remote sensing

MDPI

*Article*

# Robust Multimodal Remote Sensing Image Registration Based on Local Statistical Frequency Information

**Xiangzeng Liu *, Jiepeng Xue, Xueling Xu, Zixiang Lu, Ruyi Liu, Bocheng Zhao, Yunan Li and Qiguang Miao**

School of Computer Science and Technology, Xidian University, Xi'an 710071, China;
jpxue@stu.xidian.edu.cn (J.X.); xlxu_1@stu.xidian.edu.cn (X.X.); zxlu@xidian.edu.cn (Z.L.);
ruyiliu@xidian.edu.cn (R.L.); zhaobocheng@xidian.edu.cn (B.Z.); yunanli@xidian.edu.cn (Y.L.);
qgmiao@xidian.edu.cn (Q.M.)
* Correspondence: xzliu@xidian.edu.cn

**Abstract:** Multimodal remote sensing image registration is a prerequisite for comprehensive application of remote sensing image data. However, inconsistent imaging environment and conditions often lead to obvious geometric deformations and significant contrast differences between multimodal remote sensing images, which makes the common feature extraction extremely difficult, resulting in their registration still being a challenging task. To address this issue, a robust local statistics-based registration framework is proposed, and the constructed descriptors are invariant to contrast changes and geometric transformations induced by imaging conditions. Firstly, maximum phase congruency of local frequency information is performed by optimizing the control parameters. Then, salient feature points are located according to the phase congruency response map. Subsequently, the geometric and contrast invariant descriptors are constructed based on a joint local frequency information map that combines Log-Gabor filter responses over multiple scales and orientations. Finally, image matching is achieved by finding the corresponding descriptors; image registration is further completed by calculating the transformation between the corresponding feature points. The proposed registration framework was evaluated on four different multimodal image datasets with varying degrees of contrast differences and geometric deformations. Experimental results demonstrated that our method outperformed several state-of-the-art methods in terms of robustness and precision, confirming its effectiveness.

**Keywords:** remote sensing; multimodal image registration; phase congruency; Log-Gabor filter

check for updates

**Citation:** Liu, X.; Xue, J.; Xu, X.; Lu, Z.; Liu, R.; Zhao, B.; Li, Y.; Miao, Q. Robust Multimodal Remote Sensing Image Registration Based on Local Statistical Frequency Information. *Remote Sens.* **2022**, *14*, 1051. https://doi.org/10.3390/rs14041051

Academic Editors: Riccardo Roncella and Mattia Previtali

Received: 24 January 2022
Accepted: 18 February 2022
Published: 21 February 2022

## 1. Introduction

Nowadays, advanced remote sensing technology can realize omni-directional and multi-granularity perception of the same target scene under different imaging environment and conditions [1]. In the meanwhile, rich multimodal remote sensing images can be acquired with different sensors, or in different time periods [2,3]. A prerequisite for comprehensive utilization of multimodal image information is accurate image registration, which has been widely used in many computer vision tasks such as image mosaic, image fusion, change detection, vision-based navigation, scene matching guidance [4]. Therefore, the accuracy of image registration is critical to the application effect of the above top-level tasks.

Multimodal remote sensing image registration is the process of aligning remote sensing images of the same scene taken by different sensors, at different times, or/and from distinct viewpoints [5]. Although many related methods have been proposed [6–12], it is still very challenging work due to the contrast inconsistency caused by differences in imaging environment (different sensors, weather, and time periods) and the large geometric deformations induced by imaging conditions (different platform attitudes and positions) (Figure 1). Recent mainstream methods have been applied successfully in the situation where the geometric changes are small [6–9] or can be greatly alleviated according to the capture information [10,13]. However, automatic multimodal remote sensing image registration has not been solved effectively in complicated environments with large geometric changes and significant contrast differences. The challenges in the accurate registration of multimodal remote sensing images are specifically analyzed as follows:

1. Differences in imaging mechanisms or different weather capture conditions cause non-linear radiation changes between images, which leads to significant contrast differences, rendering traditional feature representation methods based on grayscale or gradient less effective or even invalid.
2. Large geometric deformations occur between images acquired from different azimuths (viewpoints) or different platforms (airborne camera, space camera), which makes it extremely difficult to extract invariant features.
3. Images acquired at different times or by different sensors contain some structural changes, resulting in poor consistency of the feature representation for the same target, making it difficult to achieve accurate registration.

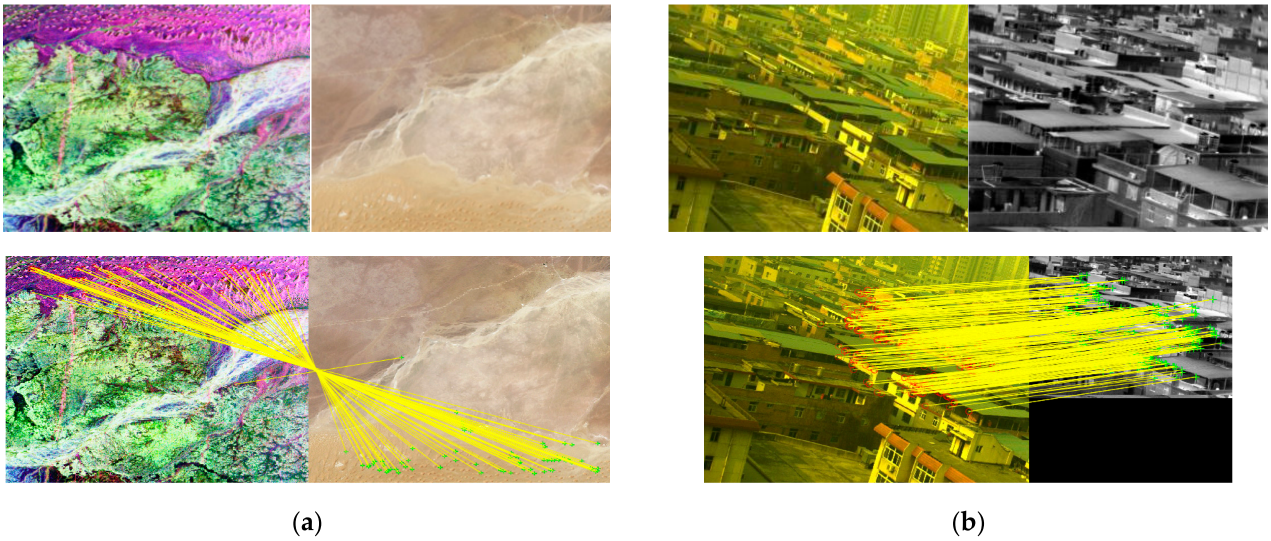


**Figure 1.** Nonlinear intensity difference and geometric deformation in multimodal images. (**a**) SAR-Optical image pair. (**b**) Optical-Infrared image pair. Original images (top line) and matching results obtained by our method (bottom line). The endpoints of the yellow lines in the matching results represent the corresponding matching point pairs.

To alleviate the difficulty mentioned above, this paper presents a robust multimodal remote sensing image registration framework using local statistical frequency information. The main contributions of our paper can be summarized as follows:

1. The maximum phase congruency optimization method is proposed, which is a guarantee for stable structural feature localization in multimodal remote sensing images and determines the center of the feature description regions.
2. To make full use of the frequency information and dig out structural features in image better, a joint local frequency information map that combines Log-Gabor filter responses over scales and orientations was constructed, which offers the main information to the feature descriptors.
3. The geometric and contrast invariant descriptors were generated through the selection of feature scales and the orientation statistics of the region to be described on the joint feature map, which is critical to achieve accurate registration.

Due to the improvements above, the proposed method is robust for multimodal remote sensing image registration with large geometric variations and significant contrast differences (Figure 1). The rest of this paper is organized as follows: The related works of registration for multimodal remote sensing images are reviewed and the main differences of our proposed method from others are described in Section 2. Detailed description of the proposed registration framework using local statistical information is given in Section 3. Comparative experiments and analysis are performed in Section 4. Finally, conclusions are drawn in Section 5.

## 2. Related Works

Over the years, a variety of multimodal remote sensing image registration methods have been proposed, which can be classified into two categories: global intensity-based methods and local feature-based methods [14]. Global intensity-based methods obtain the optimum transformation parameters via maximizing the similarity of global intensities between input images, mainly including mutual information [6], cross correlation [15], phase correlation [16], Fourier transformation [17] and wavelet features [18]. These methods perform well for the images with high correlation in global intensity or in transform domain. However, contrast reversal, occlusion, small area overlaps, and clutters occur frequently in some regions of input images, which makes the intensity-based methods unable to achieve an accurate registration.

In contrast, local feature-based methods first achieve feature matching by extracting and comparing local features in images and then compute the transformation via the correspondence of features. The advantage of these methods is that they can deal with significant geometric deformations as well as contrast differences in images. These methods can be classified into three groups: typical feature-based methods, deep learning-based methods, and local structural feature-based methods. The main idea of typical feature-based methods is to manually design different feature extraction models according to different applied scenarios. The related typical feature extraction methods include contour-based [19], line-based [20], region-based [21], and gradient distribution-based [8,22–25]. These methods have a good performance in the case that corresponding features are obvious or their gradient distribution is consistent in input images. The localization performance of several local features was compared in [26]; the best result was obtained by Root-SIFT [25]. However, they treat all image content equally, so they are highly sensitive to structural disparities caused by insignificant structures, such as shadow, illuminance, and resolution. Therefore, it will lead to severe degradation of matching performance when large contrast differences appear in multimodal remote sensing images. In recent years, with the prosperous application of deep learning in computer vision field, deep learning-based methods have been developed for remote sensing image registration [27–29], which can automatically learn high-level semantic features and get better matching performance compared to typical feature-based methods in some complicated cases. Especially, superglue [30] outperforms other learned approaches and achieves state-of-the-art results on the task of pose estimation in challenging real-world indoor and outdoor environments. However, these methods cannot deal with large geometric deformation (scale and rotation), and their performance is usually affected by the size of training data. As a result, the multimodal registration problem still cannot be effectively solved by the current deep learning-based methods.

Different from the methods of the first two groups, local structural features-based methods can extract more robust common features from different modalities and are less sensitive to contrast differences. Due to these advantages, they have been successfully applied to multimodal image registration [9,10,31–34]. As a typical representative of structural features, phase congruency of local frequency information was first proposed by Morrone et al. [35]. Subsequently, to improve the robustness of phase congruency to noise and contrast, Kovesi proposed a new sensitivity measure [36] and a highly contrast invariant localized feature detector [37]. Recently, Liu et al. [9] proposed mean local phase angle and frequency spread phase congruency by using local frequency information on Log-Gabor wavelet transformation space, which improved the robustness compared with traditional multimodal matching. To extract the structural features, Ye et al. [10]

developed the histogram of orientated phase congruency descriptor, which outperforms several methods in matching performance. Lately, they proposed channel features of orientated gradients (CFOG) [38] as an extension of [10] with superior performance in image matching and computational efficiency. Xie et al. [32] achieved multimodal image registration through combining phase correlation and multi-scale structural information.

Moreover, a structural information orientation histogram descriptor is constructed by concatenating the orientation of magnitude and the minimum moment [37]. Li et al. [33] proposed RIFT method for multimodal images registration via constructing a maximum index map (MIM) for feature description, which can achieve good performance on images with radiation-variation. However, the above-mentioned methods based on phase congruency are not enough to make full use of local frequency information, leading to limitations in processing large geometric deformation.

To address the registration problem of infrared and visible image, we constructed the maximally stable phase congruency (MSPC) descriptor using maximally stable extremal regions (MSER) [39] and local frequency information [31], nevertheless, in which MSER cannot obtain a higher repeatability for multimodal images. Subsequently, we developed a robust matching method for electro-optical by combining phase congruency with kernelized correlation filter (KCF) [40], which can first adjust the input images according to the platform parameters, and then get the registration results [13]. To deal with the multimodal remote sensing image registration with large geometric deformation and contrast difference, this paper extends our early works [13,31], by fully utilizing local frequency information. The main process of the proposed framework is shown in Figure 2. We first obtain the maximum phase congruency of local frequency information by optimizing the control parameters, and then extract the salient feature points according to their optimized phase congruency response. After that, we construct the geometric and contrast invariant descriptors (GCID) based on a joint local frequency information map (JLFM) that combines Log-Gabor filter responses over multiple scales and orientations. Finally, image matching and registration can be achieved by the correspondence of descriptors. In the experimental part, we performed a more thorough evaluation in terms of robustness and precision for the proposed registration framework with four different multimodal remote sensing image datasets.

