Robust Local Structure Visualization for Remote Sensing Image Registration

Jiaxuan Chen, Shuang Chen, Yuyan Liu, Xiaoxian Chen, Yang Yang, *Member, IEEE*, and Yungang Zhang

***Abstract*—Image registration is a fundamental and important task in remote sensing. In this article, we focus on feature-based image registration. Existing attempts often require estimating a transformation model or imposing relaxed geometric constraints to establish reliable feature correspondences. However, a parametric model cannot handle image pairs undergoing complex transformations, and relaxed methods discard a lot of structure information and the results are often coarse. To solve the above issues, we propose a local structure visualization descriptor to preserve the original structure information, and cast the feature matching task into an evaluation of the consensus of visual structure under a convolutional neural network. This strategy can effectively measure the similarity of neighborhood structure for mismatch removal. In summary, our method does not depend on a specific transformation model and can process arbitrary remote sensing images (e.g., different deformations, severe outliers, various rotations, and scaling changes). To demonstrate the robustness of our strategy for image registration, extensive experiments on various real remote sensing images for feature matching are conducted and compared against nine state-of-the-art methods, where our method gives the best performances in most scenarios.**

***Index Terms*—Feature matching, image registration, mismatch removal, remote sensing, visualization descriptor.**

I. Introduction

IMAGE registration is an important fundamental research in computer vision and pattern recognition, and it works as an essential image preprocessing step for many remote sensing tasks, such as image mosaic, image fusion, remote sensing monitoring, and image analysis. The purpose of image registration is to find an optimal alignment between the sensed image and the reference image. In real applications, image pairs to be registered are normally captured from different viewpoints, different sensors, or different times and inevitably include the following issues: image distortion, low overlap, scaling, rotation, multimodality, or their mixtures. Fig. 1 gives some registering examples for ground monitoring by small unmanned aerial vehicle (UAV). There are two common approaches used in image registration: 1) area-based and 2) feature-based methods [1]. Area-based methods directly manipulate image intensity values, and their subclasses in a broad sense include correlation-like methods [2], Fourier methods [3], mutual information methods [4], and convolutional neural network (CNN)-based methods [5], [6]. Remarkably, Fourier-based image correlation is a specific type of area-based technique, which has experienced rapid development, especially in the field of photogrammetry and remote sensing. Cross correlation in the frequency domain (CCF) and the phase correlation (PC) are two primary types of correlation forms. CCF tends to be sensitive to the intensity variations since it does not involve any type of normalization. PC weakens the dependence on image intensity and content by using the phase information solely and, thus, invariant to global linear variations in contrast and brightness. In the case of subpixel shifts, the signal power in the PC is concentrated in several coherent peaks and the most outstanding ones largely adjacent to each other, which implies that PC leads to a downsampled 2-D Dirichlet kernel [7]. To solve the above problem, various subpixel Fourier-based image correlation methods (e.g., peak centroid [8], upsampling [9], plane fitting [10], and line fitting [11]) have been proposed over the years. However, area-based methods are sensitive to image distortion, low overlap, intensity change, and training samples [1], [12], [13].

Feature-based methods, which overcome above defects [14], are, in general, more robust and have been widely used in

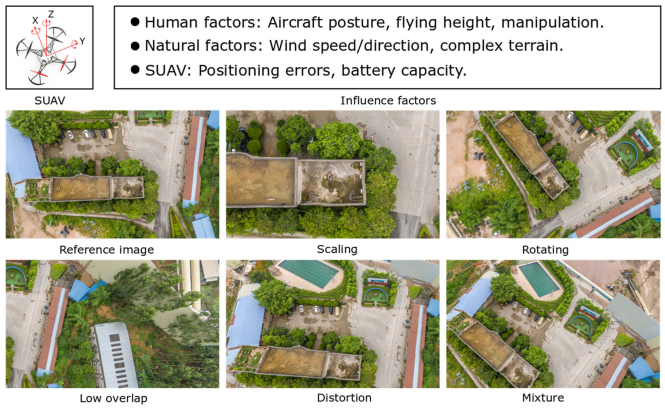


Fig. 1. Registering examples for ground monitoring by small UAV. The above inevitable factors will produce image distortion, low overlap, scaling change, various rotations, and their mixtures between images captured by small UAV.

