Rotation-Invariant Siamese Network for Low-Altitude Remote-Sensing Image Registration

Yuyan Liu, Xiaoying Gong, Jiaxuan Chen, Shuang Chen, and Yang Yang, *Member, IEEE*

***Abstract*—Multiple-view change caused by small unmanned aerial vehicles (UAVs) monitoring the ground, resulting in image distortion, multiview transformation, and low overlap. Thus, such change has a strong effect on the accuracy of image registration. In this study, we utilize a Siamese network to deal with the complexity registration of low-altitude remote-sensing images. A robust neighbor-guided patch representation is designed to describe feature points based on neighborhood relation reconstruction, and patch selection. The network is trained based on rotation-invariant layer to solve the inevitable rotation, and nonrigid deformation caused by multiview images in low-altitude remote-sensing images. With only three training images involving 4500 putative matches, the experiment results demonstrated that the learned network can process the scenarios of yaw rotation, pitch rotation, mixture, and extreme (e.g., mixture, scaling, and distortion occur simultaneously) of UAV better than other six state-of-the-art methods.**

***Index Terms*—Feature matching, low-altitude remote sensing, mismatch removal, registration, rotation-invariant.**

I. Introduction

SMALL unmanned aerial vehicles (UAVs) represent a trend in the development of airborne remote-sensing platforms in recent years. Due to their low cost, flexible flight, and high relevance, small UAVs can provide a customizable airborne platform that can be installed with a variety of sensors to quickly acquire high-resolution images in small areas where flying is difficult. With the development of related technologies, the application fields for small UAVs have expanded from early military applications to urban monitoring, postdisaster reconstruction, precision agriculture [1], ocean monitoring [2], forest resource survey [3], and change detection [4]. Image registration is an important fundamental subject in remote sensing, and its purpose is to find an optimal alignment between two or more images, which can be acquired at different times, from different viewpoints, or by different sensors.

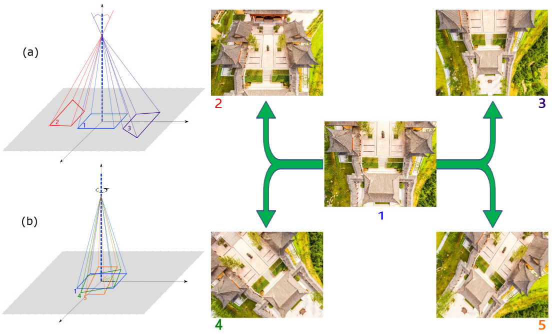

Fig. 1. Two representative scenarios are used to illustrate the multiview changes in captured images during ground monitoring. (a) Pitch rotation of UAV due to natural or human factors. Picture 2 and 3 correspond to the images rotated vertically up and down, respectively. (b) UAV is rotated clockwise and counterclockwise in horizontal direction. Picture 4 and 5 correspond to the images rotated clockwise and counterclockwise, respectively.

When the UAV is monitoring the ground, it is cannot avoid the influence of flight attitude (yaw, pitch, roll) due to the following factors:

1) natural factors such as wind speed and direction, complex terrain;
2) human factors such as improper operations, differences in flying height, speed, and posture; and
3) equipment factors such as battery issues and GPS positioning error, which cause the acquired images to be squeezed, twisted, stretched, and offset relative to the target position of the ground [5].

Pitch or roll rotation results in nonrigid deformation of the image, such as stretching and squeezing, which is prone to many-to-one feature points. Coupled with yaw rotation, the image pairs are prone to low overlap and nonrigid deformation, resulting in a large number of redundant points. In addition, the image can also be distorted, scaled, and offset relative to the target position of the ground. Therefore, image pairs of the same scene taken from different viewpoints often contain a larger rotation angle. Thus, accurate registration of UAV images is a prerequisite for subsequent applications. Fig. 1 shows some examples of multiview images when the flight attitude of UAV is affected by the aforementioned factors. Our contribution is stated in the related work.

The rest of this article is organized as follows. Section II describes research background and related work. Section III introduces the proposed registration method in detail. Section IV presents the extensive experiments and analysis of our proposal. The conclusion is drawn in Section V.

Manuscript received June 9, 2020; revised July 18, 2020, August 10, 2020, August 17, 2020, and September 7, 2020; accepted September 13, 2020. Date of publication September 18, 2020; date of current version October 1, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 41971392, 41661080 and in part by the Yunnan Ten-Thousand Talents Program. *(Yuyan Liu and Xiaoying Gong contributed equally to this work.) (Corresponding author: Yang Yang.)*

The authors are with the School of Information Science, and Technology, and Laboratory of Pattern Recognition, and Artificial Intelligence,Yunnan Normal University, Kunming 650500, China (e-mail: yuyan_l19@163.com; gxy956289900@gmail.com; jrbook_chen@foxmail.com; chen_shuang283@163.com; yyang_ynu@163.com).

Digital Object Identifier 10.1109/JSTARS.2020.3024776

II. Related Work

The following two common approaches are used in image registration: area-based and feature-based methods. Area-based methods use the original pixel intensity to find matching information between two images, in which normalized cross-correlation [6], Fourier methods [7], and mutual information [8] are widely accepted. Owing to limitations of relying on windows, area-based methods are highly sensitive to intensity change and illumination, and our study mainly focuses on the feature-based method.

Feature-based methods used image salient features derived by feature extraction algorithm rather than intensity values for matching purpose. It can recover point-to-point correspondences between an image pair through the following three different strategies:

1) feature matching;
2) point set registration; and
3) mismatch removal.

These strategies are discussed as follows.

Feature matching is the earliest strategy applied to image registration, which aims to find the correct corresponding relationship by evaluating the similarity of feature points, such us the SIFT algorithm proposed by Lowe [9] and its improved versions [10]–[14]. SIFT-OCT, proposed by [12], skips the first octave of scale space to reduce the impact of noise. AB-SIFT [15] uses an adaptive binning strategy to calculate local feature descriptors, which are calculated on a normalized region defined by the uniform robust Hessian affine algorithm. Ma *et al.* [14] proposed modified SIFT feature and feature matching method based on the new gradient definition to overcome the difference in image intensity between remote image pairs, where the feature matching combines the location, scale and direction of each key points. MS-SIFT [13] defines a new mode of scale, rotation difference, and translation for SIFT feature points, and performs reliable filtering of outlying feature correspondences by horizontal and vertical shifts among all the corresponding SIFT key points.

However, to prevent unstable inlier matches, the thresholds in such methods are normally fixed at a relatively high similarity measurement, which can lead to losses in a part of potential inlier pairs. Furthermore, complex situations such as various rotations, large deformation, and scaling changes caused by multiview changes may significantly degrade the discriminative ability of feature descriptor.

Point set registration is designed to solve the geometrical feature problem in feature matching. This strategy performs the following two alternating steps: correspondence estimation and transformation update. The key idea is to adjust the initial geometrical structure and location of the source point set (by the transformation update) so that it can gradually become more similar to the target point set, and then correspondence estimation using geometrical features can be facilitated. Diverse solutions are also available in the phase of point set registration. Thin-plate spline (TPS) robust point matching [16] established a general framework for nonrigid registration based on soft assignments [17], deterministic annealing [18], and TPS interpolation [19]. Identifying correspondence function (ICF) [20] captures the relationships between corresponding points by the mapping between a matched image pair, and the mismatch is rejected if its corresponding function is inconsistent with the estimated transformation. A typical Gaussian mixture model (GMM) based method is coherent point drift [21], which takes the moving and fixed point sets as the centroid of Gaussian components and data, respectively, and then iteratively update point locations under the expectation maximization framework. Jian *et al.* [22] employed the GMMs to represent the point sets, and two Gaussian mixtures are aligned by minimizing the discrepancy between them instead of the log-likelihood function. Global-local mixture distance [23] estimates the one-to-one correspondence from the global-local scales, a deterministic process is used to gradually control the priority between the two scales. Global-local correspondence and transformation estimation [24] approximates the correspondence using the mixed-feature Gaussian mixture model, and then updates the transformation by combining the global feature of point-to-point Euclidean distance and the local feature of shape distance. Yang *et al.* [25] proposed a method that mixes the intensity information with the geometric information to form a mixed-feature based correspondence estimation using the GMM. The gravitational approach [26], which transformed the point-set registration problem into a modified n-body problem with additional constraints, is a new astrodynamic-inspired rigid point-set registration algorithm. It mimics a template point set that moves in viscous medium under the gravitational action of a reference point set. Vector field consensus [27] interpolates a vector field between the two point sets to solve for correspondence, which involves estimating a consensus of inlier points whose matching follows a nonparametric geometrical constraint. Ma *et al.* [28] created a set of putative correspondences by feature similarity to remove outliers, and then developed locally linear transforming to preserve the local structure between adjacent feature points, which is highly robust to severe outliers.

However, in real applications, the weights between different features, model adaptability, and nonadaptive optimization parameters are extremely sensitive to various registration patterns (i.e., large deformations, various rotations, scaling change, and extreme mixture scenarios).