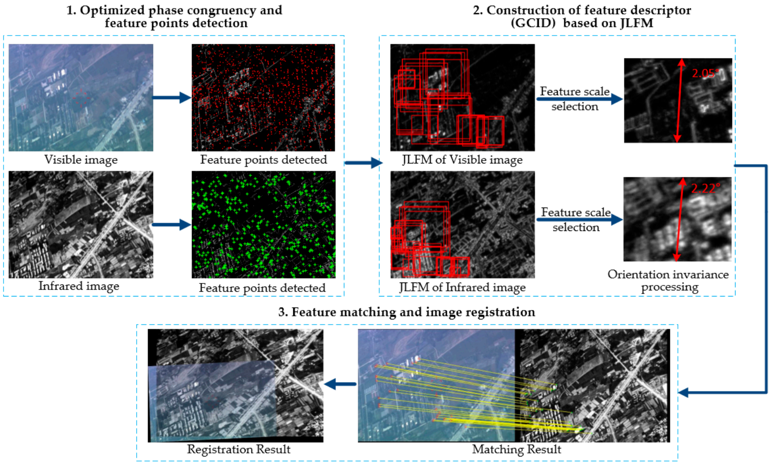


**Figure 2.** Illustration of a registration by the proposed framework. The framework consists of three parts: feature point detection using optimized phase consistency, construction of geometric and contrast invariant descriptors, feature matching and image registration.

## 3. Methodology

In this section, the robust multimodal remote sensing image registration framework using local statistical frequency information is presented. Section 3.1 introduces the calculation method of the maximum phase congruency of local frequency information through parameter optimization, and further realizes the location of salient feature points on the optimized feature map. Then, Section 3.2 presents the construction of GCID using JLFM in detail. Finally, the entire multimodal remote sensing image registration framework is given in Section 3.3.

### *3.1. Maximum Phase Congruency and Feature Detection*

Phase congruency provides a measure that is independent of the overall magnitude of the signal making it invariant to variations in image illumination and/or contrast. To improve the insensitive to noise and provide good localization, Kovesi [36] proposed a new sensitivity measure and noise compensation method for phase congruency, which can locate the features that remain constant over scales. The new measure is calculated by the following formula:

$$PC(x,y) = \frac{\sum_{s}\sum_{o} w_o(x,y)\lfloor A_{so}(x,y)\Delta\Phi_{so}(x,y) - T\rfloor}{\sum_{s}\sum_{o} A_{so}(x,y) + \varepsilon}, \tag{1}$$

where $w_o$ is a factor that weights for frequency spread, $A_{so}(x,y)$ is an amplitude of the Fourier component at position $(x,y)$, and $\Delta\Phi_{so}$ is a phase deviation function. $\varepsilon$ is a small constant to avoid the denominator being zero, and $T$ is a threshold that eliminates noise influence. The symbol $\lfloor\ \rfloor$ denotes that the enclosed quantity is equal to itself when its value is positive and zero otherwise. Based on this new measure, Kovesi [37] presented a highly localized feature detector whose responses are invariant to image contrast, which can be achieved as follows:

(1) The moment analysis equations at each point are calculated as following:

$$A = \sum_{o}(PC(\theta_o)\cos(\theta_o))^2, \tag{2}$$

$$B = 2\sum_{o}(PC(\theta_o)\cos(\theta_o))(PC(\theta_o)\sin(\theta_o)), \tag{3}$$

$$C = \sum_{o}(PC(\theta_o)\sin(\theta_o))^2, \tag{4}$$

(2) The maximum moment $M$ and minimum moment $m$ are given by,

$$M = \frac{1}{2}\left(A + C + \sqrt{B^2 + (A-C)^2}\right), \tag{5}$$

$$m = \frac{1}{2}\left(A + C - \sqrt{B^2 + (A-C)^2}\right). \tag{6}$$

A large value of $M$ indicates an edge feature point and A large value of $m$ means that point should be a corner; therefore, $M + m$ contains more features than anyone of them. To highlight features and to reduce computational complexity, we adopted the sum of phase congruency in multiple orientations as a candidate feature map (CFM) according to (7).

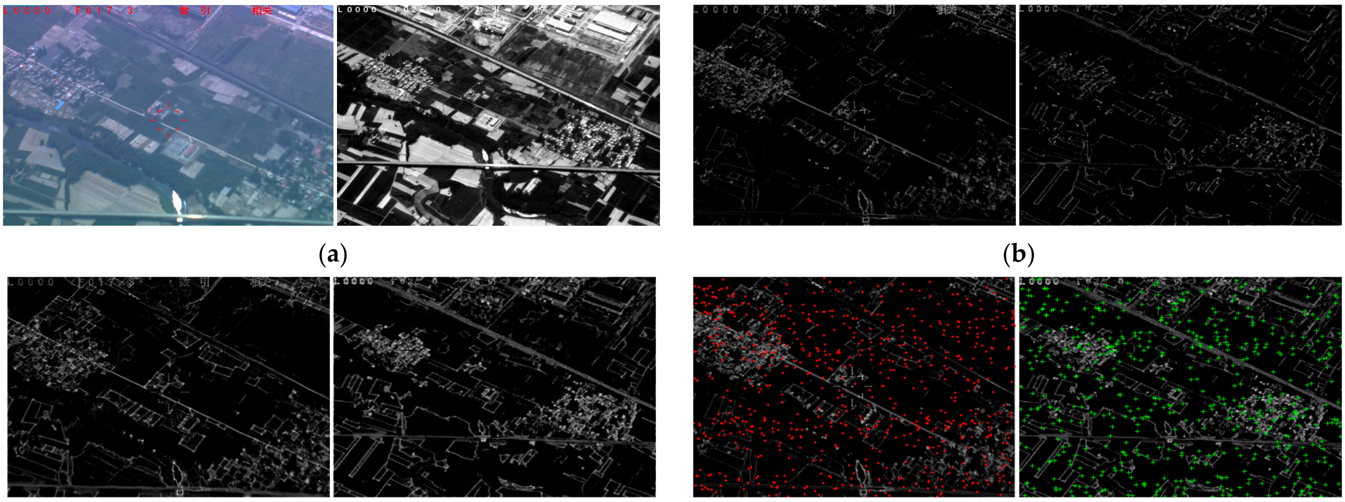
$$M + m = \sum_{o} PC^2(\theta_o) \leq \sum_{o} PC(\theta_o) \tag{7}$$

However, a very important factor rarely considered in the previous literature for phase congruency feature detection is the optimization and fine adjustment of control parameters. To improve robustness of phase congruency to nonlinear radiation deformations, RIFT [33] proposed MIM for feature description, nevertheless, other important structural information is lost. The literature [41] presented that the 2D-MSPC parameters can be optimally and automatically tuned by maximizing the norm of cost function formed by $M$ and $m$; however, their calculations are very complicated. In this paper, we use the similarity of CFMs extracted from two input images as the criterion of parameter optimization. The larger the similarity of structural features, the better the parameters.

There are eight parameters that mainly affect the performance of phase consistency, which can be denoted by $[N_s, N_o, \lambda_{\min}, \eta, \sigma, k, C_o, g]$ and described as Table 1. Among them, $N_s$ and $N_o$ are the number of filter scales and orientations respectively. Optimal values of $N_s = 4$ and $N_o = 6$ are verified through detailed experiments in [33]. The changes of $\lambda_{\min}$, $k$, $C_o$, and $g$ do not have significant effect on the phase consistency extraction results, which has been confirmed by our experiments. Therefore, the default values of those parameters are used in this paper. The values of scaling factor between successive filters $\eta$ and ratio of the standard deviation of the Gaussian $\sigma$ have a relatively large impact on the results, which can be seen from Figure 3.

**Table 1.** List of phase congruency parameters.

| Parameters | Range | Default | Meaning |
|---|---|---|---|
| $N_s$ | [3~6] | 4 | Number of wavelet scales. |
| $N_o$ | [1~6] | 6 | Number of filter orientations. |
| $\lambda_{\min}$ | ≥3 | 3 | Wavelength of smallest scale filter |
| $\eta$ | [1.3, 1.6, 2.1, 3] | 1.6 | Scaling factor between successive filters. |
| $\sigma$ | [0.1~1] | 0.55 | Ratio of the standard deviation of the Gaussian. |
| $k$ | [10~20] | 2 | Noise scaling factor. |
| $C_o$ | (0~1] | 0.5 | The fractional measure of frequency spread. |
| $g$ | [1~50] | 10 | Controls the sharpness of the transition in the sigmoid function. |



(**a**) (**b**) (**c**) (**d**)

**Figure 3.** Phase congruency feature detection by optimizing parameters. (**a**) Infrared and visible image pair. (**b**) CFMs obtained by default parameters. CFMs after parameter optimization (**c**) can extract more complete structural information than those using the default parameters, so that more high-quality feature points can be detected (**d**).

The cosine similarity of CFMs extracted from the input images was used as the criterion of optimizing parameters. The specific optimization process can be described as follows:

(1) CFMs are produced from the input images through changing the values of $\eta$ and $\sigma$, and keeping default values for other parameters.

(2) Compute the cosine similarity of CFMs by the following formula:

$$CS(CFM_{ref}, CFM_{sen}) = \frac{\overrightarrow{CFM}_{ref} \cdot \overrightarrow{CFM}_{sen}}{\left\| \overrightarrow{CFM}_{ref} \right\| \cdot \left\| \overrightarrow{CFM}_{sen} \right\|}, \quad (8)$$

where $CFM_{ref}$ and $CFM_{sen}$ are the corresponding CFMs of reference image and sensed image respectively. $\overrightarrow{CFM_{ref}}$ and $\overrightarrow{CFM_{sen}}$ are their histogram statistical vectors.

(3) The average cosine similarity of different CFMs for a group parameter is computed, and the optimal parameters can be determined according to the maximum cosine similarity of CFMs obtained by different parameter combinations.

In the algorithm above, each $\eta$ has 19 combinations by changing the value of $\sigma$ within the range of (0.1, 1) at 0.05 intervals. Therefore, the optimal parameters are determined by comparing the cosine similarity between 76 CFM pairs, those obtained from one multimodal remote sensing image pair with different parameters. In this paper, the optimal parameters for CFMs of a category image can be determined by their average cosine similarity. CFMs extracted from infrared and visible images by using optimized and default parameters are shown in Figure 3, from which we can see that the optimized CFMs (OCFMs) contained more structural information than those obtained by default parameters.

After the extraction of optimized CFMs from multimodal remote sensing image pairs, feature points will be detected from those CFMs by using FAST [42] and response ranking, which can be performed as follows:

(1) FAST is applied on OCFMs obtained from input multimodal remote sensing image pairs to get plentiful candidate feature points.

(2) To enhance the saliency of feature points, the extracted candidate feature points are ranked according to their response value in CFMs, and the top-k points will be selected as salient feature points.

(3) To ensure the uniform distribution of feature points, non-maximum suppression is implemented on their $n \times n$ neighborhood.

Among the above feature point detection methods, OCFMs and response ranking can ensure the saliency of feature points, and non-maximum suppression can promote the uniformity of their distribution. Examples of feature point detection on a multimodal image pair are shown in Figures 3d and 4, in which, the remote sensing images were acquired by different sensors (Infrared and Visible) or by the same sensor (Visible) but at different times, which leads to an obvious contrast difference between the input images. It can be seen from the results that the feature points detected by our proposed method had high saliency and repeatability and that their distributions were very uniform.