Manuscript received September 15, 2020; revised December 7, 2020; accepted January 6, 2021. Date of publication January 11, 2021; date of current version January 28, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41971392 and in part by the Yunnan Ten-thousand Talents Program. *(Jiaxuan Chen and Shuang Chen contributed equally to this work.) (Corresponding author: Yang Yang and Yungang Zhang.)*

Jiaxuan Chen, Shuang Chen, Yuyan Liu, Yang Yang, and Yungang Zhang are with the School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China (e-mail: jrbook_chen@foxmail.com; chen_shuang283@163.com; yuyan_l19@163.com; yyang_ynu@163.com; yungang.zhang@ynnu.edu.cn).

Xiaoxian Chen is with the Laboratory of Computer Networks and Information Security, China Agricultural University, Beijing 100089, China (e-mail: xxianchen@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3050459

TABLE I
RELATED WORK IN FEATURE-BASED METHODS

	Method	Designed Algorithm		Contribution
		Correspondence Estimation	Transformation Update	
Point Set Registration	TPS-RPM [15]	fuzzy correspondence	TPS/Annealing	**D/O/N**
	CPD [16]	GMM	MC constraint/EM	**D/O/N**
	GMMREG [17]	GMM	L2 distance/EM	**D/O/N/M**
	GLMDTPS [18]	global and local mixture distance	TPS/Annealing	**D/O/N**
	RPM-L2E [19]	L2E estimator	QNM/Annealing	**O/N**
	PR-GLS [20]	GMM	global and local structures/EM	**R/O**
	DSMM [21]	SMM	dirichlet distribution/EM	**O**
	VBPSM [22]	hierarchical graphical model	six groups of hyperparameters/VB/KD	**O/M/LM**
	GL-CATE [23]	mixture-feature GMM	global and local structures/EM	**O/IP**
	Yang et al. [1]	DGCD	dynamic threshold/SSCP	**IM/O**
	Zhou et al. [24]	TLMM	transformation adjustment/VB	**O/M/R**
		Additional Descriptor	Matching Algorithm	
Mismatch Removal	RANSAC [25]	null	random sample consensus	**MR**
	LLT [13]	GMM	LG constraint	**MR**
	LFGC [26]	eight-point algorithm	multi-Layer perceptron	**MR**
	GMS [27]	the number of local matches	grid-based motion statistics	**MR**
	LMR [28]	neighborhood relationship	BP network	**MR**
	GS [29]	graph	replicator equation	**MR**
	LPM [30]	neighborhood relationship	locality preserving matching	**MR**
	GLPM [31]	neighborhood relationship	guided matching strategy	**MR**
	SIR [32]	context-aware locality measure	stepwise registration strategy	**MR**
	RFM-SCAN [33]	K-dist	DBSCAN	**MR**
	Ours	local structure visualization	convolutional neural network	**MR**

Their contributions mainly involve solving the following problems: **O** (outliers), **D** (deformation), **R** (rotation); **M** (missing or occlusion), **N** (noise), **LM** (local minima), **IP** (ill-posed problem), **IM** (inliers maximization), **MR** (mismatch removal). Other abbreviations are: FM: feature matching; TPS: thin plate spline; MC: motion coherent; GMM: Gaussian mixture model; EM: expectation-maximization; QNM: quasi-Newton method; LG: local geometrical; SMM: student-t mixture model; VB: variational Bayesian; KD: Kullback–Leibler divergence; DGCD: dynamic Gaussian component density; SSCP: spatial structure and curvature preservation; TLMM: student-t latent mixture model.

remote sensing applications. In this article, we mainly focus on feature-based methods. The procedure of feature-based method for image registration can be summarized as follows: 1) detecting and extracting features points from the reference image and sensed image; 2) seeking reliable correspondence (i.e., feature matching); and 3) estimating geometric transform matrix for registration according to the matching results. The point set registration and mismatch removal are the two main techniques for seeking reliable correspondence. Obviously, the registration result depends on the matching performance. In order to introduce current representative methods and compare them more clearly, their designed algorithms and contributions are listed in Table I and are discussed as follows.

The key of feature-based methods is to recover point-to-point correspondences between an image pair. To this end, it is necessary to construct two sets of feature points from two images. Fortunately, various well-designed feature detection algorithms have been developed, such as scale-invariant feature transform (SIFT) [34] and speeded up robust features (SURF) [35]. After obtaining the reliable correspondences, the transformation parameters between two images can be calculated accordingly with a predefined transformation model. In fact, the image registration reduces to a feature matching problem [31].