1) Weights: when mixed features are used to evaluate similarity, the weight of each feature is uncertain.
2) Model adaptability: when the model is used to fit the distribution of points, the correspondence estimation may be biased if the point set distribution is irregular.
3) Nonadaptive optimization parameters: many parameters are involved in parameter optimization, these parameters have no priors, and adaptive parameter optimization cannot be implemented for different registration forms.
4) Local image information: these methods completely discard the abundant information on local image descriptors, and their performance deteriorate when the image pair undergoes nonrigid deformation.

Mismatch removal is designed to solve the intensity-based feature problem, and aims to employ one or more additional feature descriptors to further estimate inliers and outliers based on a prematching result. The random sample consensus [29] and its variants maximum likelihood estimation sample consensus (MLESAC) [30] and progressive sample consensus [31], and propose to use a hypothesize-and-verify framework to eliminate mismatch correspondences and have been widely used for automatic registration of remote-sensing images [32], [33], [34]. The resampling methods rely on a predefined parametric model; thus, these become less efficient when the underlying image transformation is nonrigid and also tend to severely degrade if the mismatch proportion increases [20]. In the field of graph matching methods, representative studies such as spectral matching [35], dual decomposition [36], graph shift [37], and deformable graph matching [38] have also been applied for image registration. Although the model selection and robust matching results have considerable flexibility, such methods are unfavored by the nonpolynomial hard nature. An efficient approach called locality preservation matching (LPM) [39] is recently proposed for mismatch removal by maintaining the local neighborhood structures of those potential true matches. Its variant-guided locality preservation matching (GLPM) [40] further employs a small putative set with a high inlier ratio to guide the matching. SIR [41] enriches the inlier pool and refines the transformation by repeated iteration to select the inliers from the candidate pool, and the refined transformation prunes inconsistent mismatches to alleviate the incoming matching ambiguity. RFM-SCAN [42] can adaptively cluster a set of putative matches into several inlier groups with motion consistency and an outlier group with linearithmic time complexity.

Nevertheless, the mismatch removal method based on handcrafted feature lacks generalization ability. No one can simultaneously guarantee deformation, scaling, rotation or extreme mixtures invariant, and the weight problem remains in the aforementioned multifeature-based mismatch removal.

In recent years, deep learning has achieved great success in the areas of computer vision [43], speech processing [44], and image processing [45], [46]. The application of deep learning to remote sensing has become a hotspot [47]–[49], and is also applied for image registration and patch matching. Zagoruyko *et al.* [50] used the following three basic models: Siamese, pseudo-Siamese, and two-channel, which are applied to learn directly from raw image pixels a general similarity function for patches. Wu *et al.* [51] proposed a learning-based image registration framework, in which feature selection uses a convolutional stacked autoencoder to identify inherent deep feature representations in image patches. Learning a two-class classifier for mismatch removal (LMR) [52] learns a classifier to distinguish false and true putative matches, and the mismatch removal problem is transformed into a two-class classification problem. However, LMR only uses the neighborhood of the point and its topological structure, and lacks the pixel information around the point. Wang *et al.* [53] proposed a deep neural network to learn the mapping between patch pairs from sensed and reference images and output their matching label for later registration. However, the network encounters difficulty in distinguishing image patches with similar backgrounds and sensitive to various rotation and extreme scenarios caused by multiview changes of low-altitude remote sensing.

To solve the aforementioned multiview issues and provide a reliable registration for subsequent applications, we propose a novel mismatch removal approach using the Siamese network. The main contributions are listed as follows.

1) To solve the problem of unstable inlier matching, lack of generalization ability and difficulty in distinguishing a similar background, a neighbor-guided patch representation is designed for feature point descriptor, and includes the following steps:
   a) neighborhood relation reconstruction;
   b) patch selection; and
   c) enhancing similar patch differences.
2) To solve the problem of weights, model adaptability, nonadaptive optimization parameters, and local image information, this study proposes a Siamese network to measure similarity between the aforementioned patches for mismatch removal.
3) To solve the aforementioned multiview issues, we designed a rotation-invariant layer in the abovementioned learning framework.

III. Method

In this section, we provide an overview of the framework. Then, we elaborate the neighbor-guided match representation and rotation-invariant Siamese network and finally provide the implementation details and pseudocode.

A. Overview of the Framework

The proposed learning framework for similarity measurement is presented in Fig. 2. Two feature point sets $X_{N\times 2} = \{x_1, x_2,, \ldots, x_N\}^T$ and $Y_{N\times 2} = \{y_1, y_2,, \ldots, y_N\}^T$ are extracted from I_X and I_Y by SIFT and NNDR [9], respectively. Our purpose is to learn a similarity measurement network of a pair of feature point descriptors, which can distinguish inliers from outliers in a putative set.

Our framework involves two major steps: neighbor-guided patch representation and rotation-invariant Siamese network. In the neighbor-guided patch representation step, we aim to establish a characteristic representation for each feature point that can adequately represent it. In the training stage, we simulate the angle change caused by the flight attitude change of UAV, obtain the characteristic representation after the angle change of each feature point, and then use the expanded training samples to train the rotation-invariant Siamese network. In the testing stage, given a new image pair, we first extract a set of putative matches and construct their patch representations, and then use the learned network to measure similarity.

B. Neighbor-Guided Patch Representation

Constructing a proper match representation is the key to the success of our learning framework. To maintain the relatively complete texture structure of feature points, we use neighborhood relation reconstruction and patch selection to construct the robust patch representation for feature points. The details are given as follows.

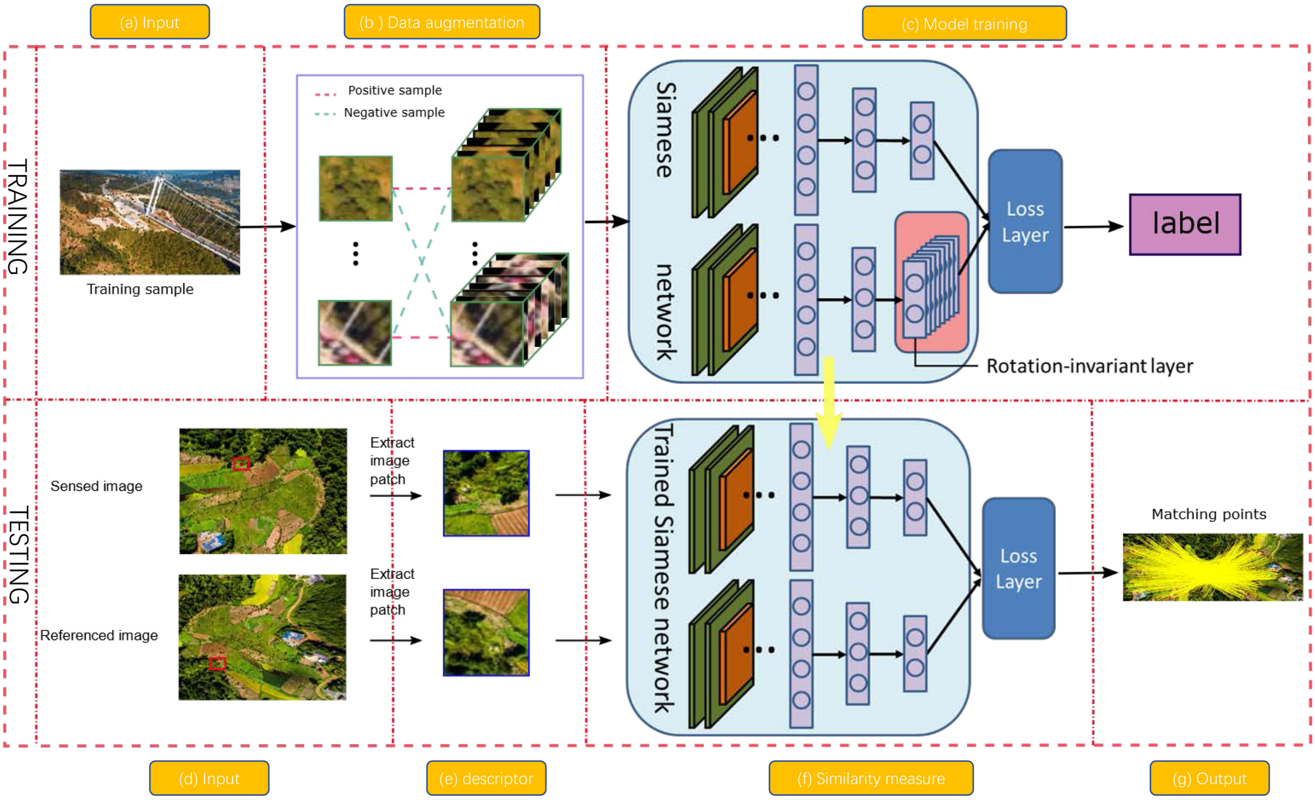

Fig. 2. Proposed framework has two stages: matching learning (upper half) and similarity measure (lower half). The learning part consists of two stages: data augmentation and Siamese model training. The first stage is to generate a set of enhanced training samples through a simple rotation operation. Then, we trained our Siamese architecture with rotation invariant on the VGG-16 model, by using the last fully connected layer of VGG-16 as the rotation invariant layer of branch 2. The measure part uses the parameters trained in the upper half to generate the matching probability of the image patch-pairs so as to get the matching result.

