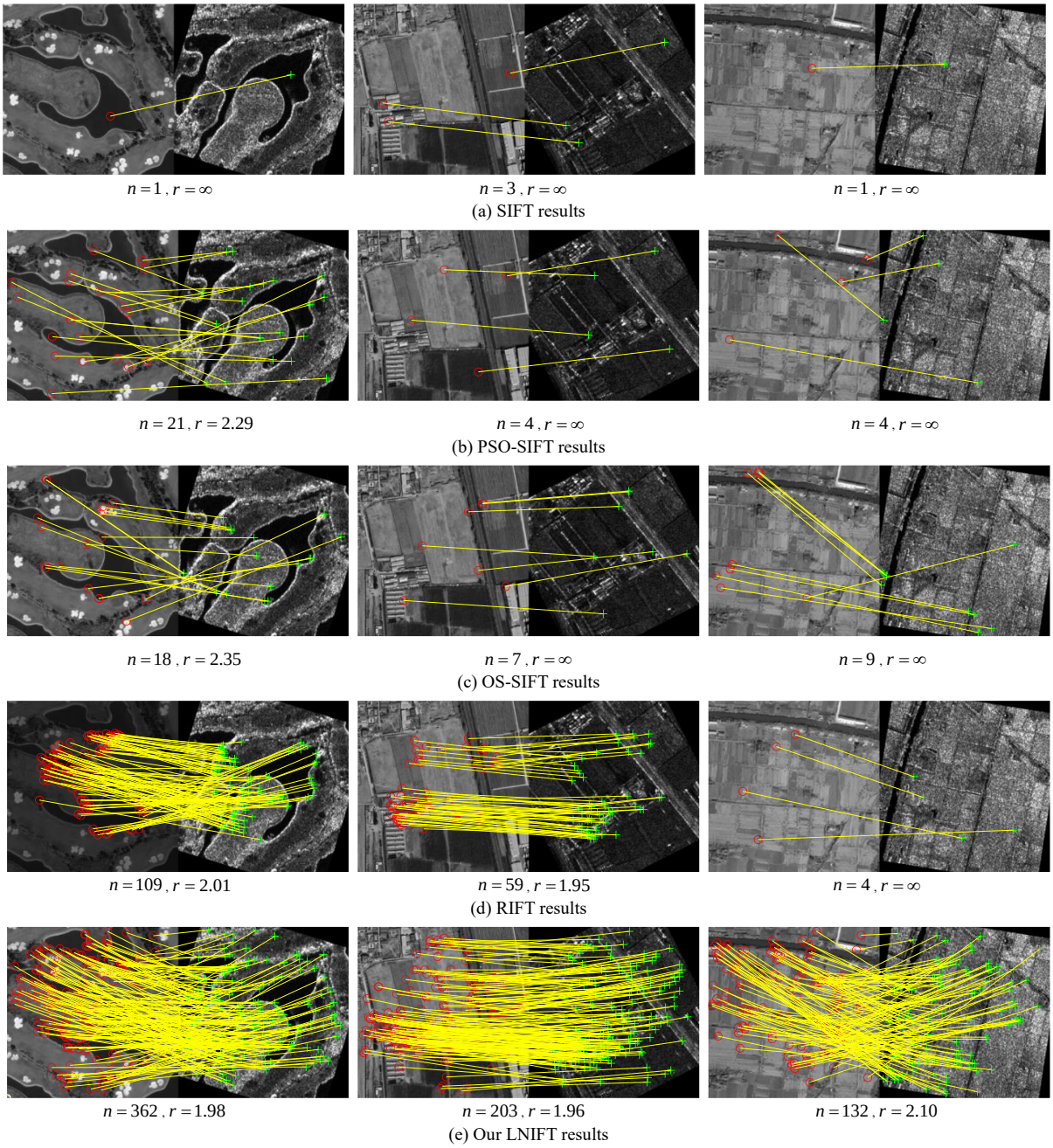


Fig. 8. Qualitative comparison results on the Dataset 2. Red circles and green crosshairs are keypoints of the reference and target images, respectively; yellow lines represent correct matches. A method with RMSE $r = \infty$ indicates that it fails to match this image pair. For better visualization, no more than 200 matches are displayed.