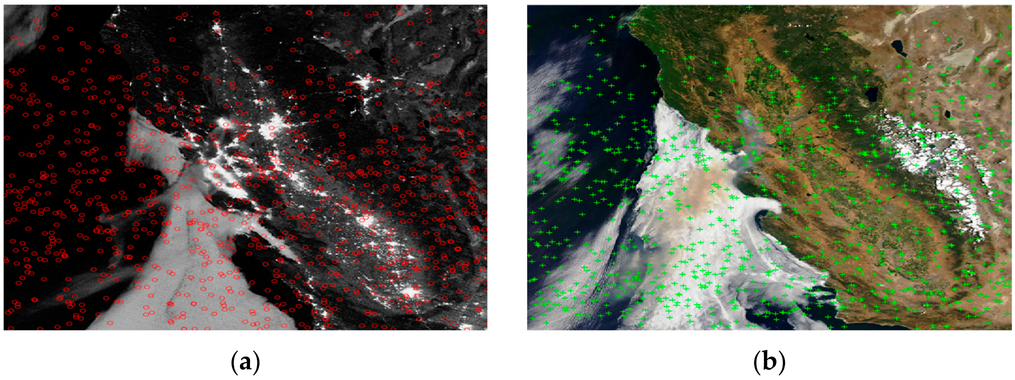


**Figure 4.** Salient feature points detection on multimodal remote sensing image pairs. (**a**,**b**) are Night image and Day image, respectively.

### 3.2. Construction of GCID

Salient feature points imply that there is important structural information around them. How to extract and describe structural information make descriptors have geometric deformation and contrast invariance, which are crucial for multimodal remote image registration. To explore the solution, GCID was constructed in this section. First, we built JLFM by combing Log-Gabor filter responses over scales and orientations, which could achieve a robust description for local feature under large contrast difference and geometric deformation. Then, the regions with scale and orientation invariance based on JLFM were located. Finally, GCID were generated by using BRIEF [43] in those located regions.

Li et al. [33] proposed that phase congruency maps are not suitable for feature description because they have small value and are sensitive to noise without obvious edge features in the image. Therefore, they presented the MIM measure instead of the PC map for feature description; however, the important frequency information that can reflect structural features is discarded. To make full use of frequency information and dig out structural features in an image better, we constructed a JLFM by combing Log-Gabor filter responses over scales and orientations as follows:

$$JLFM(x,y) = \frac{1}{S}\sum_{s}\sum_{o}\sqrt{E_{so}(x,y)^2 + O_{so}(x,y)^2}, \tag{9}$$

$$[E_{so}(x,y), O_{so}(x,y)] = [I(x,y) * LG^e(x,y,s,o), I(x,y) * LG^o(x,y,s,o)]. \tag{10}$$

where $E_{so}(x,y)$ and $\mathrm{O}_{so}(x,y)$ are response components produced by convolving the image $I(x,y)$ with the even-symmetric and the odd-symmetric Log-Gabor wavelets. JLFM has two advantages for feature description. First, it can integrate multi-orientation and multi-scale local frequency information that reflects structural features (Figure 5c) more comprehensively than phase congruency (Figure 5b). Second, the invariance of geometric deformation can be achieved through statistics of the response of JLFM at different scales and orientations, which is an important guarantee to obtain accurate matching of multimodal remote sensing images (Figure 5d).

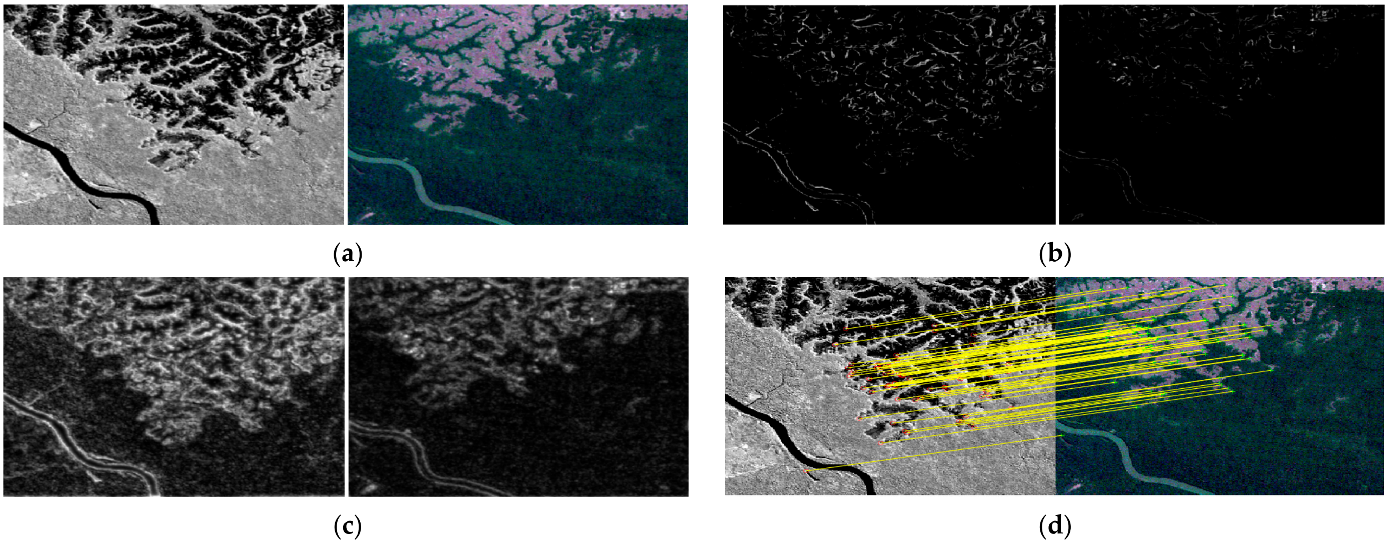


**Figure 5.** Comparison of different structural feature extraction methods. (**a**) Infrared and visible image pair. JLFMs (**c**) contains more local structural information than phase congruency maps (**b**) rendering good matching results (**d**).

### 3.2.1. Scale-Invariant Description Region

Scale and rotation are the main geometric deformation between multimodal remote images. Morel et al. [44] proposed that complex global transformations can be locally approximated by similarity transformations. Therefore, the scale and rotation invariance of structural feature descriptors are the main problems to be solved in this paper. First, the scale invariance of the description region is selected by the following steps:

(1) The scale of point $(x, y)$ can be computed as follows:

$$\sigma_p(x, y) = \underset{s \in \{1,2,\ldots N_s\}}{\text{argmax}} \left(A_s(x, y)\right), \tag{11}$$

$$A_s(x, y) = \sum_o \sqrt{E_{so}(x, y)^2 + O_{so}(x, y)^2}, \tag{12}$$

where $\sigma_p(x, y)$ is the assigned scale of point $(x, y)$ in the neighborhood of the feature point, which means that point $(x, y)$ has the maximum response at $\sigma_p(x, y)$ over $N_s$ scales. $A_s(x, y)$ is the sum of filter responses of point $(x, y)$ in all orientations at scale $s$.

(2) In the $n \times n$ neighborhood of the feature point, count the number of points that have the same $\sigma_p$ and use the scale $\sigma_F$ with the largest number of points as the scale of the feature point, which can be formulated as:

$$\sigma_F = \underset{(x,y) \in N_F, \sigma_p \in \{1,2,\ldots N_s\}}{\text{argmax}} \left(Number(\sigma_p(x, y))\right), \tag{13}$$

where $N_F$ is the $n \times n$ neighborhood of the feature point.

(3) The central frequency of Log-Gabor filter controls their scales; therefore, reciprocal of that is adopted to determine the description region radius of a feature point as follows:

$$R_F = R_0 \cdot \lambda_{\min} \cdot \eta^{\sigma - 1}, \tag{14}$$

where $\lambda_{\min}$ is the minimum wavelength, $\eta$ is the scale factor between successive filters, and $R_0$ is the initial radius. The above adaptive determination of the description regions can realize the scale invariance of the feature.

### 3.2.2. Rotation-Invariant Description Region

After the description regions are selected, the rotation invariance of the feature can be achieved by the statistics of histogram on the orientation map (ORM), which is defined as follows:

$$ORM(x, y) = \arctan\left(\frac{\sum\limits_o \cos(\theta) \sum\limits_s O_{so}(x, y)}{\sum\limits_o \sin(\theta) \sum\limits_s O_{so}(x, y)}\right), \tag{15}$$

where $\theta$ is the orientation of Log-Gabor filter and $O_{so}(x, y)$ is the response component generated by the odd-symmetric Log-Gabor wavelet.

Similar to SIFT [45], the dominant orientation of a feature point is determined by the histogram statistics in its description region on ORM, which can be described as follows:

(1) To improve the stability of description on image contrast, the range of the histogram is (0~180°); therefore, the histogram contains 36 bins, and every 5° is counted as one bin. Each bin can be calculated by ORM and Gaussian weighted JLFM as follows:

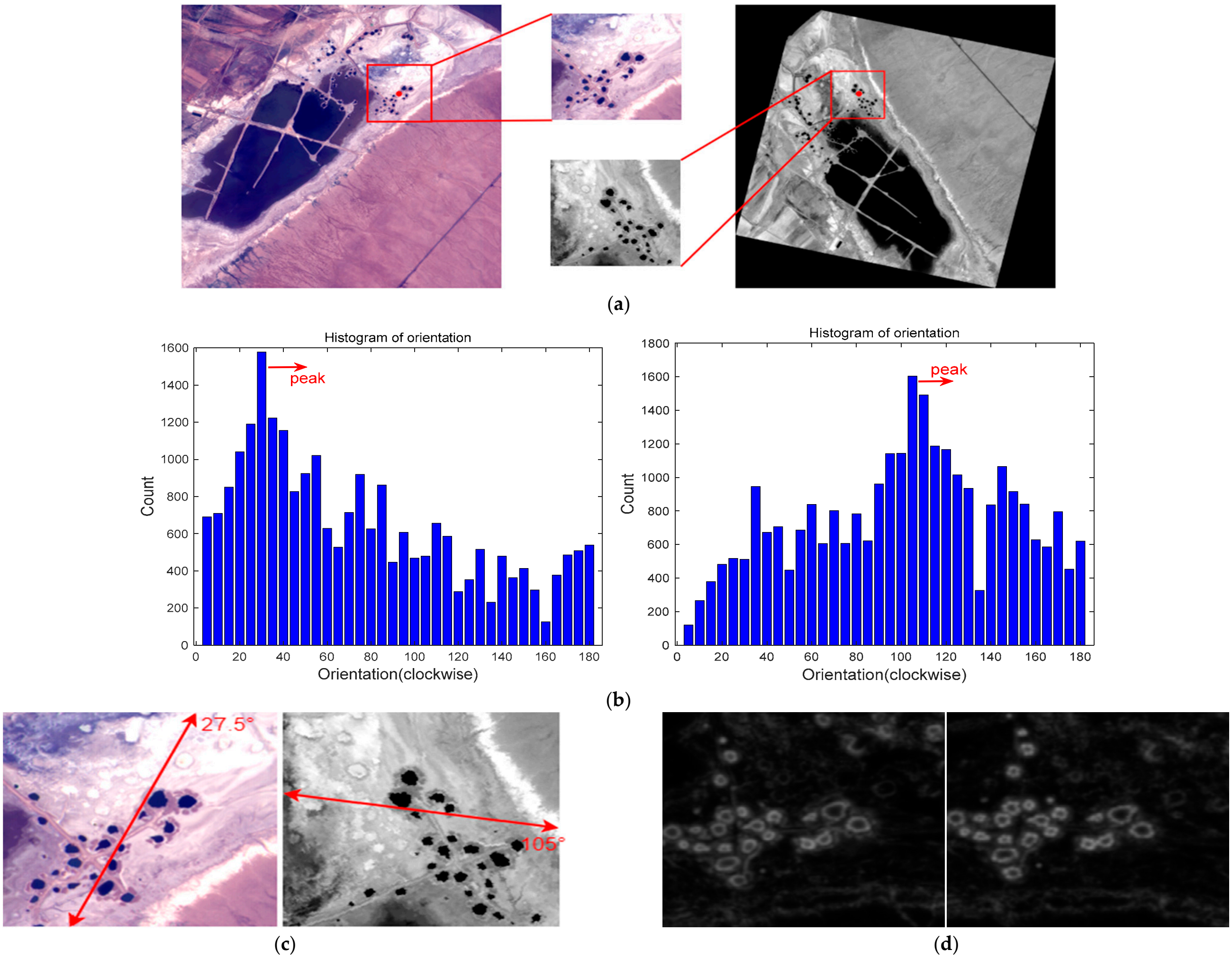
$$hist(s) = hist(s) + JLFM(i, j) * W(i, j), s \in [0, 255], (i, j) \in N_{R_F}, \tag{16}$$

$$W(i, j) = \exp(-(i^2 + j^2)/2 * (1.5 * \sigma_F))), \tag{17}$$

where $N_{R_F}$ is the description region of a feature point.