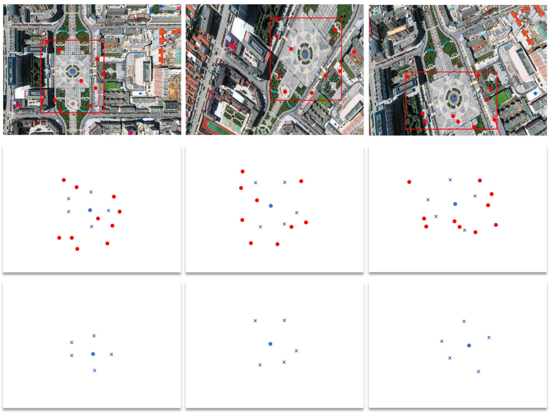

Fig. 3. Neighborhood relationship reconstruction. The first row shows the different degrees of rotation, scale, and distortion of the image. The second row shows the distribution of feature points and their neighboring points in different states. Among them, ● indicates the feature points that need to be judged, ● indicates outliers, and × indicates inliers, The third row indicates that the distribution of the neighboring points follows the neighborhood relation reconstruction.

Neighborhood relation reconstruction: In the case of the multiview changes, a large number of outliers exist, which can easily cause the local feature description to be unstable, as shown in Fig. 3. To obtain a stable description in an unstable space, we need to reconstruct the neighborhood. For each feature point, we center on it and use K-D tree [54] to search its nearest κ neighboring points. The key idea is, for each pair of points, if it is an inlier, then the distribution of corresponding neighboring points should be similar. On the contrary, if the pair is an outlier, then the distribution of corresponding neighboring points is different. To capture such a property mathematically, we apply two steps in reconstructing the neighborhood of the feature points.

1) Calculating the similarity of the neighborhood

$$r = \frac{c}{\kappa} \tag{1}$$

where $r \in [0, 1]$ represents the proportion of the common number of neighbors of the corresponding point, κ is the number of neighboring points to the feature point, and $c \leq \kappa$ is the number of common elements of two neighborhoods $\mathcal{N}_{x_i}$ and $\mathcal{N}_{y_i}$. Obviously, if the putative match is an inlier, then the value of r increases, and vice versa.

2) The neighboring points corresponding to the serial number are left from the neighboring points of each pair of feature points, and the ones that do not correspond are removed. The purpose of this method is to reduce the influence of mismatching on the reconstruction of feature point descriptors, similar to noise removal. Therefore, before extracting the pixel information, we first form a roughly judged inlier point set. Then, we select the nearest K neighboring points from $\mathcal{I}$ when looking for the

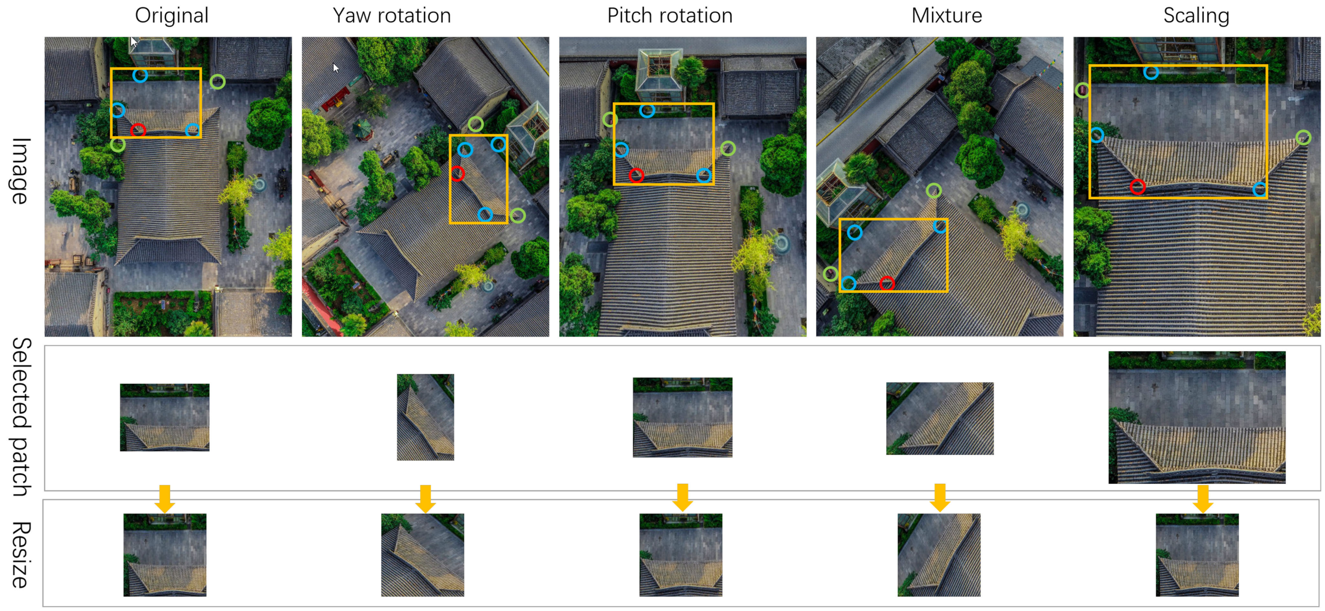

Fig. 4. Neighbor-guided region selection. The red points in the sensed and the reference image are a pair of putative corresponding feature points, and five neighbors are found for them, respectively. The blue point is the neighbor corresponding to the serial number, and the green point is the neighbor point not corresponding to the serial number. According to the blue point, the image patch is drawn as shown in yellow box. The second row is the selected patch in the yellow box in the first row, and the third row is the patch in the second row resize into 32. The figure shows the selection of patches when the images produce yaw rotation, pitch rotation, mixture, and scaling.

neighborhood of the center point and filter the noncorresponding neighbors simultaneously.

Patch selection: Selecting a fixed size of the region for each feature point is unreasonable because low-altitude remote-sensing images often have different scale structures, rotation, and deformation. However, the spatial neighborhood relation between feature points representing the local topological structure of the image scene is usually well preserved [39]. Therefore, we use the neighbor-guided feature points obtained in the previous step to delimit useful patch information. The farthest point is found in the upper, lower, left, and right directions to form the boundary of the image patch region. As shown in Fig. 4, when the image has undergone scale, rotation, and deformation, the image patch region selected according to the neighbor-guided feature points still contains similar information. Then, we resize all image patches to 32 × 32 pixels for training and learning. Compared to the conventional method of extracting fixed size, our methods contain more information of the same size.

Enhancing similar patch differences: In a pair of images, a many-to-one situation often occurs in the matching of the feature points due to the stretching or low overlap of one image. In our experiments, we found that the network has a weak ability to distinguish similar background patches, as shown in Fig. 5. $A1, A2$, and B form two sets of corresponding point pairs $\{A1, B\}, \{A2, B\}$, and the similarity measure of the network to these two similar point pairs is close, which in the case of many-to-one, causes multiple point pairs to be recognized as inliers by the neural network or all become considered as outliers, and does not select one of the most similar points as the inlier, and the rest are outliers. To solve the problem, we increase the discrimination for the image patches with the same background. Gaussian high-pass filter is added to the similar patches. Gaussian high-pass filter uses standard deviation to adjust the degree of filtering. We use the unique SIFT value of each feature point as its standard deviation, thereby increasing the similarity of outliers and inliers as follows:

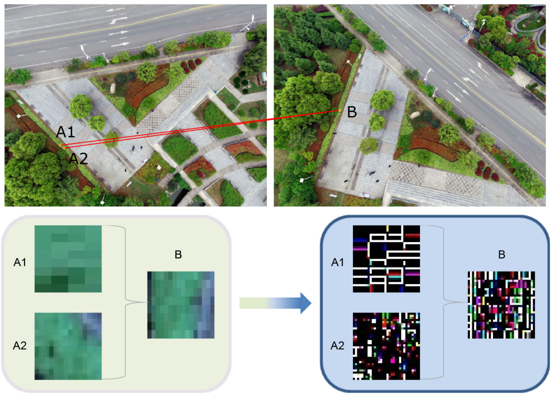

Fig. 5. Many-to-one of similar patches case. Point $A1$ and $A2$ are neighbor points and both corresponding to point B. The lower left part represents the patch region extracted according to the neighborhood, and the lower right part is obtained after Gaussian high-pass filtering.

$$I_\sigma = I * \left(1 - \frac{1}{2\pi\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}}\right) \tag{2}$$

where $*$ represents convolution operation, and to its right is the template for the Gaussian high-pass filter, where σ is standard deviation that varies according to the SIFT descriptor value of its point

$$\sigma = 1 - \frac{1}{D}\sum_{i=1}^{D} \frac{\min(\mathrm{DES}_i^x, \mathrm{DES}_i^y)}{\max(\mathrm{DES}_i^x, \mathrm{DES}_i^y)} \tag{3}$$

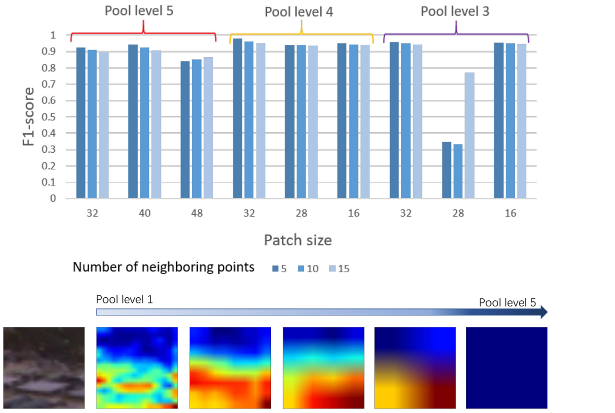

Fig. 6. Bar chart shows the results of $F1$ caused by different patch sizes and different numbers of neighboring points when using the third, fourth, and fifth pooling layers, respectively. The following figure shows the visualization results of the five pooling layers of VGG-16.

where D represents the number of SIFT descriptors, and DES represents the descriptor for each point.