For example, our LNIFT gets four times as many correct matches as RIFT. The reason may be twofold: (1) we convert the two different modalities into the same intermediate modal based on a locally normalization filter, which largely decreases the NRDs. (2) we have carefully designed the feature detector and descriptor so that they are suitable for multimodal images.

D. Quantitative Evaluations

The quantitative results on each dataset are summarized in Table IX, where the RMSE is the lower the better while the success rate and n are the higher the better. As reported, the highest success rate of SIFT is still lower than 7%. Namely, it almost fails on all the image pairs, which is expected since SIFT is not designed for multimodal image matching. PSO-SIFT gets better results than SIFT. It achieves a success rate of 47.8% on the Dataset 3. However, its success rate is no better than 30% on the other three datasets. The average number of correct matches n is only 8. This capability is far from sufficient for practical applications. OS-SIFT performs slightly better than PSO-SIFT on the Dataset 2, since it is designed for SAR-optical matching. Even so, OS-SIFT only gets a success rate of 41.3% on Dataset 2, which is very bad compared

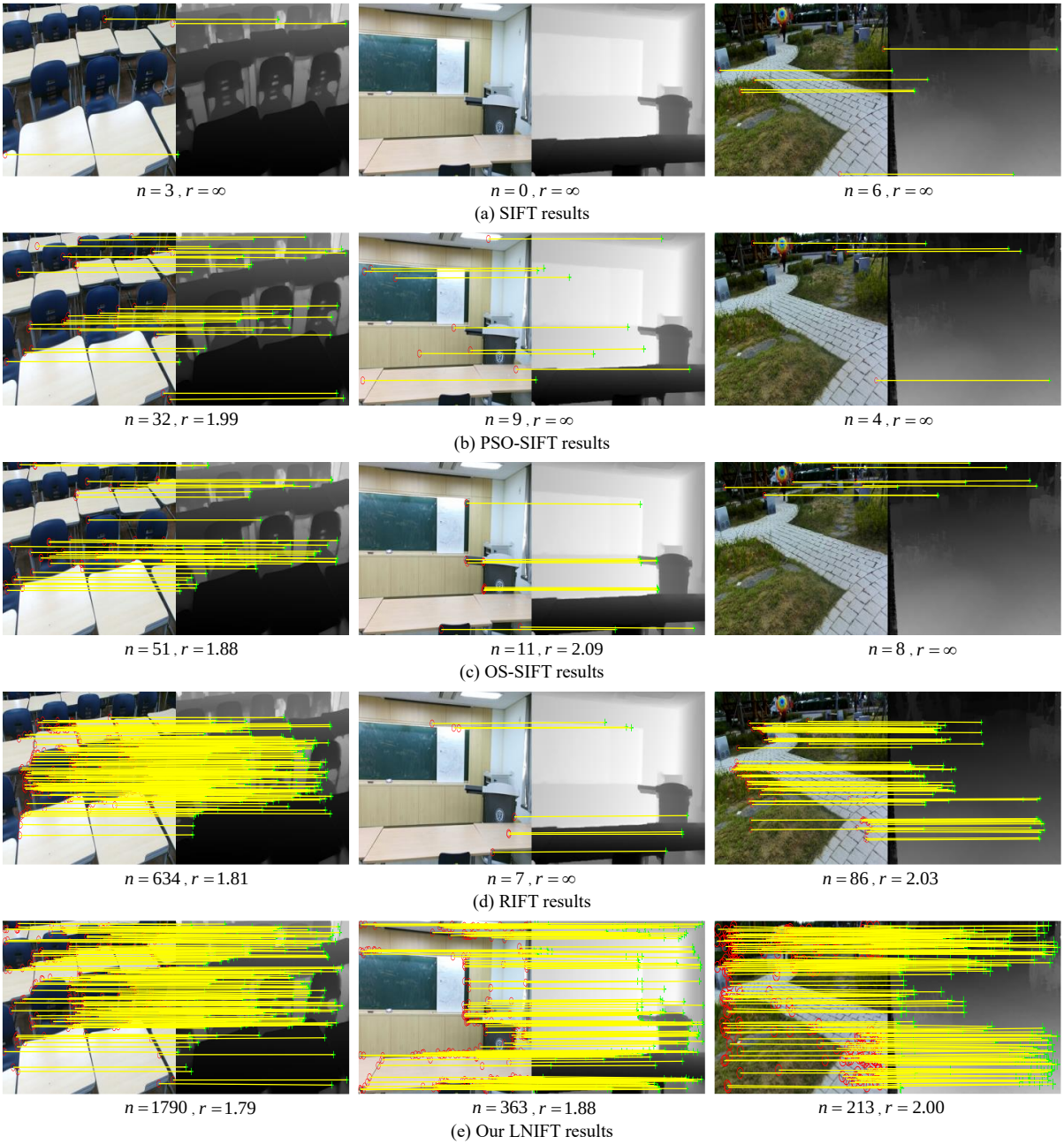


Fig. 9. Qualitative comparison results on the Dataset 3. Red circles and green crosshairs are keypoints of the reference and target images, respectively; yellow lines represent correct matches. A method with RMSE $r = \infty$ indicates that it fails to match this image pair. For better visualization, no more than 200 matches are displayed.

with RIFT and our LNIFT. Dataset 1 is extremely challenging due to severe speckle noise. SIFT, PSO-SIFT, and OS-SIFT are almost totally failed. The second best performance in success rate achieved by RIFT is still lower than 35%. RIFT ranks second among all the compared methods. It obtains very impressing results on Dataset 2 $\sim$ Dataset 4, whose success rate is higher than 90%. However, the edge information in phase congruency of RIFT is sensitive to speckle noise, this is the reason why RIFT performs so bad on Dataset 1. Compared with the above methods, our LNIFT achieves the best results on all the datasets, i.e., it gets the smallest RMSE and the highest success rate and n. Our success rate is close to 100% on all datasets.

The average success rates of SIFT, PSO-SIFT, OS-SIFT, RIFT, and our LNIFT on all the four datasets are 3.0%, 26.0%, 28.6%, 79.85%, and 99.9%, respectively. Our LNIFT improves by 20 percents compared with RIFT, and more than 70 percents compared with OS-SIFT. The average numbers of correct matches n of SIFT, PSO-SIFT, OS-SIFT, RIFT, and our LNIFT are 3, 8, 10, 119, 309, respectively. Our LNIFT gets almost three times as many correct matches as RIFT, and 30 times as many correct matches as OS-SIFT and PSO-SIFT. The average RMSE of our LNIFT is 2.05 pixels, which is enough for many remote sensing applications.