Point set registration is the process of finding one-to-one correspondence of two point sets, which includes two main types: noniterative methods and iterative methods. Noniterative methods are difficult to accomplish a good matching result under a single estimation for large nonrigid transformations [18]. Iterative methods are designed to perform alternating two 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 becomes easier. Normally, iterative registration methods need a robust estimation model (e.g., probability

model [16], [17], [19], [23], [24]) and reliable transformation update (e.g., global-local structure constraint [23], expectation-maximum [16], [17], [20], [21], and variational Bayesian inference [22], [24]). However, the weights between different features, model adaptability, and nonadaptive optimization parameters are very sensitive to different registration patterns in real applications. And they ignore the local structure information among feature point sets. Therefore, their performance very likely degrades in complicated registration patterns.

Mismatch removal involves two steps: 1) computing a set of putative matches, which is considered an easy mission; and 2) removing the outliers, which employ one or more additional descriptors to further estimate inliers and outliers (i.e., identify mismatches) based on a prematching result. Mismatch removal can be roughly divided into two categories: parametric methods and nonparametric methods. The best-known parametric methods are random sample consensus (RANSAC) [25] and its variants. However, when the underlying image transformation is nonrigid, parametric methods [25], [26] become less efficient due to the dependence on a predefined parametric model. Nonparametric methods are usually suitable for both rigid and nonrigid transformation, and the recent trend has been toward developing relaxed methods [28], [30], [31], [33] in exchange for generality. In other words, the geometric constraint is made less strict to accommodate complex matching patterns [36]. Many deep-learning-based approaches have made dramatic progress on computer vision tasks, such as keypoint detection and feature description [37], stereo matching [38], [39], and image patch matching [40], which motivates us to leverage deep learning techniques to eliminate mismatches [26], [28]. However, remote sensing images often involve local distortions due to ground relief variations or viewpoint changes, resulting in complex spatial relationships between image pairs. Existing methods do not preserve or exploit complete local structure information, which usually results in the inability to capture complex feature interactions.

In order to solve the above problems in feature-based image registration/matching, from a novel perspective, we cast the mismatch removal into a consistency evaluation of visual structure topology. First, complete local structure information is mapped to a 2-D grid and then discard redundant structure information through convolutional layers. Finally, the extracted feature vectors are classified with a fully connected network. The main contributions are listed as follows.

1) To capture complete structure topology information, we design a local structure visualization (LSV) descriptor, which maps the spatial distribution of feature points to a regular grid. Such grid data can be efficiently evaluated for similarity through CNN.
2) To screen out neighborhood mismatches, we introduce a simple way to further enhance the representation ability of LSV; we term this process as vortex-field-guided structural deformation (VFGSD). After VFGSD processing, LSV can effectively capture spatial dislocation information.
3) Our method only needs a small number of training samples to achieve satisfactory performance because the LSV descriptor is independent of the original image content.

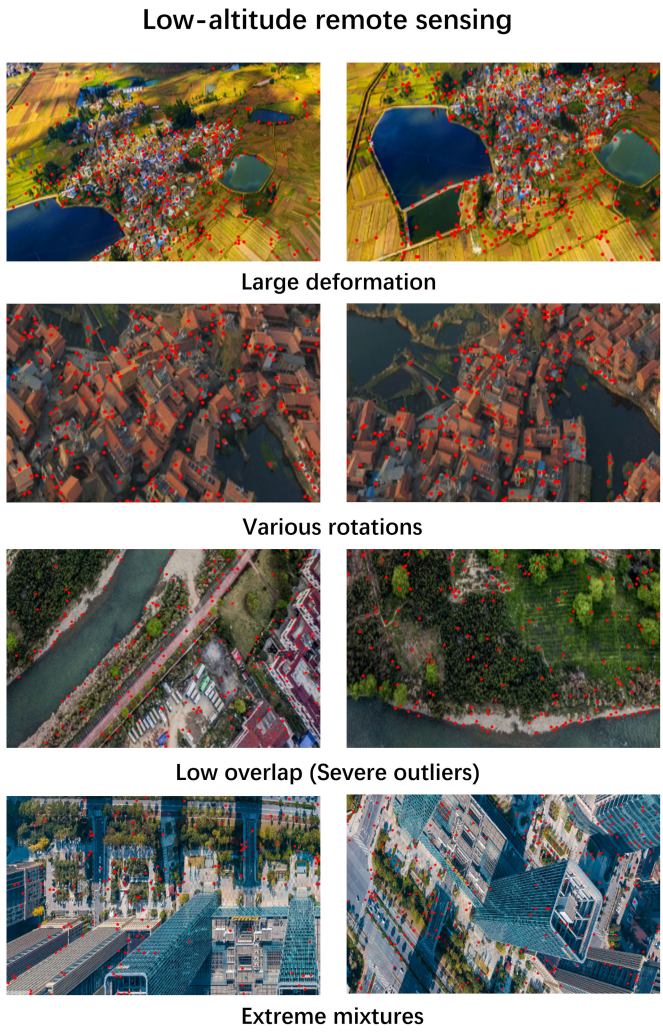


Fig. 2. Feature point set registration/matching examples. • denotes feature points extracted by SIFT.