(2) Smoothing of the histogram is performed; the highest peak of the histogram is taken as the dominant orientation; the second highest peak that exceeds 80% of the highest peak is regarded as the auxiliary orientation.

Figure 6 shows the calculation process of scale and orientation invariant features from visible and NIR images. Description regions of two corresponding points are obtained by using Formulas (10)–(13), which are shown in the middle part of Figure 6a. Orientation histograms of description regions based on ORM and JLFM are given in Figure 6b. From the results of histogram statistics, dominant orientations of their description regions were assigned as 27.5° and 105°, respectively. Rotated description regions of the two corresponding points on JLFMs are shown in Figure 6d. From those results, we can see that the content of the corresponding feature point description regions on JLFMs had good similarity and consistency. After adjusting the scale and orientation of the description region of feature points, BRIEF was used for the construction of GCID as follows:



(**a**)

(**b**)

(**c**)

(**d**)

**Figure 6.** Execution process of image feature scale and rotation invariance. (**a**) The scale invariance of the corresponding feature descriptor is achieved. (**b**) Orientation histograms of description regions obtained by using ORM and JLFM. (**c**) Dominant orientations of the description regions. (**d**) Rotated description regions of on JLFMs. (**b**–**d**) ensure the rotation invariance of the descriptor.

In the description region of a feature point on JLFM, a pair of points were selected randomly and compared by the formula as follows:

$$\tau(p;x,y) = \begin{cases} 1, & if \;\; p(x) < p(y) \\ 0, & \quad otherwise \end{cases} \tag{18}$$

where $p(x)$ and $p(y)$ are the response values of random points $x = (u_1, v_1)$ and $y = (u_1, v_1)$, respectively. According to the above criteria, N pairs of random points were selected in the description region, and the binary assignment was performed to form a binary code, that is, BRIEF descriptor. The selection of random points in this paper obeyed the anisotropic Gaussian distribution; the Hamming distances were adopted to match the BRIEF descriptors.

### *3.3. Registration Framework by Using GCID*

After the feature detection and GCID construction proposed in Sections 3.1 and 3.2, the entire multimodal remote sensing image registration framework is given in this Section.

The workflow of the proposed multimodal remote sensing image registration framework is shown in Figure 7, and its steps can be described as follows:

(1) OCFMs are first computed from $I_{Inf}$ and $I_{Sen}$ by Formulas (1)–(8), respectively and then feature points are detected by FAST and non-maximum suppression on OCFMs.

(2) JLFMs are obtained from $I_{Inf}$ and $I_{Sen}$ by combing Log-Gabor filter responses over scales and orientations by Formulas (9) and (10).

(3) GCIDs from $I_{Inf}$ and $I_{Sen}$ are generated by using JLFMs and feature points obtained by steps (1) and (2) according to Formulas (11)–(18), respectively.

(4) Matching results of GCIDs from $I_{Inf}$ and $I_{Sen}$ are computed by their distance similarity; the outliers are removed by random sample consensus (RANSAC).

(5) Transformation is estimated according to the matching results; the registration of $I_{Inf}$ and $I_{Sen}$ are achieved.

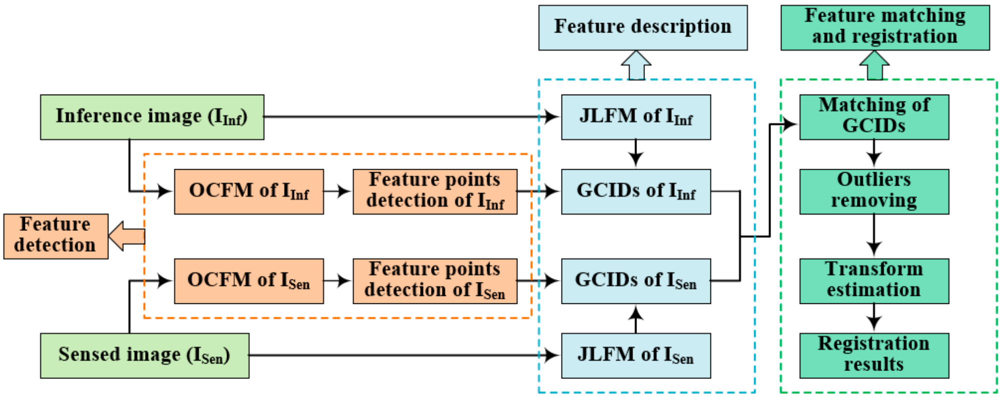


**Figure 7.** Workflow of the proposed multimodal remote sensing image registration framework.

The image registration workflow shown above was fully automatic, where parameter optimization could be done based on 2–3 image pairs from the same modal. The registration framework proposed in Figure 7 mainly included three parts: feature detection (orange), feature description (aqua green), feature matching and registration (green). Feature points localization implies that there are significant structural features around them, while the construction of GCID realizes the invariant description of structural features about the feature points. The main innovative work of this paper was the detection of salient features and the description of invariant structural features.

## 4. Experiment Results and Analysis

To demonstrate the effectiveness of the proposed method, four multimodal datasets were employed in comparative and evaluative experiments in this section. Those different sets of images are introduced in Section 4.1. To show the effect of parameter optimization, experiments with parameter optimization in image matching are implemented in Section 4.2. Then, robustness of the proposed method to geometric deformation is tested under different degrees of rotation and scale changes in Section 4.3. Finally, the comparative experimental results of the proposed method and state-of-the-art methods (Root-SIFT [25], RIFT [33], CFOG [38], SuperGlue [30]) are analyzed in Section 4.4. In addition, more visual matching results obtained by the proposed method are given in this section.

To quantitatively evaluate the matching performance, performance measures such as precision, root mean square error (RMSE), and median error (MEE) were adopted, which can be expressed as:

$$Precision = \frac{NCM}{NTM}, \tag{19}$$

where $NCM$ and $NTM$ is the number of correct matched and total matched point pairs, respectively.

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(p^{i}_{Inf} - T(p^{i}_{Sen})\right)^{2}}, \tag{20}$$

$$MEE = median\left\{\sqrt{\left(p^{i}_{Inf} - T(p^{i}_{Sen})\right)^{2}}\right\}_{i=1}^{L}, \tag{21}$$

where $\left\{p^{i}_{Inf}, p^{i}_{Sen}\right\}$ indicates the corresponding matched point pair, $p^{i}_{Inf}$ and $p^{i}_{Sen}$ are the feature point coordinates extracted from inference image and sensed image, respectively. $T$ is the ground-truth transformation between the two images, which can be obtained by manually selecting the corresponding point pairs. $L$ represents the number of matched point pairs, and $median\{\cdot\}$ returns the median value of a set.

### 4.1. Multimodal Remote Sensing Datasets

Multimodal image datasets (Figure 8) employed in comparative and evaluative experiments consist of the following four parts:

(1) Remote sensing dataset [14]: the dataset contains 78 image pairs, which can be divided into 7 modal types, such as UAV cross-season images, visible day-night, LiDAR depth-optical, infrared-optical, map-optical, optical cross-temporal, and SAR-optical images. These images have different resolutions ranging from 500 × 500 to 713 × 417, while the corresponding images have the same resolution; therefore, differences in contrast and inconsistencies in detail are the main changes between them.

(2) Computer vision dataset [14]: this dataset contains 54 image pairs, which includes 4 modal types, such as visible-infrared images, visible cross-season, day-night, and RGB-NIR images. These images have different resolutions ranging from 256 × 256 to 2133 × 1600; the corresponding images have the same resolution. Contrast difference and geometric deformation are the main changes between the image pairs.

(3) UAV dataset [13]: those visible and infrared images were captured at the same time from EOP on UAV, which consisted of 160 image pairs with discontinuous focus length change from 25 to 300 mm for the infrared camera and from 6.5 to 130.2 mm for the visible camera. The infrared images were captured by a mid-wavelength infrared camera operated in the 3–5 μm waveband with a size of 640 × 512. The visible images were captured by a lightweight CCD camera with a size of 1024 × 768. Therefore, large geometric deformation and contrast differences occurred between those image pairs.

(4) NIR-VIS dataset: this dataset was captured by Gaofen-14 satellite, which contained 40 image pairs. Near infrared (NIR) images were taken by a medium wavelength infrared camera and the visible images were taken by a visible camera; contrast difference is the main change between those image pairs.

Examples of the above four image datasets are shown in Figure 8. The dataset has a total of 9 modalities and 332 image pairs. We can see that geometric deformation and contrast changes between multimodal images are obvious. Their ground truth transformations were determined in advance by manual registration, and the match was considered correct if the RMSE was less than 3; otherwise, it was false.

### 4.2. Parameter Optimization Experiments

According to the parameter optimization description in Section 3.1, this section gives matching performance experiments before and after optimization to demonstrate the effectiveness of the method. The parameters to be optimized and their ranges are given in Table 2; from the table, we can see that the determination of a pair of optimal parameters $[\sigma, \eta]$ required the calculation of 76 combinations. In this experiment, CFMs were first extracted by calculating the sum of phase congruency in multiple orientations, and then the optimal parameters could be determined according to the maximum of the cosine similarity of CFMs obtained by different parameter combinations.

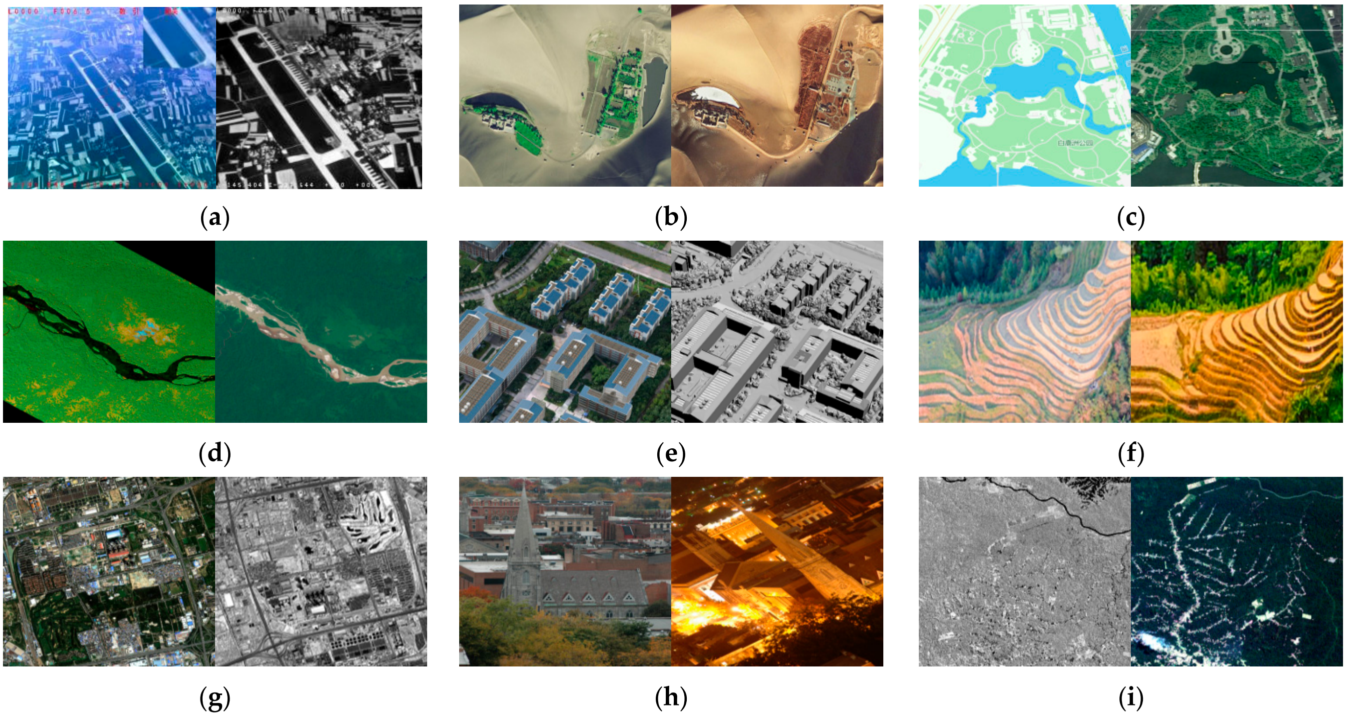


**Figure 8.** Examples of four multimodal image datasets employed in comparative and evaluative experiments. The image pair types from (**a**–**i**) are Visible-Infrared, Optical-Optical, Map-Optical, SAR-Optical, Visible-Depth, Visible Cross-Season, Optical-NIR, Visible Day-Night and Infrared-Optical, respectively.