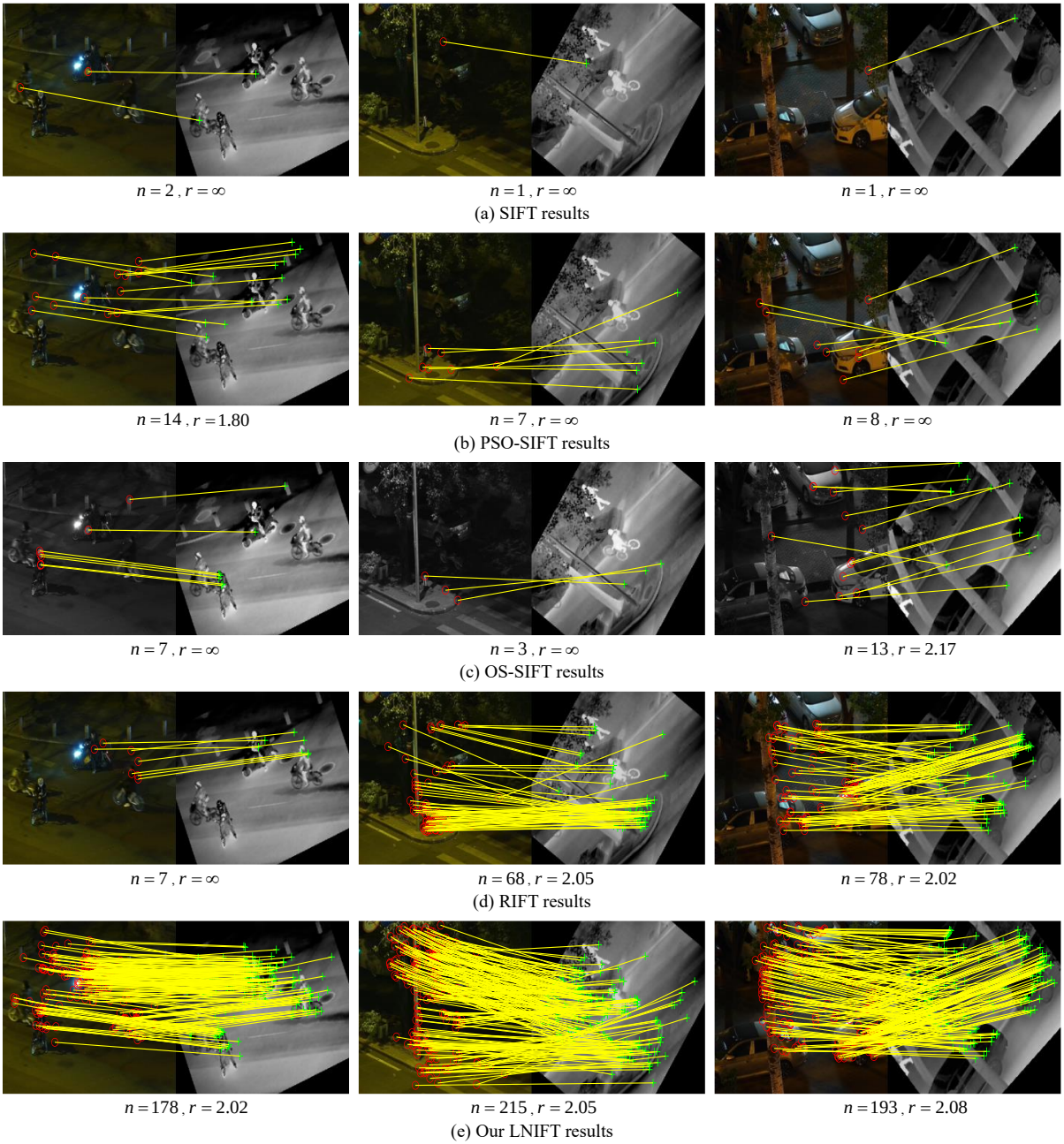


Fig. 10. Qualitative comparison results on the Dataset 4. Red circles and green crosshairs are keypoints of the reference and target images, respectively; yellow lines represent correct matches. A method with RMSE $r = \infty$ indicates that it fails to match this image pair. For better visualization, no more than 200 matches are displayed.

Further, template-based methods can be applied to refine the results of LNIFT, which is a typical strategy called coarse-to-fine in image registration.

E. Ablation study

We conduct an ablation experiment on the Dataset 3 (depth-optical) to demonstrate the necessity of each improvement in the proposed LNIFT algorithm. Table X summarizes the quantitative results with different component configurations, where ORB_{ori}+HOG_{ori} is the baseline, IORB represents our improved ORB detector, IHOG is our HOG-like descriptor, subscript ori represents operations performed on the original image, and subscript lni represents operations performed on our locally normalized image (LNI). As can be seen, removing any improvement of our LNIFT decreases the performance, which proves that our designs could boost the proposed method. Although our ORB detector mainly focuses on