C. Rotation-Invariant Siamese Network

Siamese network: In this study, a Siamese network is designed to process complex low-altitude remote-sensing images that lead to undesirable feature representation and matching. We use the network to judge whether a pair of feature points corresponds to each other according to the similarity of the neighbor-guided patch representation.

In the Siamese network, as the name implies, we use two identical models (e.g., VGG-16 [55]) to input the sensed image patch and the referenced image patch to facilitate the similarity calculation between the two patch representation. We select the fourth pooling layer of VGG-16 to connect with the full connection layer, as shown in Fig. 6. The shallow network extracts texture and detail features, and the deep network extracts contour, shape, and strongest features. The deeper the layers, the more representative the extracted features are.

This model consists mainly of the following two parts: 1) the two branches obtain their respective feature representations through a series of convolution operations, and share weights among the branches to reduce the parameters, and 2) a measurement component by combining the features of two branches. The aforementioned two parts are taken as a whole to implement supervised learning, and we presented this structure in the lower half of Fig. 2.

We assume that both I_X and I_Y have m pairs of putative feature points, then the hybrid feature descriptor patches of I_X and I_Y are denoted as $p^X = \{p_1^X, p_2^X, \ldots, p_m^X\}$ and $p^Y = \{p_1^Y, p_2^Y, \ldots, p_m^Y\}$. By combining the patch of image I_X and I_Y, we can acquire a series of patch pairs $\{(p_i^X, p_i^Y)\}$, where $i = 1, 2, \ldots, m$. We take each patch pair as input to the Siamese architecture, of which the output is the similarity between them. We define O_a and O_b as two branches of output features of the last fully connected layer. Mathematically, we use the following function to calculate the similarity

$$S = \max(0, (\|O_a(p_i^X) - O_b(p_i^Y)\|_2^2 - (\min + (\max - \min) \times \tau))). \tag{4}$$

Min and max represent the minimum and maximum values of Euclidean distance, and τ is a probability. S is a matrix and we find all indexes with $S = 0$. The feature points corresponding to them are inliers, and the rest are outliers.

Rotation-invariant layer: However, the CNN features show that the images still encounter difficulty in disposing the challenges of image rotation, which are crucial sources of detection error. In this situation, learning a more powerful feature representation with rotation insensitivity is highly desirable. In the training stage, inspired by [56], we modify one of the branches of the Siamese architecture slightly, changing the last full connection layer of VGG-16 to the rotation-invariant layer, and the other branch becomes invariant, with both branches sharing weights and biases. As shown in the upper half of Fig. 2, we refer to a rotation-invariant descriptor to enhance feature representation.

In the training step, we simulated the shooting angle of the UAV and defined the rotation operation of K angles in both yaw and pitch directions $T_\phi = \{T_1, T_2, \ldots, T_K\}$. For the input image I_t, K extended training samples of the input image $\{I_t^{T_1}, I_t^{T_2}, \ldots, I_t^{T_K}\}$ are obtained. When image I_t has n feature points $\{x_1, x_2, \ldots, x_n\}$, the feature points corresponding to $I_X^{T_k}$ are $\{x_1^{T_k}, x_2^{T_k}, \ldots, x_n^{T_k}\}$. Then, we can obtain matching point pairs $\{(x_i, x_j^{T_k}), i = j\}$ and nonmatching point pairs $\{(x_i, x_j^{T_k}), i \neq j\}$, where $k = 1, 2, \ldots, K, i$ and $j = 1, 2, \ldots, m$.

As mentioned above, the pixel information patch p_i can be selected by the neighbor field of the feature point x_i. Therefore, the corresponding transformed patch $\{p_1^{T_k}, p_2^{T_k}, \ldots, p_m^{T_k}\}$ can be obtained. Thus, we can also obtain matching patch pairs $\{(p_i, p_j^{T_k}), i = j\}$ and nonmatching patch pairs $\{(p_i, p_j^{T_k}), i \neq j\}$, and the training samples are $\{(p_i^X, p_j^{T_k}), y_{ij}^{T_k}\}$, where

$$y_{ij}^{T_k} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases} \tag{5}$$

Patch p_i in I_t has only one matched patch in the transformed image, and $n - 1$ nonmatched patches that lead to imbalances on both sides. Thus, we randomly selected only one nonmatched patch pairs to feed the network.

The last full connection layer of branch 1 and the rotation invariant layer of branch 2 also share the same weights and biases, but the difference between them is that the latter denotes the average features of all rotated image patches. The final loss layer, which computes the Euclidean distance between the features of the sensed and averaged patches, tries to minimize the distance for matching pairs to enforce sharing of the similar features and maximize it for nonmatching pairs. $O_a(p_i)$ represents the output features of layer FC of branch 1, and $O_b(p_j)$ is the output of the rotation-invariant layer of branch 2. Thus, the Euclidean distance between them is

$$D = \left\| O_a\left(p_i\right) - O_b\left(p_j^{T_k}\right) \right\|_2^2 \tag{6}$$

where

$$O_b\left(p_j^{T_k}\right) = \frac{1}{K}\sum_{k=1}^{K} O_b\left(p_j^{T_k}\right). \tag{7}$$

Inspired by [57], we modified contrastive loss function to more suitable for our method, which is defined as follows:

$$L = \frac{1}{2}\left(yD^2 + (1-y)\left(D-1\right)^2\right) \tag{8}$$

where y is a binary label, $y = 1$ when a pair of image patches match, and $y = 0$ when they do not. We hope that the similarity of the mismatched image patches approaches 1 and the similarity of the matched image patches approaches 0, so that we can distinguish the matched from the mismatched.

D. Implementation Details

In this study, we select the fourth pooling layer of the VGG-16 model and discard the convolution layer and fifth pooling layer after it. As mentioned earlier, to obtain the hybrid feature descriptor patch, we establish a set of neighborhoods, which is better if all the neighbor points are inlier. Therefore, we calculated the ratio of the common number of matching point pairs in the neighborhood under the condition of $\kappa = 5$ through (2), and only selected $r > 0.2$ to construct the neighborhood point set. In the training stage, we select three images with rich scene contents (including building, vegetation, lake, and others.) as the training set, and select the Adam optimizer as gradient descent method. The learning rate is $1\mathrm{e}-5$, batch size is 100, and rotation transformation $K = 8$. In the loss layer, we set $m = 10$ to increase the difference between matches and mismatches. Finally, in the test phase of the similarity function, we set $\tau = 0.9$. The pseudocode is provided in Algorithm 1.

TABLE I
QUANTITATIVE COMPARISON WITH OR WITHOUT ROTATION-INVARIANT LAYER

Category	Rotation-invariant Layer	Recall	Precision	F1
Yaw	Add	**0.96 ± 0.04**	**0.98 ± 0.02**	**0.97 ± 0.02**
	Not	0.56 ± 0.09	0.99 ± 0.01	0.71 ± 0.08
Pitch	Add	**0.99 ± 0.01**	**0.98 ± 0.02**	**0.98 ± 0.01**
	Not	0.78 ± 0.24	0.99 ± 0.01	0.85 ± 0.19
Mixture	Add	**0.99 ± 0.01**	**0.98 ± 0.01**	**0.99 ± 0.01**
	Not	0.86 ± 0.09	0.99 ± 0.01	0.92 ± 0.05

Algorithm 1: Rotation-Invariant Siamese Network for Image Registration.

```
input: Training data $S$, testing data: the sensed image $I_X$
    and reference image $I_Y$
output: optimized correspondence $\{x_i, y_i\}$
Training stage:
Extract feature points on $S$;
Construct the neighbor-guided patch representation for
    each feature point;
Construct each training sample with one image patch
    and K rotation transformation;
Training similarity measure function using a
    supervised learning technique;
Testing stage:
Extract putative match point on testing data;
Construct the neighbor-guided patch representation for
    each feature point;
Measure similarity between image patch pairs by
    Siamese architecture;
Obtain the optimized correspondence from the
    similarity matrix.
```

IV. Experimental Result

The experiments are divided into the following two categories: algorithm contribution test and quantitative comparison. In the first category, we test the effect of rotation-invariant layer and Gaussian high-pass filter in the algorithm. In the second category, we evaluate the performance of our algorithm against six state-of-the-art methods: GLPM [40], LPM [39], ICF [20], GS [58], LMR [52], and SIR [41] in the comprehensive rotation and extreme cases for feature matching and registration. The experiments are conducted on a desktop with 3.0 GHz Intel Core i5-8500 CPU, 16-GB RAM, and MATLAB code.