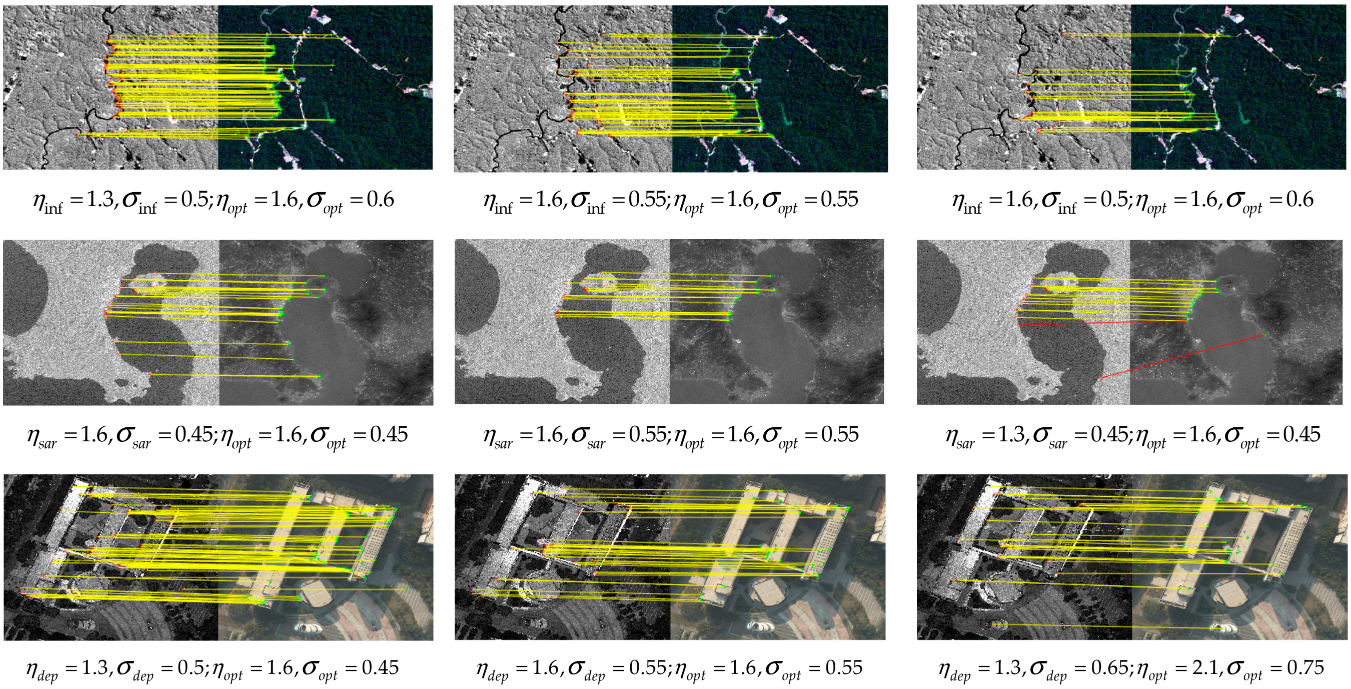
**Table 2.** The range of optimization parameters.

| Parameters | Optimization Range |
|---|---|
| $\sigma$ | [0.1, 0.15, . . . , 0.95, 1] (Interval 0.05) |
| $\eta$ | [1.3, 1.6, 2.1, 3] |

Three types of multimodal remote sensing images (infrared-optical, SAR-optical, and depth-optical) were selected to test the matching performance in the case of optimized parameters adopted in the process of feature extraction. The matching results before and after parameter optimization are shown in Figure 9; from the results, we can see that the number of correct matching point pairs was significantly increased, and that the distribution was more uniform after optimization. In addition, four types of multimodal remote sensing images were employed to measure the performance of optimized parameters in terms of average NCM. The comparative results are shown in Table 3, which shows a double times improvement in average NCM after parameter optimization over the pre-optimization.

### *4.3. Geometric Deformation Resistance Experiments*

To verify the robustness of the proposed method to geometric deformations, visible and infrared image dataset captured from UAV and remote sensing image dataset were used as experimental data, which had large geometric deformations and contrast differences. Since rotation and scale transformation are the most important geometric deformations, this section mainly tested the robustness of our method to these two deformations.



$\eta_{\inf} = 1.3, \sigma_{\inf} = 0.5; \eta_{opt} = 1.6, \sigma_{opt} = 0.6$

$\eta_{\inf} = 1.6, \sigma_{\inf} = 0.55; \eta_{opt} = 1.6, \sigma_{opt} = 0.55$

$\eta_{\inf} = 1.6, \sigma_{\inf} = 0.5; \eta_{opt} = 1.6, \sigma_{opt} = 0.6$

$\eta_{sar} = 1.6, \sigma_{sar} = 0.45; \eta_{opt} = 1.6, \sigma_{opt} = 0.45$

$\eta_{sar} = 1.6, \sigma_{sar} = 0.55; \eta_{opt} = 1.6, \sigma_{opt} = 0.55$

$\eta_{sar} = 1.3, \sigma_{sar} = 0.45; \eta_{opt} = 1.6, \sigma_{opt} = 0.45$

$\eta_{dep} = 1.3, \sigma_{dep} = 0.5; \eta_{opt} = 1.6, \sigma_{opt} = 0.45$

$\eta_{dep} = 1.6, \sigma_{dep} = 0.55; \eta_{opt} = 1.6, \sigma_{opt} = 0.55$

$\eta_{dep} = 1.3, \sigma_{dep} = 0.65; \eta_{opt} = 2.1, \sigma_{opt} = 0.75$

**Figure 9.** Comparison of matching performance of three different groups of parameters: optimized (**left column**), default (**middle column**), randomly selected (**right column**). The top-to-bottom rows show matching results of infrared-optical, SAR-optical, and depth-optical, respectively.

**Table 3.** Average NCM in different modal datasets by using optimized and default parameters.

| Datasets | Optimized Parameters $[\eta_{Inf}, \sigma_{Inf}, \eta_{Sen}, \sigma_{Sen}]$ | Average NCM | Default Parameters $[\eta_{Inf}, \sigma_{Inf}, \eta_{Sen}, \sigma_{Sen}]$ | Average NCM |
|---|---|---|---|---|
| Infrared-Optical | [1.6, 0.50, 1.6, 0.60] | 86 | [1.6, 0.70, 1.6, 0.70] | 63 |
| Depth-Optical | [1.3, 0.50, 1.6, 0.45] | 204 | [1.6, 0.70, 1.6, 0.70] | 101 |
| SAR-Optical | [1.6, 0.45, 1.6, 0.45] | 96 | [1.6, 0.70, 1.6, 0.70] | 58 |
| Map-Optical | [1.6, 0.45, 2.1, 0.65] | 147 | [1.6, 0.70, 1.6, 0.70] | 61 |

#### 4.3.1. Rotation Robustness Test

In this part, five images pairs were randomly selected from remote sensing and UAV datasets first, and then, for each pair of images, the sensed image was rotated from 30° to 180° at intervals of 30° to generate 6 rotated images. Therefore, 30 pairs of new images were obtained and used to test the robustness of the proposed method to rotational changes. The number of correct matches (NCM) of those images is shown in Figure 10; from the results, we can see that rotation had little effect on the matching results. The average fluctuation of NCM with rotation did not exceed 30%, which was mainly due to the rotation-invariant design of the proposed method. The matching results of the fifth image pair at different rotation angles are shown in Figure 11. The average NCM of the six rotated image pairs was more than 50, which was enough for accurate image registration. The proposed method successfully matched the rotated image pairs, although the sensed images had scale and rotation changes, which confirmed the effectiveness of the proposed method on rotation changes.

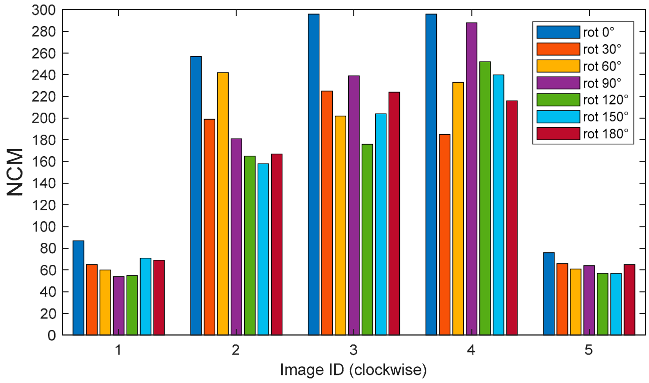


**Figure 10.** Rotation invariance test of the proposed method on five images pairs randomly selected from remote sensing and UAV datasets.

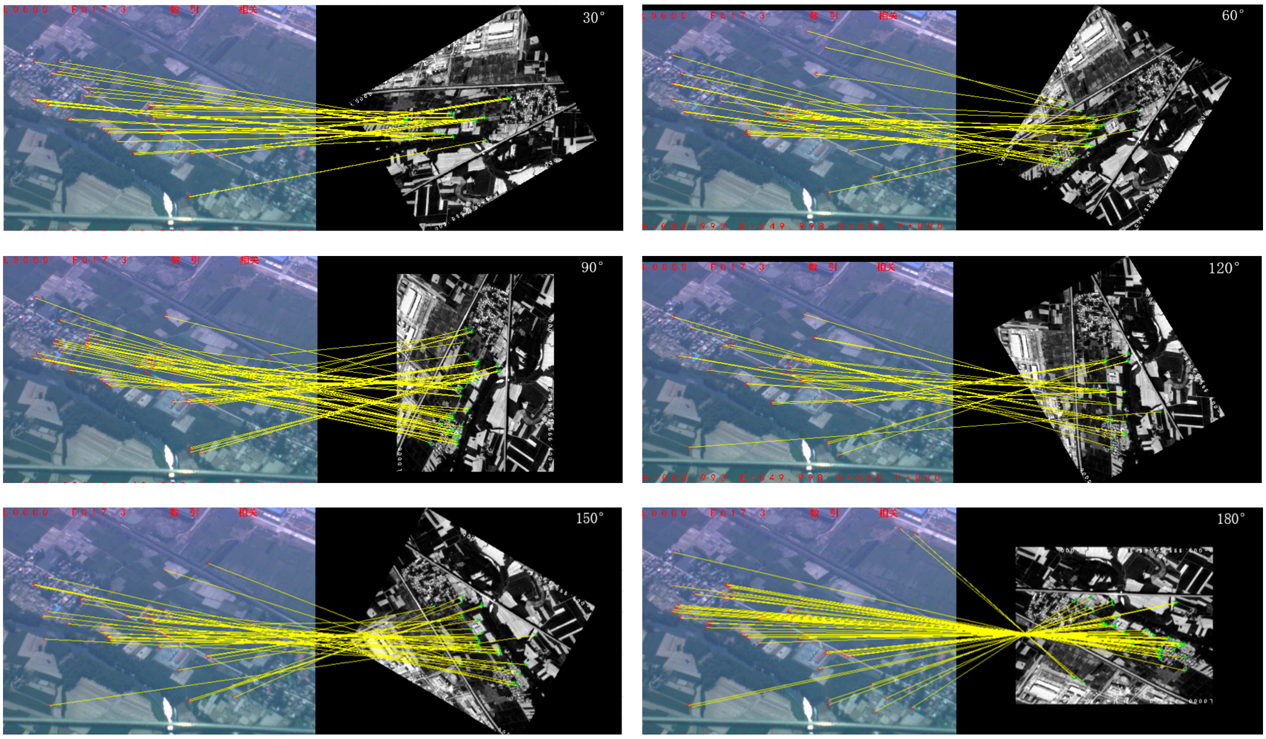


**Figure 11.** Matching results of image pairs with different rotation angles from 30° to 180° (at intervals of 30°) by using the proposed method.

### 4.3.2. Scale Robustness Test

Robustness to scale changes is key to testing the performance of registration methods; therefore, 20 visible and infrared image pairs from the UAV dataset were employed to evaluate the matching performance of the proposed method in this section.