improving the distribution of keypoints, it still increases the success rate γ by 10 percents. With our improvement on the HOG, the number of correct matches has doubled. The major contribution, i.e., LNI, increases the performance by a large of margin.

TABLE IX
QUANTITATIVE COMPARISON RESULTS ON THE FOUR DATASETS. EACH REPORTED VALUE IS THE AVERAGE OF ALL MATCHING PAIRS.

Data	Metric	Method				
		SIFT	PSO-SIFT	OS-SIFT	RIFT	Our LNIFT
Dataset 1 SAR-optical	RMSE r (pixels)↓	20	19.69	19.05	14.05	**1.99**
	success rate γ (%)↑	0.0	1.7	5.3	33.0	**100**
	correct match number n ↑	2	3	4	11	**219**
Dataset 2 SAR-optical	RMSE r (pixels)↓	19.73	15.59	12.60	3.36	**2.13**
	success rate γ (%)↑	1.5	24.5	41.3	92.7	**99.7**
	correct match number n ↑	2	7	10	88	**213**
Dataset 3 depth-optical	RMSE r (pixels)↓	18.84	11.30	11.01	2.39	**1.98**
	success rate γ (%)↑	6.3	47.8	49.5	97.2	**99.8**
	correct match number n ↑	4	15	19	262	**577**
Dataset 4 infrared-optical	RMSE r (pixels)↓	19.24	14.58	16.71	2.63	**2.08**
	success rate γ (%)↑	4.2	30.0	18.3	96.5	**100**
	correct match number n ↑	3	8	7	114	**226**
Average accuracy	RMSE r (pixels)↓	19.45	15.29	14.84	5.61	**2.05**
	success rate γ (%)↑	3.0	26.0	28.6	79.85	**99.9**
	correct match number n ↑	3	8	10	119	**309**