II. Methodology

In order to establish accurate correspondences between two feature point sets extracted, respectively, from two images, we focus on the construction of LSV descriptors and formulate it as a two-class classification problem. In other words, the confidence of putative correspondences is transformed into visual similarity evaluation via CNN.

A. Overview of the Proposed Method

First, we present Fig. 2 to show large deformation, various rotations, severe outliers, and their mixtures between two feature point sets in low-altitude remote sensing images. Fortunately, several off-the-shelf feature descriptors (e.g., SIFT and SURF) can efficiently establish putative matches, which consider all possible matches between two feature sets. It is advisable to

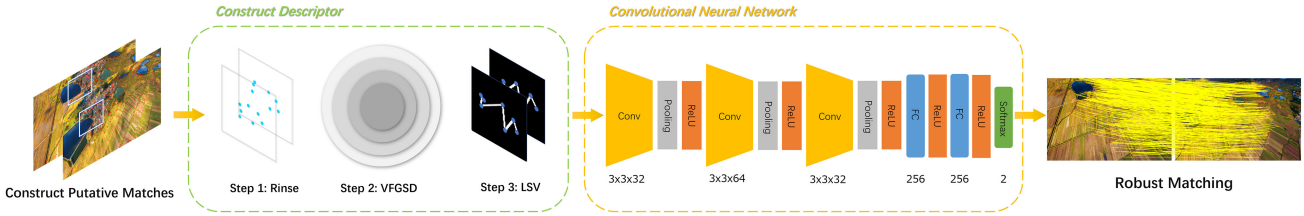


Fig. 3. Proposed learning framework. First, constructing a set of putative matches and searching the k nearest neighbors for putative match (x_i, y_i). Then, rinse neighborhood points and use vortex field to guide the local structure deformation. Subsequently visualize the local structure. Finally, a well-trained CNN produces matching results.

filter out sufficiently different matches of the feature descriptor vector.

Suppose a set of putative matches $S = \{(x_i, y_i)\}_{i=1}^{n}$ is extracted from the sensed image I^s and the reference image I^r by SIFT. There are a number of mismatches in S due to various nonrigid transformations. Our goal is to eliminate mismatches in S. Due to the physical constraints in a small region around a point, most feature points would keep the distribution of their neighboring point pairs after transformation [41], the key of which is to recognition the local neighborhood structures of those potential inliers. Human vision can easily identify local neighborhood structures of those potential inliers. Thus, we leverage LSV and CNN to simulate this recognition process.

There is no obvious local structure definition for a point set. Searching the k nearest neighbors for each point in the corresponding feature set under the Euclidean distance is our strategy, i.e., for each putative match (x_i, y_i), $N_{x_i}^k$ and $N_{y_i}^k$ ($k = 20$ in this article) are used to build LSV_{x_i} and LSV_{y_i} descriptors. Confidence probability of (x_i, y_i) is formulated as

$$\text{MCP}_i = \mathcal{CNN}(\text{LSV}_{x_i}, \text{LSV}_{y_i}) \tag{1}$$

where $\mathcal{CNN}(\cdot)$ denotes the trained convolutional neural network. The cross-entropy loss function L is used to train the network, and it is defined as

$$L = -\frac{1}{n}\sum_{i=1}^{n}[t_i \cdot \log(\text{MCP}_i) + (1 - t_i) \cdot \log(1 - \text{MCP}_i)] \tag{2}$$

where $t_i \in \{0, 1\}$ represents the correctness of putative match (x_i, y_i). Specifically, $t_i = 1$ indicates an inlier, and an outlier otherwise.

From what has been discussed above, the framework of our proposed method consists of the following three steps, as shown in Fig. 3.

1) *Image feature extraction:* An image pair (the sensed image I^s and the reference image I^r) is given, and a set of putative matches $S = \{(x_i, y_i)\}_{i=1}^{n}$ is then calculated using the SIFT algorithm, where x_i and y_i denote the spatial position of putative match.
2) *Local structural visualization:* It consists of mapping the local structure of putative match (x_i, y_i) to a 2-D grid by the midpoint Bresenham algorithm [42].
3) *Matching using CNN:* LSV_{x_i} and LSV_{y_i} are combined to a series of two-channel images and given to a modified LeNet-5 network [43]. Finally, more accurate matches between I^s and I^r are determined.