Our experiments are conducted on a publicly available small UAV image registration dataset (SUIRD) dataset,[1] which is provided for image registration/matching research. The SUIRD_v1.1 includes 50 pairs of images and their groundtruth (each pair contains 274–2385 pairs of feature points). These image pairs contain viewpoint changes in yaw, pitch, their mixture, and extreme patterns, which produce problems of low overlap, image distortion, and severe outliers. The open-source VLFeat toolbox is used to extract SIFT putative matches and K nearest neighbors are recovered using K-D tree.

We use precision, recall, and $F1$ score for quantitative comparisons. Precision represents the ratio of the true inlier to the assumed inlier of the algorithm, recall represents the ratio of the true inlier to the total inlier in the original putative set, and $F1$ score characterizes the matching performance defined as the harmonic mean of precision and recall [59].

A. Results of Test on Rotation-Invariant Layer

In the first series of experiments, we randomly select five images in each of the three categories of yaw, pitch, and mixture, and a total of 15 images to test the effect.

The ground truth is established by manually examining each putative match in each image pair, and we make the benchmark before conducting experiments to ensure its objectivity. Precision, recall, and $F1$ score are used to evaluate the performance of algorithms with and without the rotation invariant layer. Table I lists the quantitative comparison on the 15 pair of images from UAV image and 720 cloud data set. Without rotation-invariant layer tend to have higher precision, but the sacrifice of most of the inliers make a lower recall. The comprehensive evaluation

[1] [Online]. Available: https://github.com/yyangynu/SUIRD

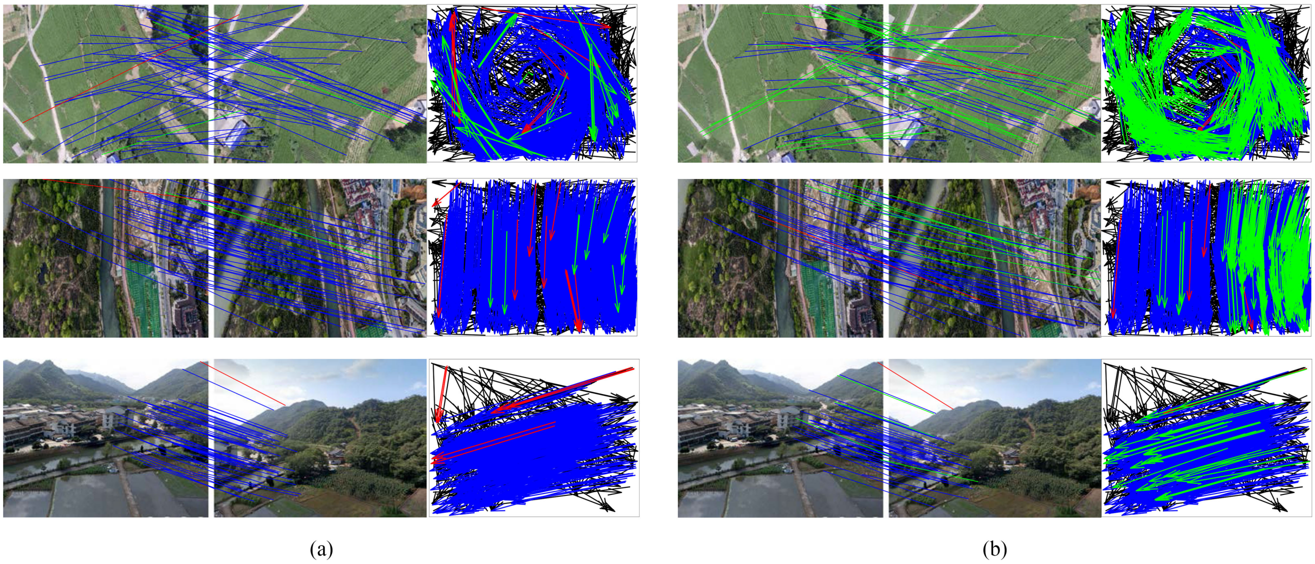

Fig. 7. Representative results of feature matching on three typical image pairs (blue = true positive, black = true negative, green = false negative, red = false positive). The left side represents the results of the algorithm in the case of yaw, pitch, and mixed rotation after the rotation-invariant layer is added, and the right side represents the results of the algorithm in the case of three scenarios without the rotation-invariant layer. For visual convenience of image pairs, at most 50 randomly selected matches are drawn.

$F1$ is far lower than with a rotation-invariant layer. Experiments show that after add a rotation-invariant layer, the algorithm greatly improves the performance of yaw rotation, pitch rotation, and mixture. The representative matching results are shown in Fig. 7.

B. Results of Test on Gaussian High-Pass Filter

In the second series of experiments, we test the effect of Gaussian high-pass filter. We use the same dataset as in the previous experiment.

Hundreds of many-to-one situations in an image on average, whereas dozens of many-to-one situations in neighboring points. As they have similar backgrounds, we add Gaussian high-pass filter to increase their discrimination. As shown in Fig. 8, when the Gaussian high-pass filter is added, the algorithm retrieves more inliers of neighboring points to improve the recall.

C. Results of Comprehensive Quantitative Comparison

In the third series of experiments, we conducted quantitative comparison with six state-of-the-art algorithms: GLPM [40], LPM [39], ICF [20], GS [58], LMR [52], and SIR [41]. Then, 27 pairs of yaw, pitch, and mixture scenarios in the dataset are used for testing. Precision, recall, and $F1$ score are used to evaluate the performance and summarized in Table II.

Table II shows that ICF and GS usually have high precision in all three scenarios, but did not perform well in recall and $F1$ score. They strictly eliminated most of the points and kept only a few, so that the precision was very high and the recall rate was low, which would make the transformed image deviate from the reference image during registration, as shown in Fig. 10. LPM is a neighboring preserving method that aims to find all interior points as far as possible, resulting in extremely high recall. However, LPM cannot distinguish the outliers that are similar to inliers, thereby resulting in lower precision. Similarly, extremely low precision results in unsatisfactory registration effect. GLPM is the advanced version of LPM, in which the matching result on a small putative set with a high inlier ratio guides the matching on a large putative set to obtain a stable local structure. SIR is a method that enriches the inlier pool through continuous iteration. This method can be taken as a progressively generalized version of the GLPM. LMR is a method that formulates the mismatch removal into a two-class classification problem. LMR, GLPM, and SIR perform better on the three scenarios than the other three methods, but degraded slightly in the mixture scenario. Furthermore, due to its requirement on the number of neighboring points, LMR cannot adapt to the feature matching of image pairs with insufficient feature points. As shown in Fig. 10, SIR performs well in registration, whereas relatively low recall leads to extremely few feature points in some regions to completely overlap. Our algorithm exhibits the most stable performance in all scenarios and outperforms the other six methods because the proposed neighbor-guided patch representation and rotation-invariant layer are more robust to various rotations. To more intuitively present the effect of our algorithm, we show some representative examples of feature matching and registration, respectively, in Figs. 9 and 10.

D. Results of Quantitative Comparison of Extreme Scenarios

In the fourth series of experiments, considering the diversity of UAV images, we add an extreme scenario, in which not only rotation occurred, but also low overlap, distortion, and scaling are combined, thereby increasing the difficulty of registration. We conducted quantitative comparison of extreme scenarios with six state-of-the-art algorithms: GLPM [40], LPM [39], ICF [20], GS [58], LMR [52], SIR [41]. Then, 23 pairs of images on extreme scenarios were used for testing. Precision, recall, and $F1$ score were used to evaluate the performance and the results are summarized in Table III.

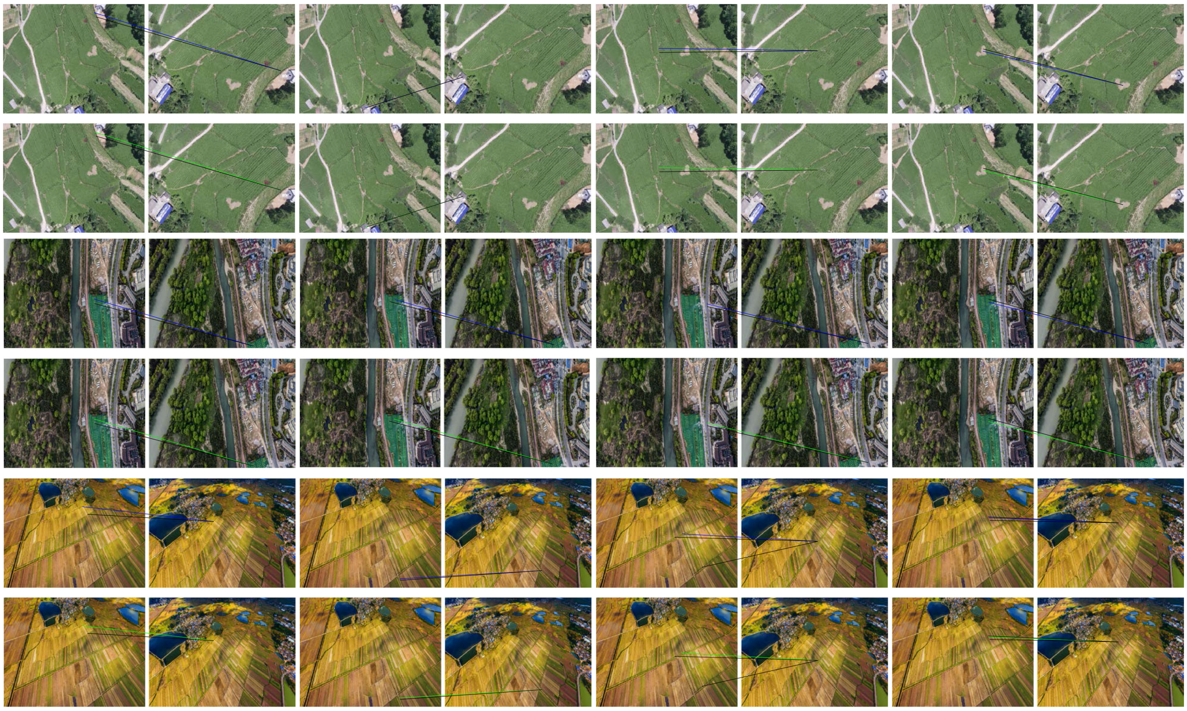

Fig. 8. Representative results with and with out Gaussian high-pass filter on three typical image pairs (blue = true positive, black = true negative, green = false negative). For visual convenience, only one many-to-one neighboring point is shown for each pair of images. In each case, the first rows represents the result with the Gaussian high-pass filter, and the second rows represents the result without the Gaussian high-pass filter.