We first chose a visible image as the inference image and then sequentially selected 20 infrared images with discontinuously varying focus lengths from 25 to 300 mm in the same scene as the sensed images. The scale difference between the input images changed from 0.8 to 3.2, examples of which are shown in Figure 12. The upper left image is the reference image, and the other images with scale changes are the sensed images, the scale of which gradually became larger. The NCM of visible image and 20 infrared image pairs selected from the UAV dataset is given in Figure 13. Although the NCM decreases as the scale becomes larger, the NCM obtained by our method was all above 20, which could meet the needs of image registration. Matching results of Figure 12 by using the proposed method are shown in Figure 14; from the results, we can see that the proposed method could handle large-scale changes due to the scale invariance processing in the descriptor GCID. The registration results in the lower right corner demonstrate the good performance of the proposed method under large scale differences.

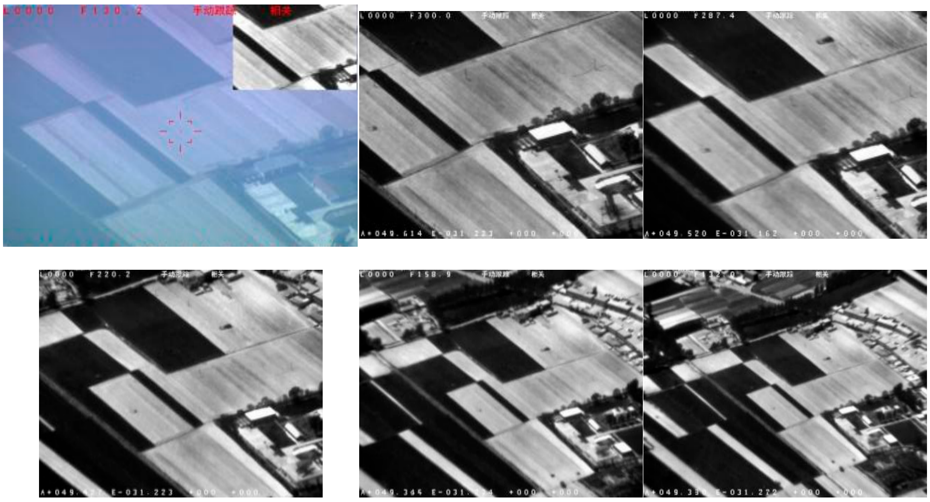


**Figure 12.** Visible and infrared image pairs examples from UAV dataset.

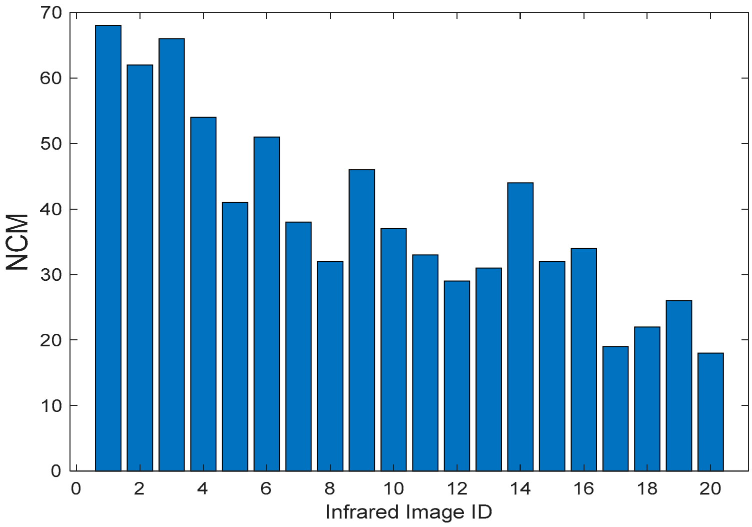


**Figure 13.** NCM obtained by our method for 20 image pairs with large scale changes.

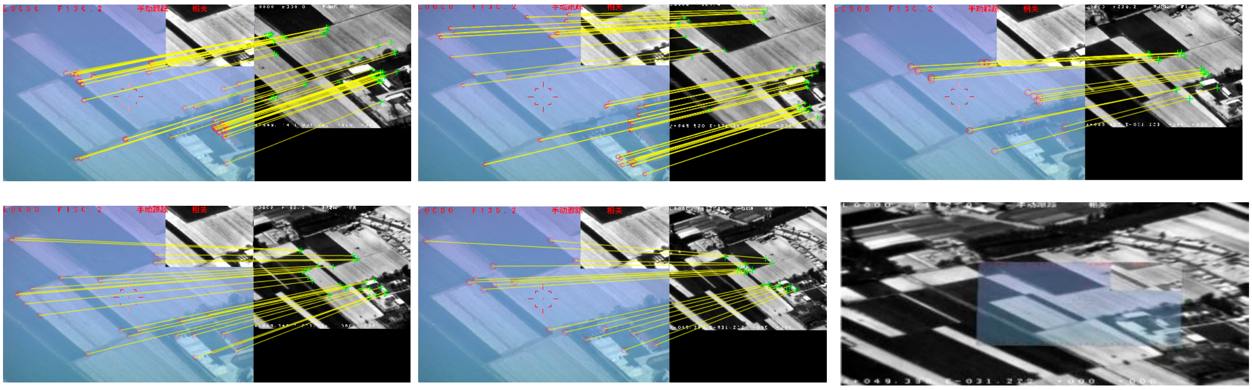


**Figure 14.** Matching results of Figure 12 and the registration result of the last pair of matching images.

### *4.4. Comparative Experiments*

To comprehensively evaluate the performance of the proposed method, comparative experiments were achieved via a comparison with state-of-the-arts (Root-SIFT [25], RIFT [33], CFOG [38], SuperGlue [30]) in terms of precision, RMSE, and MEE. The four methods chosen for comparison are excellent algorithms for image matching based on local features, among which Root-SIFT is a typical method for local feature description based on gradient distribution that has excellent performance in visual place recognition and localization. RIFT and CFOG are the representative works of multimodal remote sensing image registration in the past two years. SuperGlue won IMC-2020, as well as two other competitions at CVPR 2020.

For intuitive evaluation of matching performance, six multimodal remote sensing image pairs with geometric deformation and contrast difference were selected from a remote sensing dataset, a UAV dataset, and an NIR-VIS dataset. Figure 15 shows the matching results of six image pairs by using the five different methods, respectively. The matching result consists of two groups, each with five rows. The first group of images mainly contained contrast differences, subtle structural changes, and rotational changes, while the second group of images contained geometric deformations and contrast differences. The matching results of each group were obtained from top to bottom by Root-SIFT, RIFT, CFOG, SuperGlue, and the proposed method, respectively. For images with large contrast changes from the same sensor (the first pair of the first group), all five methods could obtain good matching point pairs, among which Root-SIFT obtained the least NCM, and RIFT obtained the most NCM. However, CFOG and SuperGlue failed for rotated images from different sensors (the third pair of the first group), mainly because they could not cope with large geometric transformations. Especially for SAR-Optical images with small structural changes and large contrast differences (the second pair of the first group), only the proposed method worked well, which was mainly due to the design of JLFM and the construction of GCID in our method. Due to the lack of ability to handle geometric changes, CFOG could not cope with infrared and visible images containing scale changes (the first pair of the second set), while other methods worked well. For multimodal images with large contrast differences (the second and third pairs of the second group), only RIFT and the proposed method were effective; obviously, our method obtained more NCM. Overall, the proposed method had the best matching performance. The registration results obtained by using the proposed method for six image pairs used above are shown in Figure 16; our method could achieve good registration results for multimodal images with large contrast differences and geometric deformations.

In addition, we selected ten representative images from the four datasets to compare the matching accuracy of the above five related algorithms. The comparative results are given in Figure 17. From the matching accuracy of remote sensing image datasets with contrast differences and inconsistencies in details, we could see that the average precisions obtained by CFOG and Root-SIFT were both below 65%, and those obtained by RIFT and SuperGlue were also below 80%. While the average precision of the proposed method was higher than 85%, because CFOG could not deal with inconsistencies in details and Root-SIFT was sensitive to the contrast differences. Due to the large geometric deformation in the UAV dataset, CFOG and RIFT had much lower average precisions than Root-SIFT and SuperGlue; however, none of these four methods achieved an average precision above 85%, while the proposed method achieved average precision over 90%. On the NIR-VIS dataset with mainly contrast difference, RIFT and the proposed method obtained higher precision than the other three methods. Due to the contrast difference and geometric deformation contained in the computer vision dataset, the matching accuracy of the other four methods fluctuated greatly, and their average precisions was lower than 80%, except that the precision of our method was higher than 85%.

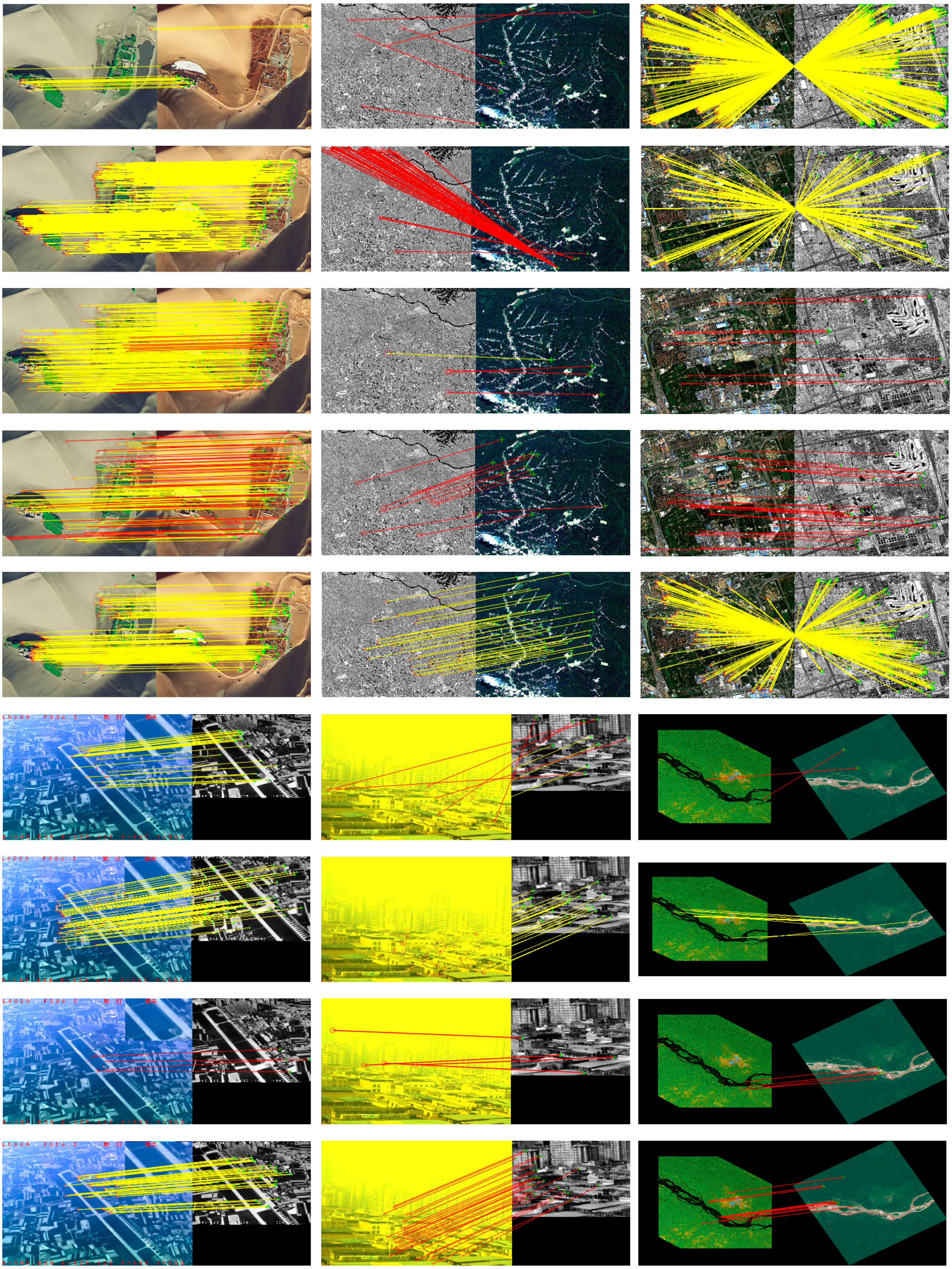


**Figure 15.** *Cont.*

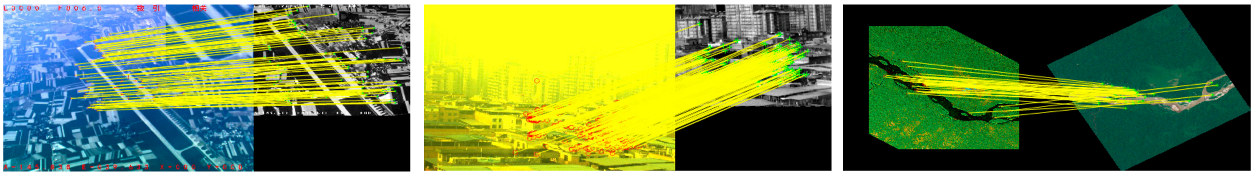


**Figure 15.** Comparison of matching performance of five methods on six image pairs. Five rows are in one group, and the top-to-bottom matching results of each group are obtained by Root-SIFT, RIFT, CFOG, SuperGlue, and the proposed method, respectively. The red lines indacate incorrect matching point pairs.