B. Local Structure Visualization

The key of our method is the LSV descriptor, which is constructed by the following three steps.

Step 1: Reconstruct a reliable neighborhood relation by rinse of neighborhood points for rejecting most outliers.

Step 2: VFGSD is designed to further solve the neighborhood mismatches, which can be considered as a special outlier problem.

Step 3: Retain topology information by LSV, which can solve other remaining problems such as deformations, rotations, and scales.

1) Rinse of Neighborhood Points: In complex nonrigid transformations, $N_{x_i}^k$ and $N_{y_i}^k$ are unstable. In other words, many $x_j \in N_{x_i}^k$ cannot find their corresponding points in $N_{y_i}^k$. This kind of noncorrespondence in a fixed neighborhood is usually a mismatch because of disobedience to the smoothness of motion. Therefore, removing noncorresponding points can reject most outliers and improve the stability of neighborhood relation. Toward this goal, we use a simple set operations to rinse such noncorresponding points in local region:

$$\text{RLS}_i = S \cap (N_{x_i}^k \times N_{y_i}^k) \tag{3}$$

$$\text{RLS}_{x_i} = \{x_j | (x_j, y_j) \in \text{RLS}_i\} \tag{4}$$

$$\text{RLS}_{y_i} = \{y_j | (x_j, y_j) \in \text{RLS}_i\} \tag{5}$$

where $\times$ is the Cartesian product, and RLS_{x_i} and RLS_{y_i} represent a more reliable neighborhood relation for x_i and y_i, respectively.

The neighborhood relations RLS_{x_i} and RLS_{y_i} have the following properties: for a putative match (x_i, y_i), if it is an inlier, they will have a similar neighborhood feature point structure. Conversely, neighborhood points of an outlier will be different in their spatial structures. To enhance the visual effect, we apply a simple way to represent the local structure for x_i and y_i, as shown in Fig. 4, which is simply to connect selected neighbors (RLS_{x_i} and RLS_{y_i}) in turn, and these neighbors are not necessary to be sorted. However, the jth connected points in neighbors of x_i and y_i are corresponding to each other. It can be seen from Fig. 4 that most common mismatches in putative matches $S = \{(x_i, y_i)\}_{i=1}^{n}$ have relatively large and different local structures or nonstructures [see Fig. 4(d) and (e)] and can

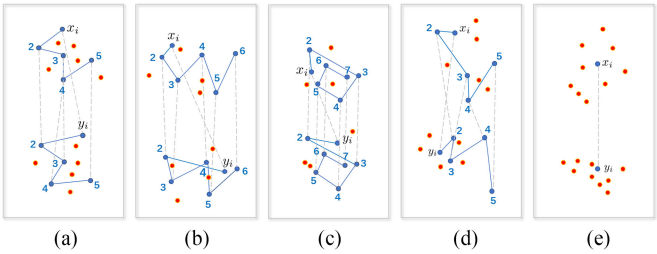


Fig. 4. Neighborhood structures of inliers and outliers. ● denotes noncorresponding point. (a) Inlier. (b) and (c) Special outliers (neighborhood mismatches). (d) and (e) Common outliers.

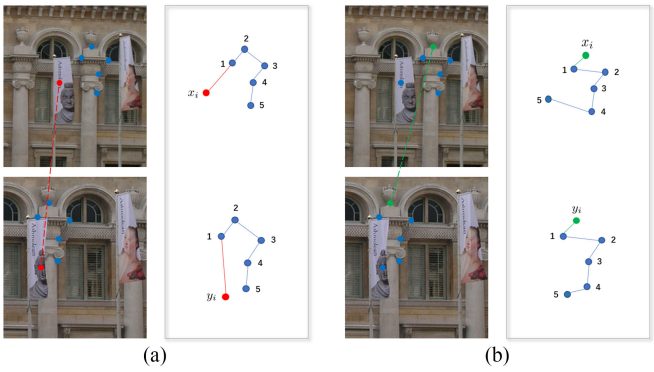


Fig. 5. Neighborhood mismatch problem. In (a), local regions of outlier (●) have very-similar-structure-based feature points and the tiny differences appear in the first connection. In (b), inlier (●) also have similar structures, but the tiny differences appear in the last connection.

be easily removed by comparing the structure similarity between RLS_{x_i} and RLS_{y_i}.