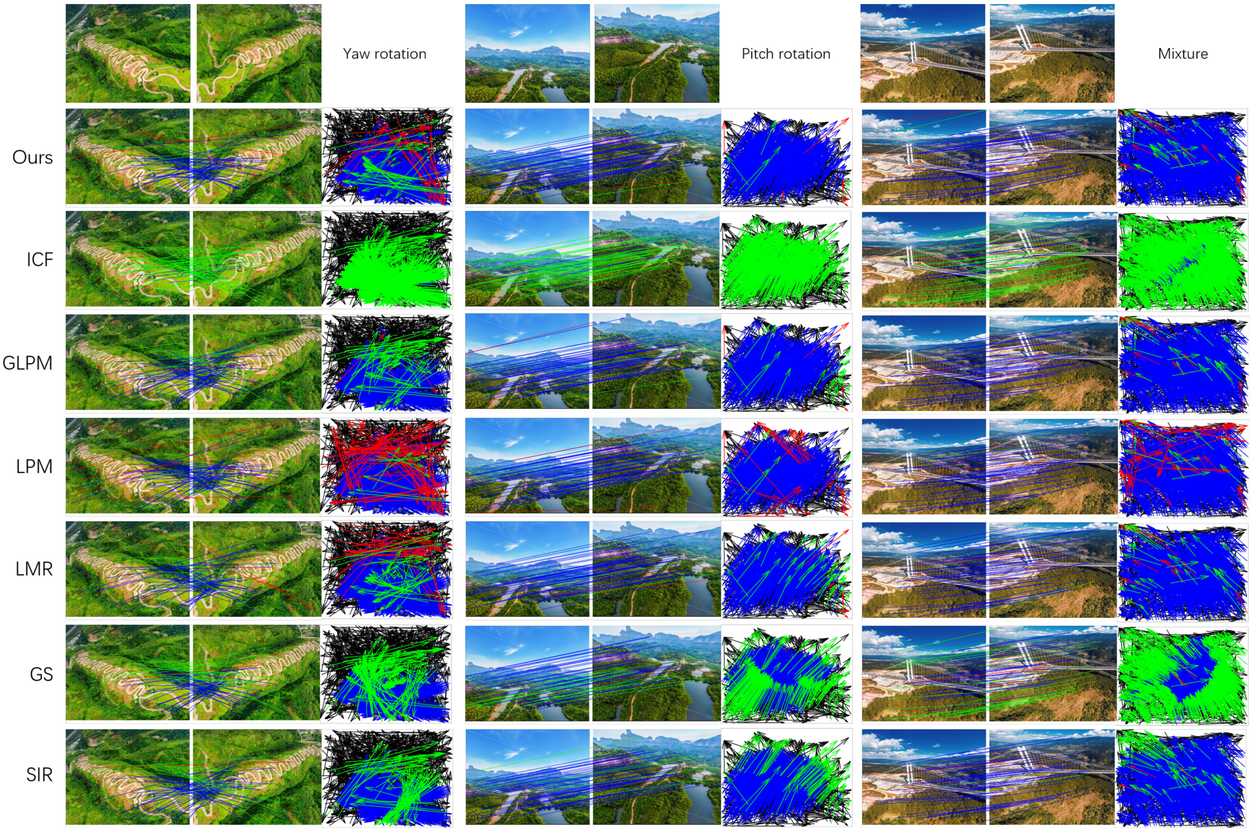

Fig. 9. Representative results of feature matching on three typical scenarios (blue = true positive, black = true negative, green = false negative, red = false positive). For visual convenience of image pairs, at most 50 randomly selected matches are drawn.

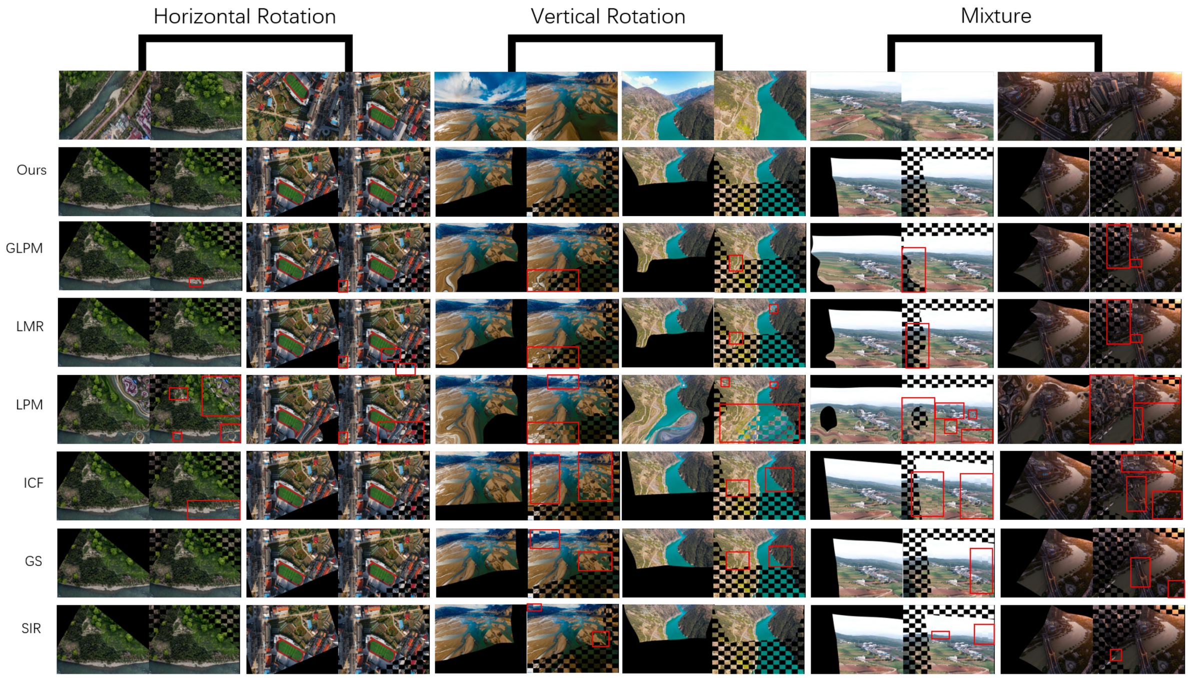

Fig. 10. Representative image registrations of the seven methods on six UAV image pairs in yaw rotation, pitch rotation, and mixture scenarios. Red rectangles indicate the misalignments.

TABLE II
Quantitative Comparison With GLPM [40], LPM [39], ICF [20], GS [58], LMR [52], SIR [41] on Three Scenarios of Experimental Data

Cat	Met	Rec	Pre	F1
Yaw	Ours	**0.99** ± **0.01**	**0.94** ± **0.11**	**0.96** ± **0.07**
	LPM	0.91 ± 0.11	0.94 ± 0.07	0.93 ± 0.08
	GLPM	0.95 ± 0.12	0.96 ± 0.07	0.95 ± 0.08
	ICF	0.97 ± 0.08	0.42 ± 0.26	0.53 ± 0.19
	GS	0.96 ± 0.12	0.83 ± 0.12	0.81 ± 0.13
	LMR	0.95 ± 0.12	0.95 ± 0.07	0.94 ± 0.08
	SIR	0.96 ± 0.12	0.95 ± 0.08	0.94 ± 0.08
Pitch	Ours	**0.97** ± **0.02**	**0.98** ± **0.01**	**0.98** ± **0.01**
	LPM	0.99 ± 0.01	0.94 ± 0.03	0.97 ± 0.01
	GLPM	0.98 ± 0.02	0.99 ± 0.01	0.98 ± 0.01
	ICF	0.15 ± 0.08	0.99 ± 0.02	0.25 ± 0.11
	GS	0.65 ± 0.12	0.99 ± 0.01	0.78 ± 0.09
	LMR	0.97 ± 0.02	0.98 ± 0.01	0.98 ± 0.01
	SIR	0.93 ± 0.05	0.99 ± 0.01	0.96 ± 0.03
Mix	Ours	**0.99** ± **0.01**	**0.93** ± **0.09**	**0.96** ± **0.06**
	LPM	0.99 ± 0.01	0.88 ± 0.13	0.93 ± 0.09
	GLPM	0.92 ± 0.20	0.95 ± 0.07	0.93 ± 0.16
	ICF	0.27 ± 0.21	0.96 ± 0.07	0.38 ± 0.21
	GS	0.70 ± 0.13	0.96 ± 0.01	0.79 ± 0.09
	LMR	0.95 ± 0.07	0.95 ± 0.09	0.95 ± 0.08
	SIR	0.85 ± 0.29	0.97 ± 0.06	0.88 ± 0.27