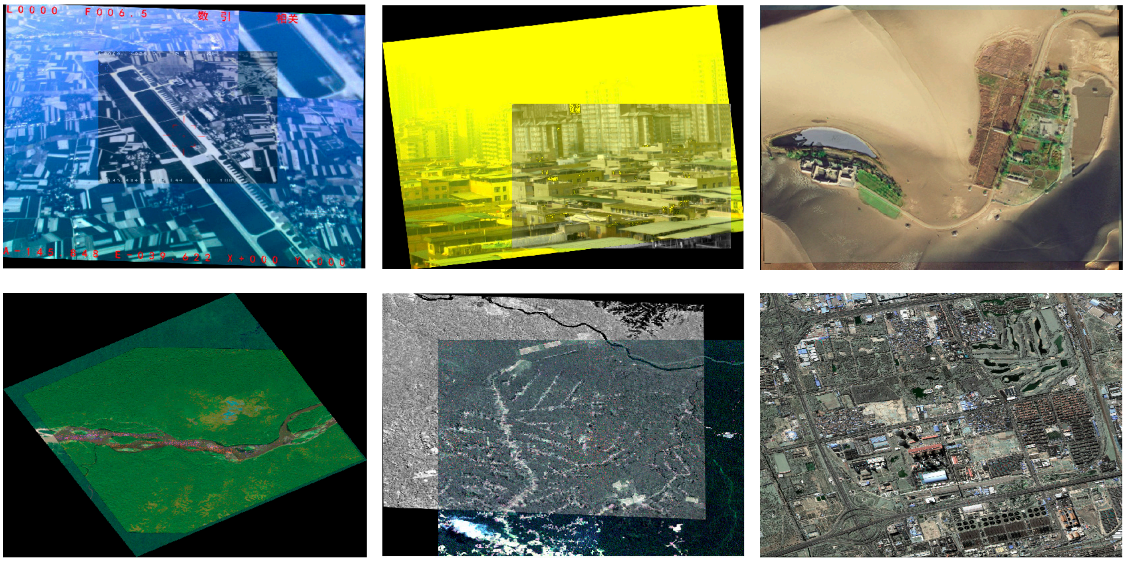


**Figure 16.** Registration results of six image pairs by using the proposed method.

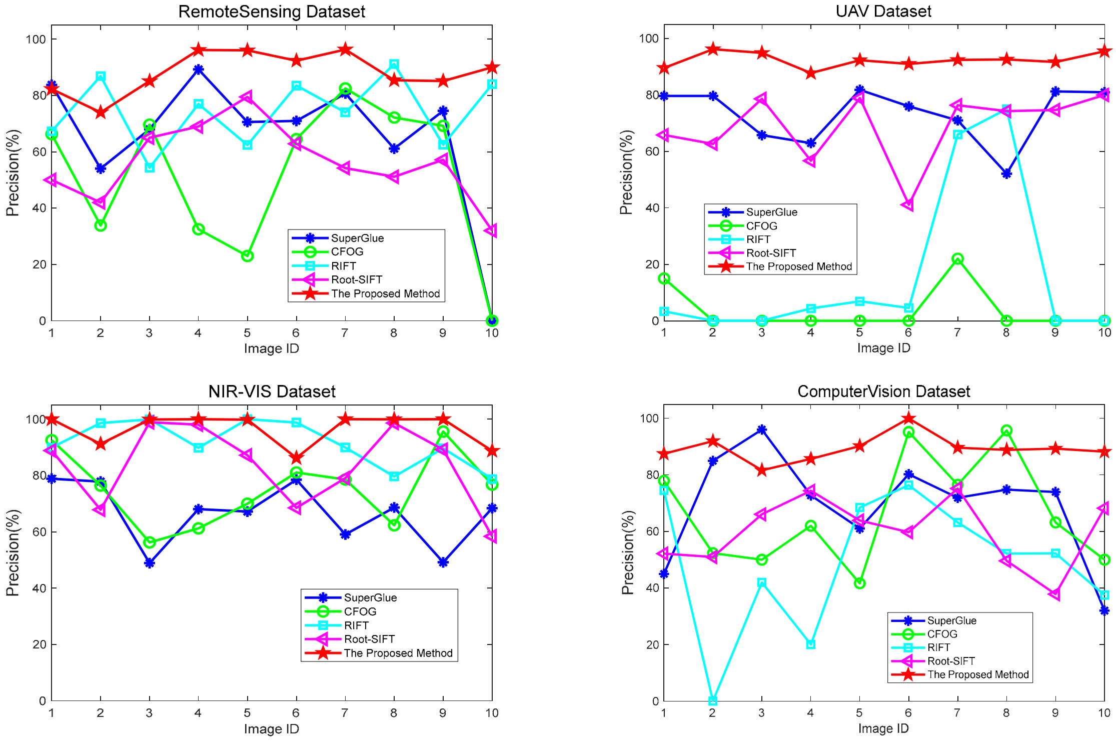


**Figure 17.** Matching precisions of five methods on ten image pairs selected from different datasets.

To evaluate the proposed method more comprehensively, we compared the matching precision of the five methods on all image pairs of the four datasets. The comparison results are given in Table 4, the performance of the proposed method is significantly better than other methods on UAV, remote sensing, and computer vison datasets. The average precision of our method on the three datasets was 99.14%, 79.31%, 80.13% and 81.48%, respectively. On the NIR-IR dataset only, the average precision of our method was slightly lower than that of CFOG, but still higher than the other three methods. Table 5 shows the quantitative evaluation results of the proposed method on different types of images. The average precision of the proposed method on different types of multimodal images was higher than 75%; the average NCM was more than 100. The average RMSE of registration on the datasets was less than 3 pixels; the MEE was less than 1.5 pixels, which met the requirements of practical application. Finally, RMSE of registration obtained by the five methods on different types of images are shown in Figure 18. The RMSEs of SuperGlue and Root-SIFT for different types of image data fluctuated greatly; their respective average values were higher than 5 pixels. The RMSEs of CFOG and RIFT were relatively stable; however, their respective average values were still higher than 4 pixels. The average RMSE of the proposed method was less than 3 on each type of images, which indicates that our method was robust to contrast differences and large geometric deformation between the multimodal remote sensing images.

**Table 4.** Comparison of Average Precision for five different methods.

| METHOD | AVERAGE PRECISION/% | | | |
|---|---|---|---|---|
| | NIR-IR | UAV | Remote Sensing | Computer Vision |
| SUPERGLUE | 76.46 | 40.24 | 44.72 | 63.31 |
| CFOG | 99.17 | 0 | 61.50 | 63.29 |
| RIFT | 99.10 | 12.57 | 77.68 | 39.31 |
| ROOT-SIFT | 98.15 | 53.69 | 41.14 | 47.98 |
| OURS | 99.14 | 79.31 | 80.13 | 81.48 |

**Table 5.** Quantitative evaluation of proposed method on different types of images.

| Metric | Cross-Season | Day-Night | Opti-Opti | Depth-Opti | Map-Opti | SAR-Opti | IR-Opti | NIR-Opti | VIS-IR |
|---|---|---|---|---|---|---|---|---|---|
| MEE | 1.446 | 0.9649 | 1.1478 | 1.026 | 1.3842 | 0.8799 | 1.1324 | 1.089 | 1.50 |
| NCM | 99 | 135 | 154 | 204 | 147 | 96 | 183 | 668 | 125 |
| RMSE | 2.4648 | 1.9649 | 1.8412 | 1.4847 | 2.378 | 1.2567 | 1.3177 | 1.6547 | 2.629 |
| Precision | 0.7678 | 0.8508 | 0.7816 | 0.8266 | 0.7595 | 0.8716 | 0.9347 | 0.9914 | 0.7931 |

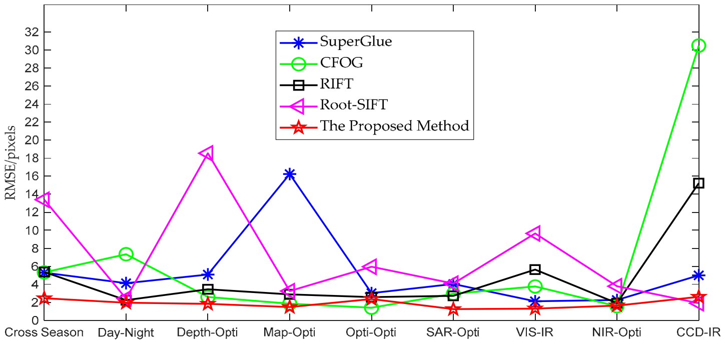


**Figure 18.** RMSEs of registration obtained by the five methods on different types of images.

The above verification and comparison experiments show that the proposed method had good adaptability to contrast difference and geometric deformation, which was mainly due to the following two points: (1) Parameter optimization in the process of local structural information extraction; (2) Construction of joint local frequency information map and design of scale and rotation invariance in the feature description process. More matching results of the proposed method on multimodal image pairs are shown in Figure 19, which demonstrates the effectiveness of our method for multimodal image matching with different variations.