2) Vortex-Field-Guided Structural Deformation: After the above local structure reconstruction, there are still some special mismatches. Such mismatches generally have very similar local structures [see Fig. 4(b) and (c)], which easily lead to a neighborhood mismatch by using current similarity measurement approaches. Even trickier, both inliers and outliers may have this tiny differences between two local structures. This issue is one of the most challenging tasks in feature representation and matching, as shown in Fig. 5.

In neighborhood mismatches, x_i and y_i have a very similar neighborhood structure since a more reliable neighborhood relation for x_i and y_i is reconstructed, and elements in RLS_{x_i} and RLS_{y_i} correspond to each other. Based on our observation, similar neighborhood structures means that most of the neighborhood elements are inliers, and the tiny differences are always caused by individual outliers. Therefore, if putative match is an outlier, the relative position between putative match and most neighborhood features (inliers) is quite different; conversely, if putative match is an inlier, this difference of relative position only exists in a fraction of the features (outliers). For instance, in Fig. 5(a), the sequence of neighboring elements of x_i from near to far is $[1, 5, 4, 2, 3]$, while that of y_i is $[5, 4, 1, 3, 2]$, i.e., none of the numbers match. On the contrary, in Fig. 5(b), the sequence of neighboring elements of x_i from near to far is $[1, 2, 3, 4, 5]$, and that of y_i is also $[1, 2, 3, 4, 5]$, i.e., all of the serial numbers match.

In order to enhance the tiny differences in these special mismatches [see Fig. 5(a)] and maintain the structural similarity of inlier pairs [see Fig. 5(b)]. We propose a VFGSD to solve the above problems so that the similarity measurement can easily identify these neighborhood mismatches. Simultaneously, the similarity measurement of inlier is not affected. To guide this structural deformation, each point $x_j \in \mathrm{RLS}_{x_i}$ can be transformed by

$$\tilde{x}_j = (1-\alpha_i^j)\begin{pmatrix} I_{2\times 2} & \alpha_i^j x_i \\ 0 & 1 \end{pmatrix}\begin{pmatrix} R_i^j(\theta_{ij}) & \mathbf{O}_{2\times 1} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_j \\ 1 \end{pmatrix} \tag{6}$$

where α_i^j is the decay coefficient, I denotes an identity matrix, and $R_i^j(\theta_{ij}) \in \mathbb{R}^{2\times 2}$ represents the rotation matrix. The movement of each point consists of simple rotation and translation, and the degree of transformation is determined by its relative position to the center point. This difference in relative position is captured by α_i^j, which is formulated by

$$\alpha_i^j(\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\lambda_i^j)^2}{2\sigma^2}\right) \tag{7}$$

where λ_i^j is the normalized Euclidean distance between x_i and x_j. If (x_i, y_i) is an inlier, corresponding points have close decay coefficient due to similar spatial distribution. On the contrary, if (x_i, y_i) is neighborhood mismatch, the relative position difference between the neighborhood point and the center point will result in different decay coefficients. In our experiments, we empirically fixed σ as 0.5.

For $R_i^j(\theta_{ij})$, θ_{ij} can be defined as

$$\theta_{ij}(\sigma) = \alpha_i^j(\sigma)\|\mathcal{D}_{x_i} - \mathcal{D}_{y_i}\|_2 \tag{8}$$

where $\mathcal{D}_{x_i}$ and $\mathcal{D}_{y_i}$ represents standardized SIFT feature descriptors (128-D vector) of x_i and y_i, respectively. By applying the same transformation to $y_j \in \mathrm{RLS}_{y_i}$, $\tilde{y}_j$ can be obtained. Intuitively, for neighborhood mismatch, moving track of neighborhood points will be different due to the larger SIFT feature distance and spatial dislocation, thereby destroying its original structure. Conversely, the deformation of inlier tends to be consistent, as shown in Fig. 6. This design can effectively solve the most challenging neighborhood mismatch problem and decreases the difficulty of training the neural network.

3) Visualization of Local Structure: After the VFGSD, the new positions of RLS_{x_i} and RLS_{y_i} are updated, and the differences between neighborhood mismatches are enhanced. The LSV descriptor is visualized in two steps. It is shown in Fig. 7.