TABLE III
Quantitative Comparison With GLPM [40], LPM [39], ICF [20], GS [58], LMR [52], SIR [41] on Extreme Scenarios

Met	Rec	Pre	F1
Ours	**0.97** ± **0.06**	**0.89** ± **0.17**	**0.92** ± **0.13**
LPM	0.98 ± 0.05	0.82 ± 0.20	0.88 ± 0.16
GLPM	0.89 ± 0.19	0.93 ± 0.13	0.89 ± 0.17
ICF	0.36 ± 0.28	0.90 ± 0.22	0.42 ± 0.21
GS	0.72 ± 0.09	0.93 ± 0.15	0.80 ± 0.10
LMR	0.91 ± 0.13	0.91 ± 0.15	0.91 ± 0.14
SIR	0.89 ± 0.12	0.89 ± 0.13	0.91 ± 0.12

According to the standard deviation, we can observe that all the algorithms exhibit large fluctuations in the performance of these 11 pairs of images, and each algorithm performs well for specific problems, whereas most of the algorithms show a decline in the performance when confronted with comprehensive extreme scenes. Specifically, when yaw rotation and distortion occur at the same time, the performance of GLPM, LPM, LMR, and SIR are all greatly reduced, of which SIR is obvious, with its $F1$ 10% lower than ours. GLPM is weak on yaw rotation and low overlap natural landscape images, and its $F1$ is 30% lower than ours. LMR is weak in yaw, pitch, and severely distorted scenarios, and its $F1$ is 12% lower than ours. Our algorithm shows the most stable performance and outperforms the other six methods. To more intuitively present the performance of our algorithm, we show some representative examples of feature matching and registration, respectively in Figs. 11 and 12.

E. Results of Image Registration

Finally, we randomly select ten pairs of images from the entire data set to compare the registration accuracy. We follow the same evaluation in [60] and [61]: the root mean square error (RMSE), maximum error (MAE), and median error (MEE) are used for measuring the accuracy of image registration. Their definitions are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{n=1}^{M}(a_n - b_n)^2}$$

$$\text{MAE} = \max\left\{\sqrt{(a_n - b_n)^2}\right\}_{n=1}^{M}$$

$$\text{MEE} = \text{median}\left\{\sqrt{(a_n - b_n)^2}\right\}_{n=1}^{M} \tag{9}$$

where a_n and b_n are the corresponding landmarks of the sensed images and the reference images, respectively, M represents the number of selected landmarks, max, and median return

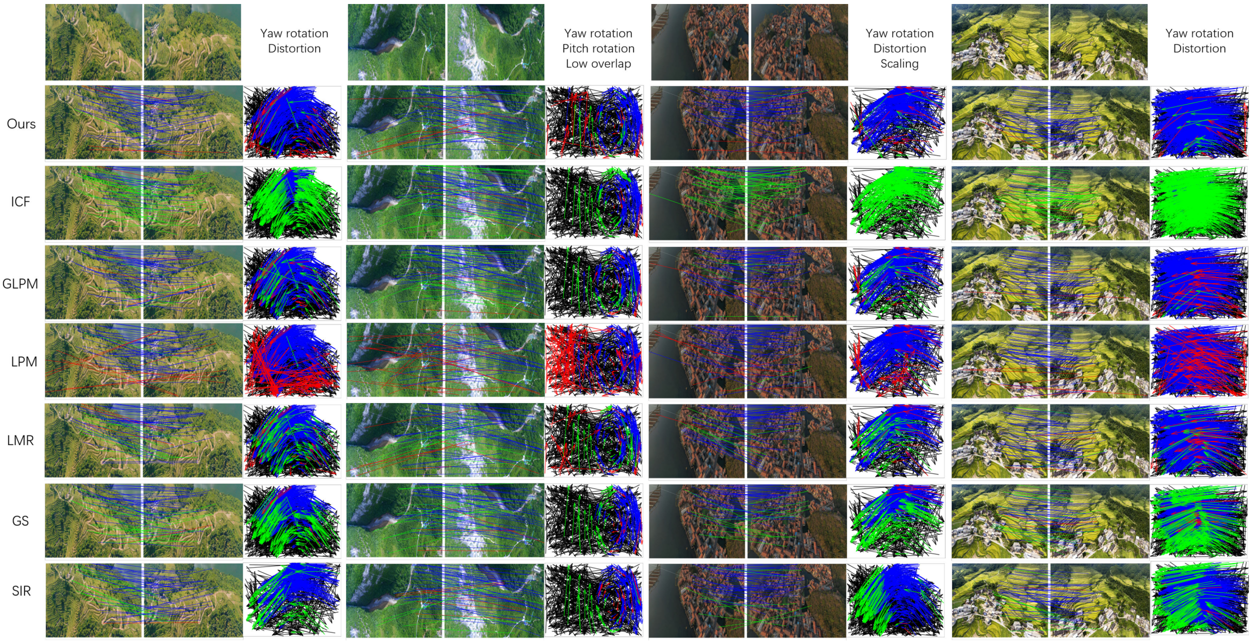

Fig. 11. Representative results of feature matching on extreme scenarios (blue = true positive, black = true negative, green = false negative, red = false positive). For visual convenience of image pairs, at most 50 randomly selected matches are drawn.

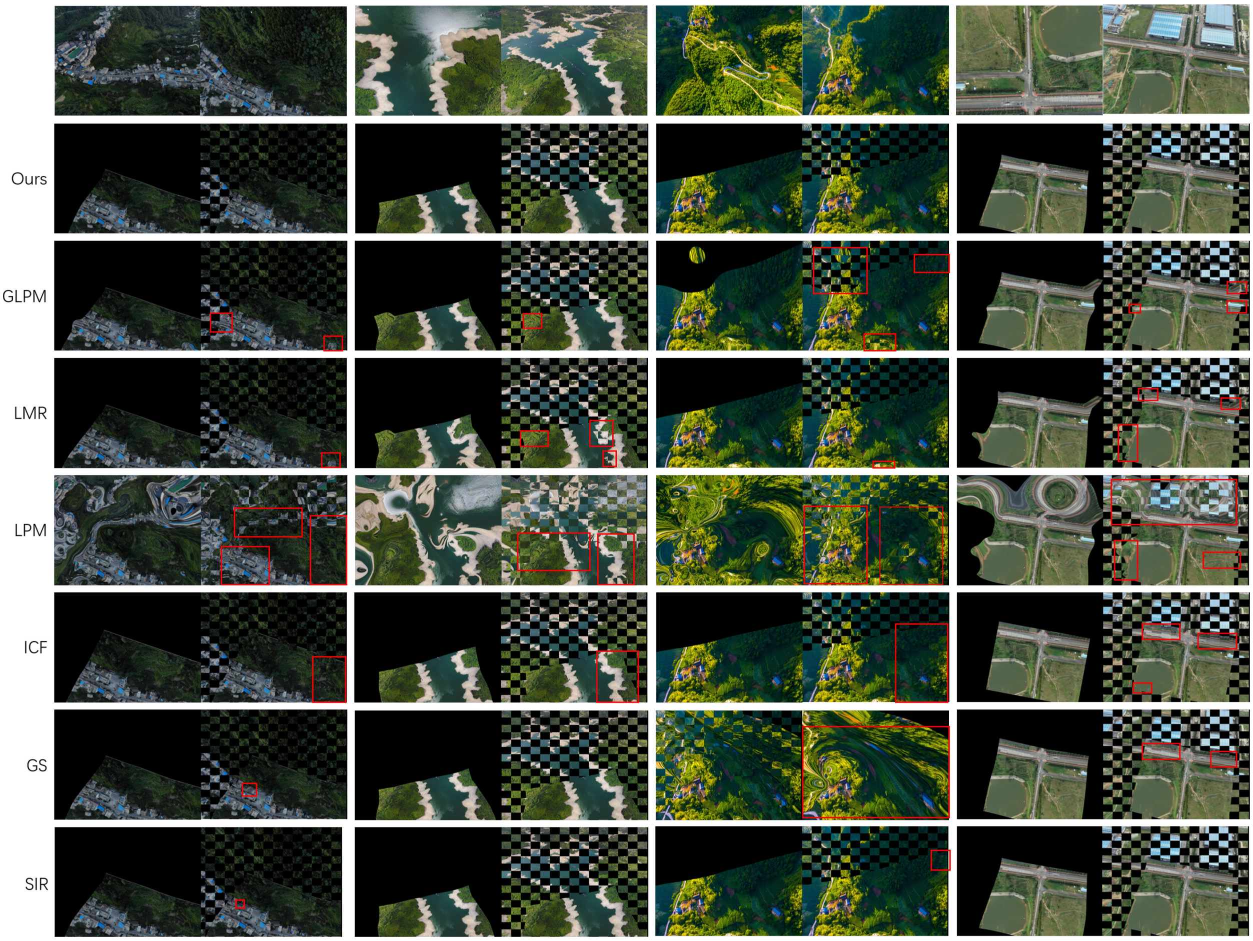

Fig. 12. Representative image registrations of the seven methods on four UAV image pairs in extreme scenarios. Red rectangles indicate the misalignments.