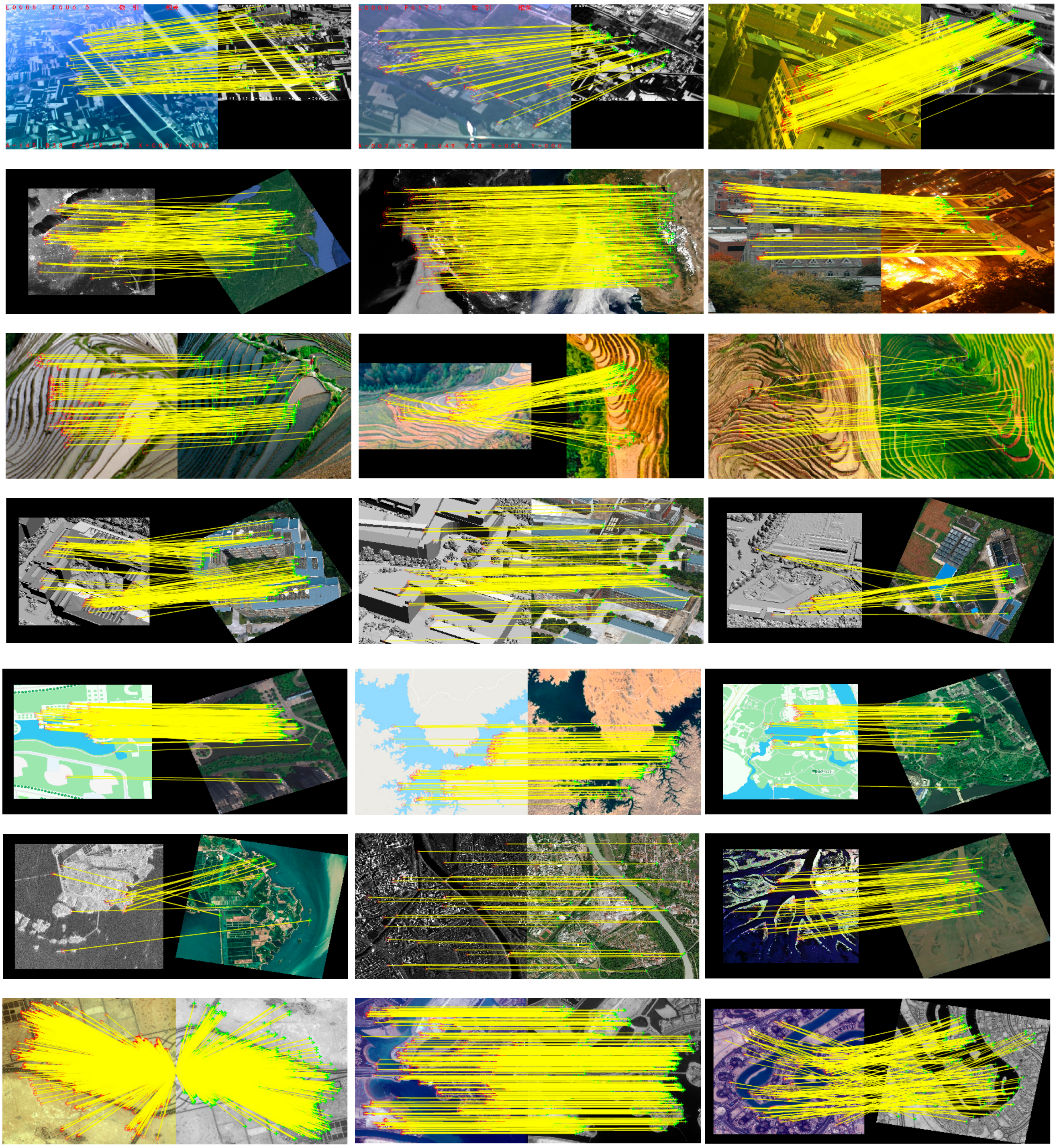


**Figure 19.** Matching results of the proposed method on different types of images.

In addition, the calculation of our algorithm mainly consisted of three parts: feature detection, feature descriptor construction, feature matching and image registration. The calculation of the feature detection part was mainly the generation of OCFMs. Assume that the maximum resolution of the image is $M \times N$, then computational complexity of the feature detection is $O(M * N * \log(M * N))$. In the feature description part, the computational complexity of GCID is $O(L * R^2)$, which is related to the number of extracted feature points $L$ and their description regions $R \times R$. For the last part, the complexity of feature matching is $O(L^2)$ and image registration are $O(M * N)$, respectively. Therefore, the total computational complexity of our method is $O(M * N * \log(M * N)) + O(L * R^2)$. The running time of the proposed registration method on the i7-9700@3.00GHz computer was less than 3 s when the resolution of the image was lower than 1000 × 1000 and the number of feature points was less than 1000.

## 5. Conclusions

To improve the robustness of multimodal remote sensing image registration with large contrast differences and geometric deformations, a robust local statistical information-based registration framework was developed in this paper. Salient feature points were firstly located according to the phase congruency response map that were optimized by control parameters. Then the geometric and contrast invariant descriptors were constructed based on a joint local frequency information map that combines Log-Gabor filter responses over multiple scales and orientations. Finally, geometric and contrast invariant descriptors are used to match the multimodal remote sensing image pairs and the registration can be achieved by the matching results. Four different multimodal image datasets were used to verify the effectiveness of the proposed method; and the results show that our method was robust to contrast and geometric variations. Through the comparative experimental analysis with the current popular four methods, it is shown that the matching accuracy and registration accuracy of the proposed algorithm are better than the four current popular methods, which confirms the superiority of the algorithm.

**Author Contributions:** X.L. conceived of and designed the experiments and wrote the paper; J.X. and X.X. performed the experiments and original draft preparation; Z.L. analyzed the data; R.L. and Y.L. offered the data curation; B.Z. provided the review and editing; Q.M. supervised the study and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key Research and Development Prpgram of China (No. B019030051), the Natural Science Foundation of Shannxi Province (No. 2020JM-195), the Fundamental Research Funds for the Central Universities (No. 20101216855, XJS210310), the National Natural Science Foundations of China (No. 61902296, 62002271), the Province Key R&D Program of Shaanxi (No. 2020LSFP3-15), the Key R&D Projects of Qingdao Science and Technology Plan (No.21-1-2-18-xx), and Guangxi Key Laboratory of Trusted Software (No. KX202061).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data that used in this study can be requested by contacting the first author.

**Acknowledgments:** We sincerely thank the authors of Root-SIFT, RIFT, CFOG, and SuperGlue for providing their algorithm codes to facilitate the comparative experiments, and thanks to the authors of reference [26] for collecting multimodal image data that enriches the diversity of experimental scenarios in this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Ma, Y.; Wu, H.; Wang, L.; Huang, B.; Ranjan, R.; Zomaya, A.; Jie, W. Remote sensing big data computing: Challenges and opportunities. *Future Gener. Comput. Syst.* **2015**, *51*, 47–60. [CrossRef]
2. Zhu, Z.; Luo, Y.; Qi, G.; Meng, J.; Li, Y.; Mazur, N. Remote sensing image defogging networks based on dual self-attention boost residual octave convolution. *Remote Sens.* **2021**, *13*, 3104. [CrossRef]
3. Zhu, Z.; Luo, Y.; Wei, H.; Li, Y.; Qi, G.; Mazur, N.; Li, Y.; Li, P. Atmospheric light estimation based remote sensing image dehazing. *Remote Sens.* **2021**, *13*, 2432. [CrossRef]

4. Paul, S.; Pati, U.C. A comprehensive review on remote sensing image registration. *Int. J. Remote Sens.* **2021**, *42*, 5396–5432. [CrossRef]
5. Zitova, B.; Flusser, J. Image registration methods: A survey. *Image Vis. Comput.* **2003**, *21*, 977–1000. [CrossRef]
6. Suri, S.; Reinartz, P. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. *IEEE Trans. Geosci. Remote Sens.* **2009**, *48*, 939–949. [CrossRef]
7. Shi, Q.; Ma, G.; Zhang, F.; Chen, W.; Qin, Q.; Duo, H. Robust image registration using structure features. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 2045–2049.
8. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 453–466. [CrossRef]
9. Liu, X.; Lei, Z.; Yu, Q.; Zhang, X.; Shang, Y.; Hou, W. Multi-modal image matching based on local frequency information. *EURASIP J. Adv. Signal Process.* **2013**, *2013*, 3. [CrossRef]
10. Ye, Y.; Shan, J.; Bruzzone, L.; Shen, L. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2941–2958. [CrossRef]
11. Fang, D.; Lv, X.; Yun, Y.; Li, F. An InSAR fine registration algorithm using uniform tie points based on Voronoi diagram. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1403–1407. [CrossRef]
12. Yang, K.; Pan, A.; Yang, Y.; Zhang, S.; Ong, S.H.; Tang, H. Remote sensing image registration using multiple image features. *Remote Sens.* **2017**, *9*, 581. [CrossRef]
13. Liu, X.; Ai, Y.; Tian, B.; Cao, D. Robust and fast registration of infrared and visible images for electro-optical pod. *IEEE Trans. Ind. Electron.* **2019**, *66*, 1335–1344. [CrossRef]
14. Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; Guo, X. A review of multimodal image matching: Methods and applications. *Inf. Fus.* **2021**, *73*, 22–71. [CrossRef]
15. Almonacid-Caballer, J.; Pardo-Pascual, J.E.; Ruiz, L.A. Evaluating Fourier cross-correlation sub-pixel registration in landsat images. *Remote Sens.* **2017**, *9*, 1051. [CrossRef]
16. Dong, Y.; Jiao, W.; Long, T.; He, G.; Gong, C. An extension of phase correlation-based image registration to estimate similarity transform using multiple polar Fourier transform. *Remote Sens.* **2018**, *10*, 1719. [CrossRef]
17. Tong, X.; Ye, Z.; Xu, Y.; Gao, S.; Xie, H.; Du, Q.; Liu, S.; Xu, X.; Liu, S.; Luan, K.; et al. Image registration with Fourier-based image correlation: A comprehensive review of developments and applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 4062–4081. [CrossRef]
18. Zavorin, I.; Le Moigne, J. Use of multiresolution wavelet feature pyramids for automatic registration of multisensor imagery. *IEEE Trans. Image Process.* **2005**, *14*, 770–782. [CrossRef]
19. Yang, Y.; Gao, X. Remote sensing image registration via active contour model. *Int. J. Electron. Commun.* **2009**, *63*, 227–234. [CrossRef]
20. Zhao, C.; Zhao, H.; Lv, J.; Sun, S.; Li, B. Multimodal image matching based on multimodality robust line segment descriptor. *Neurocomputing* **2016**, *177*, 290–303. [CrossRef]
21. Okorie, A.; Makrogiannis, S. Region-based image registration for remote sensing imagery. *Comput. Vis. Image Underst.* **2019**, *189*, 102825. [CrossRef]
22. Chang, H.H.; Wu, G.L.; Chiang, M.H. Remote sensing image registration based on modified SIFT and feature slope grouping. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1363–1367. [CrossRef]
23. Hong, Y.; Leng, C.; Zhang, X.; Pei, Z.; Cheng, I.; Basu, A. HOLBP: Remote sensing image registration based on histogram of oriented local binary pattern descriptor. *Remote Sens.* **2021**, *13*, 2328. [CrossRef]
24. Yan, X.; Zhang, Y.; Zhang, D.; Hou, N. Multimodal image registration using histogram of oriented gradient distance and data-driven grey wolf optimizer. *Neurocomputing* **2020**, *392*, 108–120. [CrossRef]
25. Arandjelovic, R.; Zisserman, A. Three things everyone should know to improve object retrieval. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2911–2918.
26. Sun, X.; Xie, Y.; Luo, P.; Wang, L. A dataset for benchmarking image-based localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7436–7444.
27. Zhang, H.; Ni, W.; Yan, W.; Xiang, D.; Wu, J.; Yang, X.; Bian, H. Registration of multimodal remote sensing image based on deep fully convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3028–3042. [CrossRef]
28. Zhang, J.; Ma, W.; Wu, Y.; Jiao, L. Multimodal remote sensing image registration based on image transfer and local features. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1210–1214. [CrossRef]
29. Li, Z.; Zhang, H.; Huang, Y. A rotation-invariant optical and SAR image registration algorithm based on deep and gaussian features. *Remote Sens.* **2021**, *13*, 2628. [CrossRef]
30. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning feature matching with graph neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4938–4947.
31. Liu, X.; Ai, Y.; Zhang, J.; Wang, Z. A novel affine and contrast invariant descriptor for infrared and visible image registration. *Remote Sens.* **2018**, *10*, 658. [CrossRef]
32. Xie, X.; Zhang, Y.; Ling, X.; Wang, X. A novel extended phase correlation algorithm based on Log-Gabor filtering for multimodal remote sensing image registration. *Int. J. Remote Sens.* **2019**, *40*, 5429–5453. [CrossRef]

33. Li, J.; Hu, Q.; Ai, M. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Trans. Image Process.* **2019**, *29*, 3296–3310. [CrossRef]
34. Yu, G.; Zhao, S. A new feature descriptor for multimodal image registration using phase congruency. *Sensors* **2020**, *20*, 5105. [CrossRef] [PubMed]
35. Morrone, M.C.; Ross, J.; Burr, D.C.; Owens, R. Mach bands are phase dependent. *Nature* **1986**, *324*, 250–253. [CrossRef]
36. Kovesi, P. Phase congruency: A low-level image invariant. *Psychol. Res.* **2000**, *64*, 136–148. [CrossRef]
37. Kovesi, P. Phase congruency detects corners and edges. In Proceedings of the Australian Pattern Recognition Society Conference, Sydney, SA, Australia, 10–12 December 2003; pp. 309–318.
38. Ye, Y.; Bruzzone, L.; Shan, J.; Bovolo, F.; Zhu, Q. Fast and robust matching for multimodal remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9059–9070. [CrossRef]
39. Donoser, M.; Bischof, H. Efficient maximally stable extremal region (MSER) tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 553–560.
40. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef] [PubMed]
41. Alavi, S.M.M.; Zhang, Y. Phase congruency parameter optimization for enhanced detection of image features for both natural and medical applications. *arXiv* **2017**, arXiv:1705.02102.
42. Rosten, E.; Drummond, T. Machine learning for high-speed corner detection. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 430–443.
43. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary robust independent elementary features. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 778–792.
44. Morel, J.M.; Yu, G. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sci.* **2009**, *2*, 438–469. [CrossRef]
45. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]