3) Adaptive coordinate system: The coordinates based on the original image are very sensitive to rigid deformation, especially the rotation problem. Thus, we use an adaptive coordinate system for angle compensation, e.g., for LSV_{x_i}: 1) the geometric center of RLS_{x_i} as the origin, 2) $\overrightarrow{o_{x_i}x_i}$ as the y-axis positive direction, where o_{x_i} is the geometric center, and, finally, 3) the direction of the x-axis are determined by the distance between the neighbor points and the y-axis; if the sum of the distances from the feature points that locate in the right region of the y-axis

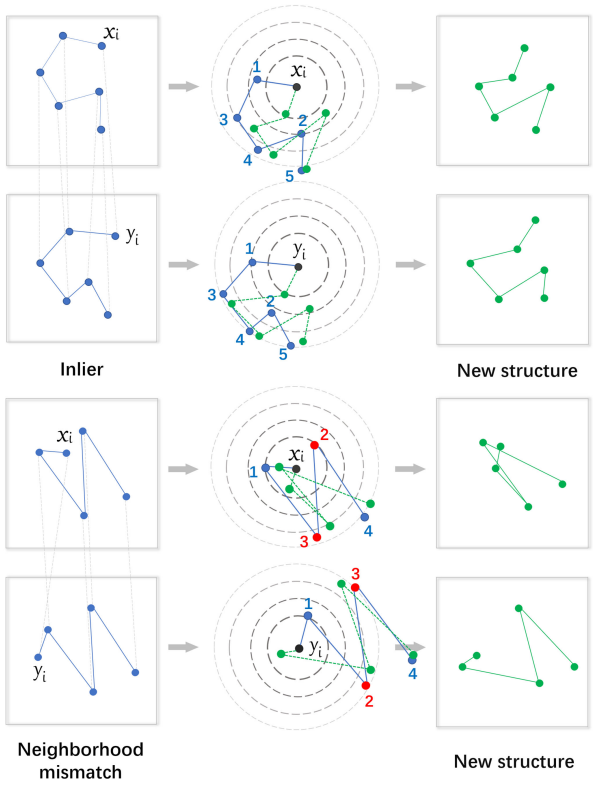


Fig. 6. Schematic illustration of the VFGSD. For each putative match $(x_i, y_i) \in S$, x_i and y_i are the centers of the vortex field. In the left group, the deformation of inlier tends to be consistent due to similar spatial distribution. In the right group, the visual similarity of outliers is destroyed. Note that the numbers are sorted according to the distance from the center point, and ● denotes the order change due to the misalignment of the center point.

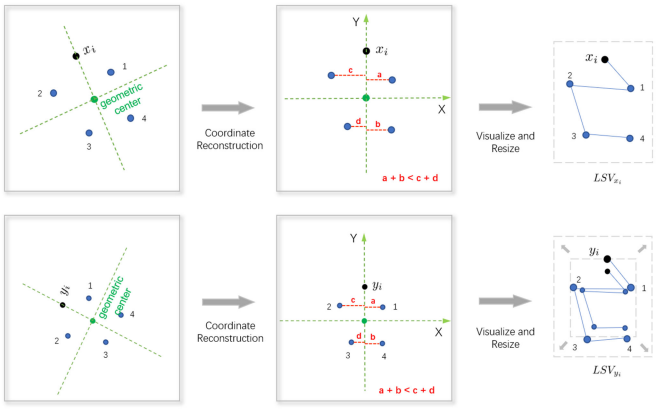


Fig. 7. Schematic illustration of the visualization of local structure. The first row represent visualization of RLS_{x_i}, where ● represent RLS_{x_i} and ● is used to represent the geometric centers of RLS_{x_i}. The second row represent visualization of RLS_{y_i}, where ● represent RLS_{y_i} and ● is used to represent the geometric centers of RLS_{y_i}.

to the y-axis is less than the left region, the right region is the positive direction of the x-axis; otherwise, the left region is the positive direction of the x-axis. Reconstruct the coordinates of LSV_{y_i} in the same way.

Fig. 8. Two image pairs used for training in our network, which contains in total 2072 SIFT putative matches (1349 inliers) as the training samples with 1349 positive samples and 1349 (including 626 misalignment samples) negative samples.

3) Visualize and resize: We initiate at x_i and y_i points to connect their neighbors in turn, respectively, and output the two images LSV_{x_i} and LSV_{y_i}. Due to scaling, the size of the LSV descriptor is not consistent; thus, all LSV descriptors are resized to the same size. To avoid floating-point calculations, we use the midpoint Bresenham algorithm [42] to draw the LSV. The basic principle of the midpoint Bresenham algorithm is to take one step in the main displacement direction each time, and the other direction depends on the value of the midpoint error term.

C. Training and Testing

There are several widely used network structures for image similarity measurement, such as siamese network [44], pseudo-siamese network, and triple network [45]. In this article, we choose the two-channel network as an instance. This network provides greater flexibility compared to the other models as it starts by processing the two patches jointly. Furthermore, it is fast to train [46].