TABLE IV
REGISTRATION COMPARISON WITH GLPM [40], LPM [39], ICF [20], GS [58], LMR [52], SIR [41] ON ALL SCENARIOS

Met	RMSE	MAE	MEE
SIR	1.96 ± 0.55	5.77 ± 1.08	2.30 ± 0.94
LPM	2.21 ± 0.58	5.69 ± 0.90	2.68 ± 1.14
GLPM	2.01 ± 0.59	5.89 ± 1.14	2.30 ± 0.88
ICF	43.43 ± 103.20	488.23 ± 158.39	3.83 ± 140.95
GS	6.61 ± 9.05	70.13 ± 26.98	2.42 ± 0.88
LMR	1.97 ± 0.56	5.49 ± 1.01	2.30 ± 0.94
Ours	**1.81 ± 0.40**	**4.82 ± 0.89**	**1.94 ± 0.76**

the maximal and median of a set, respectively. The results are summarized in Table IV.

V. CONCLUSION

In this study, we have introduced a network composed of Siamese architecture to deal with the complex characteristics of low-altitude remote-sensing images. A neighbor-guided patch representation and rotation-invariant layer enable our algorithm to effectively address yaw rotation, pitch rotation, mixture, and extreme scenarios. Experimental results show that our method provides the most stable performance and outperforms the six state-of-the-art methods on feature matching and registration. However, the running time of our algorithm, at approximately 12 s for an image, is slightly longer than that of other algorithms, probably because our input is image patch, which consumes more memory than vector input (e.g., LMR). In future work, we will focus on solving the issue of time consumption.

ACKNOWLEDGMENT

The authors would like to thank D. G. Lowe, B. Jian, Y. Yang, J. Ma, H. Liu, S. Yan, X. Li, Z. Hu, B. C. Vemuri, and A. Rangarajan for providing their implementation source codes and experimental datasets, which facilitate the comparison experiments greatly.

REFERENCES

[1] S. Candiago, F. Remondino, M. De Giglio, M. Dubbini, and M. Gattelli, "Evaluating multispectral images and vegetation indices for precision farming applications from UAV images," *Remote Sens.*, vol. 7, no. 4, pp. 4026–4047, 2015.

[2] H. Kim, L. Mokdad, and J. Ben-Othman, "Designing UAV surveillance frameworks for smart city and extensive ocean with differential perspectives," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 98–104, Apr. 2018.

[3] C. Yuan, Y. Zhang, and Z. Liu, "A survey on technologies for automatic forest fire monitoring, detection, and fighting using unmanned aerial vehicles and remote sensing techniques," *Can. J. Forest Res.*, vol. 45, no. 7, pp. 783–792, 2015.

[4] R. Qin, "An object-based hierarchical method for change detection using unmanned aerial vehicle images," *Remote Sens.*, vol. 6, no. 9, pp. 7911–7932, 2014.

[5] F. Song, M. Li, Y. Yang, K. Yang, X. Gao, and T. Dan, "Small UAV based multi-viewpoint image registration for monitoring cultivated land changes in mountainous terrain," *Int. J. Remote Sens.*, vol. 39, no. 21, pp. 7201–7224, 2018.

[6] William K. Pratt, *Digital Image Processing*. Hoboken, NJ, USA: Wiley, 1978.

[7] R. N. Bracewell and R. N. Bracewell, *The Fourier Transform and Its Applications*, vol. 31999. New York, NY, USA: McGraw-Hill, 1986.

[8] 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 Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 603–615, Jan. 2014.

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

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

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

[12] P. Schwind, S. Suri, P. Reinartz, and A. Siebert, "Applicability of the sift operator to geometric SAR image registration," *Int. J. Remote Sens.*, vol. 31, no. 8, pp. 1959–1980, 2010.

[13] B. Kupfer, N. S. Netanyahu, and I. Shimshoni, "An efficient sift-based mode-seeking algorithm for sub-pixel registration of remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 379–383, Feb. 2015.

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

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

[16] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Comput. Vis. Image Understanding*, vol. 89, no. 2/3, pp. 114–141, 2003.

[17] A. Rangarajan, H. Chui, and F. L. Bookstein, "The softassign procrustes matching algorithm," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.*, 1997, pp. 29–42.

[18] A. L. Yuille, "Generalized deformable models, statistical physics, and matching problems," *Neural Comput.*, vol. 2, no. 1, pp. 1–24, 1990.

[19] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.

[20] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, 2010.

[21] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, Dec. 2010.

[22] B. Jian and B. C. Vemuri, "Robust point set registration using Gaussian mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1633–1645, Aug. 2011.

[23] Y. Yang, S. Ong, and K. Foong, "A robust global and local mixture distance based non-rigid point set registration," *Pattern Recognit.*, vol. 48, pp. 156–173, Jan. 2015.

[24] S. Zhang, Y. Yang, K. Yang, Y. Luo, and S.-H. Ong, "Point set registration with global-local correspondence and transformation estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2669–2677.

[25] K. Yang, A. Pan, Y. Yang, S. Zhang, S. H. Ong, and H. Tang, "Remote sensing image registration using multiple image features," *Remote Sens.*, vol. 9, no. 6, 2017, Art no. 581.

[26] V. Golyanik, S. Aziz Ali, and D. Stricker, "Gravitational approach for point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5802–5810.

[27] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

[28] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

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

[30] P. H. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[31] O. Chum and J. Matas, "Matching with PROSAC-progressive sample consensus," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 220–226.

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

[33] T. Kim and Y. Im, "Automatic satellite image registration by combination of matching and random sample consensus," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1111–1117, May 2003.

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

[35] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. 10th IEEE Int. Conf. Comput. Vis. Volume 1*, 2005, vol. 2, pp. 1482–1489.

[36] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 596–609.

[37] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1609–1616.

[38] F. Zhou and F. De la Torre, "Deformable graph matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2922–2929.

[39] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, 2019.

[40] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, Aug. 2018.

[41] S. Zhang, W. Zhao, X. Hao, Y. Yang, and C. Guan, "A context-aware locality measure for inlier pool enrichment in stepwise image registration," *IEEE Trans. Image Process.*, vol. 29, pp. 4281–4295, 2020.

[42] X. Jiang, J. Ma, J. Jiang, and X. Guo, "Robust feature matching using spatial clustering with heavy outliers," *IEEE Trans. Image Process.*, vol. 29, pp. 736–746, 2020.

[43] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 118–126.

[44] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6645–6649.

[45] O. Russakovsky, *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[46] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[47] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[48] G. J. Scott, M. R. England, W. A. Starms, R. A. Marcum, and C. H. Davis, "Training deep convolutional neural networks for land–cover classification of high-resolution imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 549–553, Apr. 2017.

[49] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 144.

[50] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4353–4361.

[51] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1505–1516, Jul. 2016.

[52] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4045–4059, Aug. 2019.

[53] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, and L. Jiao, "A deep learning framework for remote sensing image registration," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 148–164, 2018.

[54] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–14.

[56] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[57] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1735–1742.

[58] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1609–1616.

[59] Y. Lin, Z. Lin, and H. Zha, "The shape interaction matrix-based affine invariant mismatch removal for partial-duplicate image search," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 561–573, Feb. 2017.

[60] W. Zhao, X. Ma, L. Liang, L. Liang, and S. H. Ong, "Remote sensing image registration based on dynamic threshold calculation strategy and multiple-feature distance fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 4049–4061, Oct. 2019.

[61] Z. Yang, Y. Yang, K. Yang, and Z. Wei, "Non-rigid image registration with dynamic Gaussian component density and space curvature preservation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2584–2598, May 2019.

Yuyan Liu is currently working toward the B.S. degree in network engineering with the School of Computer Science and Information Engineering, Shanghai University of Applied Technology, Kunming, China.

Her research interest includes computer vision and remote sensing image processing.

Xiaoying Gong is currently working toward the B.S. degree in software engineering with the School of Computer Science and Technology, Southwest University of Science and Technology, Kunming, China.

His research interest covers computer vision and image processing.

Jiaxuan Chen is currently working toward the B.S. degree in information management and information system with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

His research interest covers computer vision and image processing.

Shuang Chen is currently working toward the B.S. degree in information management and information system with the School of Mathematics and Information, Xihua Normal University, Kunming, China.

Her research interest includes computer vision and image registration.

Yang Yang (Member, IEEE) received the master's degree from Waseda University, Tokyo, Japan, in 2007, and the Ph.D. degree from National University of Singapore, Singapore, in 2013, both in computer science.

He is currently a Professor with the School of Information Science and Technology, Yunnan Normal University, Kunming, China. His research interest covers computer vision, remote sensing, geography information system, and medical imaging.