TABLE X
THE RESULTS OF ABLATION STUDY (IORB: OUR IMPROVED ORB; IHOG: OUR HOG-LIKE DESCRIPTOR; SUBSCRIPT ori: OPERATIONS PERFORMED ON THE ORIGINAL IMAGE; SUBSCRIPT lni: OPERATIONS PERFORMED ON OUR LOCALLY NORMALIZED IMAGE; SR: SUCCESS RATE; CMN: CORRECT MATCH NUMBER)

Method	RMSE r (pixels)↓	SR γ (%)↑	CMN n ↑
ORB_{ori}+HOG_{ori}	6.52	74.7	41
IORB_{ori}+HOG_{ori}	4.77	84.5	55
IORB_{ori}+IHOG_{ori}	4.02	88.3	98
IORB_{lni}+IHOG_{lni} (LNIFT)	1.98	99.8	577

TABLE XI
RUNNING TIME ANALYSIS

Method	Image size (pixel)			
	256×256	512×512	768×768	1024×1024
SIFT	0.24	0.86	2.10	3.89
PSO-SIFT	0.32	1.11	2.81	4.77
OS-SIFT	0.51	3.77	12.71	27.16
RIFT	3.30	21.22	33.94	47.80
LNIFT_{5000}	0.39	0.41	0.46	0.49
LNIFT_{2500}	**0.22**	**0.25**	**0.29**	**0.32**

F. Running time analysis

We perform an experiment to compare the running time of each method on the Dataset 3. Specifically, we resize the images of Dataset 3 to 256×256, 512×512, 768×768, and 1024×1024 pixels, which generate four datasets with different resolutions. The results are summarized in Table XI, where the LNIFT_{5000} and LNIFT_{2500} represent that the numbers of extracted keypoints of our LNIFT are 5000 and 2500, respectively.

As reported, when the image size is small, SIFT, PSO-SIFT, and OS-SIFT are comparable with our LNIFT_{5000} in running time. However, they become much slower than our method as the image size increases. For example, LNIFT_{5000} is about 10 times faster than SIFT and PSO-SIFT, and 50+ times faster than OS-SIFT on a 1024×1024 image. The reason may be twofold: First, SIFT, PSO-SIFT, and OS-SIFT can only detect a small number of keypoints (much smaller than 5000) on small-size images. Then, their feature description costs less time. Second, the calculations of these methods cost more than $O(N)$ time. RIFT is the slowest among these methods, since it is a frequency domain method. Our LNIFT is the fastest. LNIFT_{5000} and LNIFT_{2500} are about 100 times and 150 times faster than RIFT on a 1024×1024 image, respectively.

In our LNIFT, after the integral image (costs $O(N)$) is calculated, the remaining operations involved in our locally normalization filter are $O(1)$. The primary running time of the feature detection and description of LNIFT depends on the number of keypoints, but not the image size. This can also be reflected in Table XI. As can be seen, when the image size increases from 256×256 to 512×512, the running time only slightly increases. In contrast, when the number of keypoints increases from 2500 to 5000, the running time almost increases by 2 times. Therefore, if a task does not have high requirements on the number of correct matches, we can decrease the keypoint number to achieve real-time performance.

G. Limitations

Our LNIFT has mainly two drawbacks: First, LNIFT has no scale invariance since we discard the scale space of the ORB for efficiency. Currently, LNIFT is only invariant to translation and rotation variations. Hence, it is not suitable for image pairs that suffer from complex geometric distortions, such as large perspective changes and non-rigid distortions. The scale invariance can be achieved based on the GSD prior of satellite images or a Gaussian scale-space. Second, the CMR of LNIFT on SAR-optical images is not high (generally smaller than 10%) due to severe speckle noise. Thus, we need to extract

many keypoints to guarantee the number of correct matches. If we extract a small number of keypoints (e.g., 500), our LNIFT may get only a few correct matches. This problem can be alleviated by a SAR denoising preprocessing stage.

V. Conclusions

In this paper, we developed a multimodal feature matching method in spatial domain, called LNIFT, that is robust to severe NRDs. LNIFT achieves rotation invariance and can run in near real-time on a 1024×1024 image. It first tries to convert different modalities into the same one based on a locally normalization filter. Then, an improved ORB detector was adopted to extract evenly distributed keypoints and a HOG-like descriptor was designed for feature description. Both detection and description were performed on the locally normalized images. We evaluated LNIFT on four large-scale datasets with a total of 4000 multimodal image pairs. Comprehensively experiments demonstrated that LNIFT is far superior to current methods, i.e., our success rate is 20% higher than RIFT and 70% higher than OS-SIFT; our number of correct matches is about three times of the one of RIFT; and our running time is almost 100 times faster than RIFT on a moderate-size image. In the future, we will re-implement our LNIFT algorithm based on parallel computing and GPU to achieve real-time performance on large-size images, and achieve scale invariance by adding a scale space construction stage.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[4] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9059–9070, 2019.

[5] X. Xiong, G. Jin, Q. Xu, and H. Zhang, "Self-similarity features for multimodal remote sensing image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 12 440–12 454, 2021.

[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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4063–4073, 2021.

[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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2021.

[8] J. Li, Q. Hu, and M. Ai, "Rift: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Transactions on Image Processing*, vol. 29, pp. 3296–3310, 2020.

[9] ——, "Robust feature matching for geospatial images via an affine-invariant coordinate system," *The Photogrammetric Record*, vol. 32, no. 159, pp. 317–331, 2017.

[10] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[11] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[12] Y. Ye and L. Shen, "Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, p. 9, 2016.

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

[14] J.-C. Yoo and T. H. Han, "Fast normalized cross-correlation," *Circuits, systems and signal processing*, vol. 28, no. 6, pp. 819–843, 2009.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[17] C. Studholme, D. L. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern recognition*, vol. 32, no. 1, pp. 71–86, 1999.

[18] J. Liang, X. Liu, K. Huang, X. Li, D. Wang, and X. Wang, "Automatic registration of multisensor images using an integrated spatial and mutual information (smi) metric," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 1, pp. 603–615, 2013.

[19] H. Rivaz, Z. Karimaghaloo, V. S. Fonov, and D. L. Collins, "Nonrigid registration of ultrasound and mri using contextual conditioned mutual information," *IEEE transactions on medical imaging*, vol. 33, no. 3, pp. 708–725, 2013.

[20] J. Öfverstedt, J. Lindblad, and N. Sladoje, "Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment," *arXiv preprint arXiv:2106.14699*, 2021.

[21] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE transactions on image processing*, vol. 11, no. 3, pp. 188–200, 2002.

[22] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and sar images via improved phase congruency model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5847–5861, 2020.

[23] H. Zhang, W. Ni, W. Yan, D. Xiang, J. Wu, X. Yang, and H. Bian, "Registration of multimodal remote sensing image based on deep fully convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3028–3042, 2019.

[24] H. Zhang, L. Lei, W. Ni, T. Tang, J. Wu, D. Xiang, and G. Kuang, "Optical and sar image matching using pixelwise deep dense features," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[25] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for sar and optical images using multiscale convolutional gradient features," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[26] 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 Sensing*, vol. 9, no. 6, p. 586, 2017.

[27] Y. Fang, J. Hu, C. Du, Z. Liu, and L. Zhang, "Sar-optical image matching by integrating siamese u-net with fft correlation," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[28] L. H. Hughes, M. Schmitt, and X. X. Zhu, "Mining hard negative samples for sar-optical image matching using generative adversarial networks," *Remote Sensing*, vol. 10, no. 10, p. 1552, 2018.

[29] Y. Xiang, N. Jiao, F. Wang, and H. You, "A robust two-stage registration algorithm for large optical and sar images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[30] M. Uss, B. Vozel, V. Lukin, and K. Chehdi, "Efficient discrimination and localization of multimodal remote sensing images using cnn-based prediction of localization uncertainty," *Remote Sensing*, vol. 12, no. 4, p. 703, 2020.

[31] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[32] C. Harris, M. Stephens *et al.*, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[33] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[34] J. Li, Q. Hu, and M. Ai, "Point cloud registration based on one-point ransac and scale-annealing biweight estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9716–9729, 2021.

[35] J. Li, Q. Hu, M. Ai, and S. Wang, "A geometric estimation technique based on adaptive m-estimators: Algorithm and applications," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5613–5626, 2021.

[36] J. Li, Q. Hu, Y. Zhang, and M. Ai, "Robust symmetric iterative closest point," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 185, pp. 219–231, 2022.

[37] J. Li, Q. Hu, and M. Ai, "Robust geometric model estimation based on scaled welsch q-norm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5908–5921, 2020.

[38] ——, "Robust feature matching for remote sensing image registration based on $l\_\{q\}$-estimator," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1989–1993, 2016.

[39] J.-M. Morel and G. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM journal on imaging sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[40] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *European conference on computer vision*. Springer, 2016, pp. 467–483.

[41] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[42] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[43] A. Sedaghat and N. Mohammadi, "Illumination-robust remote sensing image matching based on oriented self-similarity," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 153, pp. 21–35, 2019.

[44] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine, "A partial intensity invariant feature descriptor for multimodal retinal image registration," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1707–1718, 2010.

[45] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, and L. Liu, "Remote sensing image registration with modified sift and enhanced feature matching," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 1, pp. 3–7, 2016.

[46] 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 Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3078–3090, 2018.

[47] 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," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 171, pp. 1–17, 2021.

[48] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of sar and optical imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 166–179, 2020.

[49] S. Cui, M. Xu, A. Ma, and Y. Zhong, "Modality-free feature detector and descriptor for multimodal remote sensing image registration," *Remote Sensing*, vol. 12, no. 18, p. 2937, 2020.

[50] D. Firmenichy, M. Brown, and S. Süsstrunk, "Multispectral interest points for rgb-nir image registration," in *2011 18th IEEE international conference on image processing*. IEEE, 2011, pp. 181–184.

[51] M. Huang, Y. Xu, L. Qian, W. Shi, Y. Zhang, W. Bao, N. Wang, X. Liu, and X. Xiang, "The qxs-saropt dataset for deep learning in sar-optical data fusion," *arXiv preprint arXiv:2103.08259*, 2021.

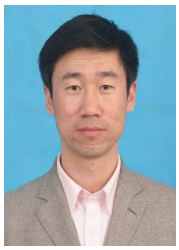
[52] J. Cho, D. Min, Y. Kim, and K. Sohn, "Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes," *arXiv preprint arXiv:2110.11590*, 2021.

[53] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3496–3504.

Jiayuan Li received the B.Eng., M.Eng., and Ph.D. degrees from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. He is currently an Associate Researcher with Wuhan University. He has authored more than 30 peer-reviewed articles in international journals. His research is mainly focused on SLAM, image matching, and point cloud registration. He was awarded the Best Youth Author Award by ISPRS in 2021 and the Talbert Abrams Award by ASPRS in 2018.

Wangyi Xu received the B.Eng. degree from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. Currently, he is pursuing the M.S. degree with Wuhan University. His research is mainly focused on multimodal image matching.

Pengcheng Shi received the B.S. degree in remote sensing science and technology from Liaoning Technical University, China, in 2018, and the M.S. degree in surveying and mapping engineering from Tongji University, China, in 2021. Currently, he is pursuing the Ph.D. degree with the School of Computer Science, Wuhan University. His research interests include simultaneous localization and mapping (SLAM) and point cloud registration.



Yongjun Zhang received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively. He is currently the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 180 research articles and one book. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, object information extraction and modeling with artificial intelligence, integration of LiDAR point clouds and images, and 3D city model reconstruction. He is the Co-Editor-in-Chief of The Photogrammetric Record.

Qingwu Hu received the B.Eng. and M.Eng. degrees in photogrammetry and remote sensing from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2007. He has authored more than 60 peerreviewed articles in international journals. His research interests include methods, techniques and applications of remote sensing, GIS and GPS integration, and photogrammetry.