The LSV descriptors (48×48 binary image) are used to represent the structural features of point x_i and y_i and are combined as a series of two-channel images (i.e., $\mathrm{LSV}_i \in \mathbb{Z}^{2\times48\times48}$) given into a two-channel modified LeNet-5 network for matching. This visualization of local structure and matching using a CNN can avoid the difficulties in similarity measurement by handcrafted features. For the two-channel network, specific architecture and parameters are as follows:

1) $(3 \times 3) \times 32$ convolutional layer with 0 padding, 2×2 max pooling, ReLU activation function;
2) $(3 \times 3) \times 64$ convolutional layer with 0 padding, 2×2 max pooling, ReLU activation function;
3) $(3 \times 3) \times 32$ convolutional layer with 0 padding, 2×2 max pooling, ReLU activation function;
4) for three full connection layers, the number of neurons was 256 (ReLU), 256 (ReLU), and 2 (Softmax), respectively.

Our training scheme does not depend on different image types since only neighborhood structures are used without original pixel information of the image. Therefore, a small training set (normally two to five image pairs) is enough to train our network, and negative samples (i.e., outliers) can be easily generated by a series of misalignments, such as (LSV_{x_1}, LSV_{y_2}) and (LSV_{x_1}, LSV_{y_3}). In the training phase, we selected two image pairs from the Small UAV Image Registration Dataset (SUIRD) as the training set (as shown in Fig. 8, from which we extract approximately 2000 SIFT putative matches as training samples) and used mini-batch stochastic gradient descent for optimizing cross entropy loss function (2). Specifically, the learning rate is 1e-3, the batch size is 50, and the max iteration is 5000.

Algorithm 1: Robust Local Structure Visualization.

Input: Training image pairs $\mathcal{S}$, testing image pairs $\mathcal{T}$
Output: Matching results on $\mathcal{T}$

Training phase:
Extract putative matches on $\mathcal{S}$;
Generate training samples using local structural visualization descriptors;
Training the network;
Testing phase:
Extract putative matches on $\mathcal{T}$;
Main process of LSV descriptor generation;
    Step 1 Rinse of neighborhood points;
    Step 2 Vortex field guided structural Deformation;
    Step 3 Visualization of local structure;
Matching with the network;

During the testing phase, we extract a set of putative matches S and construct a series of LSV descriptors. Then, we use the trained CNN to generate a output ($[\mathrm{MCP}_i, 1-\mathrm{MCP}_i]$) for $\mathrm{LSV}_i (i = 1, 2, \ldots, |S|)$. The MCP_i can be seen as the confidences of the putative match being an inlier. The final decision is whether the MCP_i is greater than 0.5. The pseudocode of the proposed method is outlined in Algorithm 1.

III. Experimental Results

In order to evaluate the performance of our LSV, we conduct experiments on feature matching for various real image pairs and apply it to the image registration. The open-source VLFeat toolbox [47] is employed to determine the putative correspondence of SIFT and to search the k nearest neighbors using K-D tree [48]. The experiments are conducted on a laptop with 2.80-GHz Intel Core i7-7700HQ CPU, 16-GB RAM, MATLAB, and C++ code. Nine representative algorithms are used for comparison, including GS [29], VFC [49], KVLD [50], MODS [51], GLPM [31], LPM [30], LMR [28], SIR [32], and LFGC [26], where all the competitors are implemented based on their publicly available codes and their own parameter settings.

A. Results on Feature Matching

In this section, we focus on establishing feature correspondences for real images. First, we further analyze the contribution of each component to LSV. Then, we test the robustness of our method to different viewpoint changes. Finally, we demonstrate the performance of LSV in low-inlier-ratio scenarios. Recall, precision and F1-score are employed as the criteria. To achieve a direct and fair comparison, we provide the experimental results in the following three datasets.

1) *SUIRD [36]:* The test dataset is provided for image registration/matching research. The SUIRD includes 60 pairs of images (800×600) and their ground truth (each pair contains 274–2385 pairs of feature points). These image pairs contain viewpoint changes in horizontal, vertical, mixture, and extreme patterns, which produce problems of severe outliers, illumination variations, various rotations, and image deformation.
2) *The Oxford Buildings Dataset (OBD) [52]:* The OBD[1] consists of 5062 images collected from Flickr,[2] which were collected by searching for specific Oxford landmarks. Image pairs taken under different extreme conditions (i.e., low inlier ratio) can be found in this dataset.
3) *The Mixture-Type Image Registration Dataset (MTIRD):* The dataset consists of small UAV image pairs, remote sensing image pairs, fingerprint image pairs, hyperspectral and visible image pairs, and multimodal MR image pairs that involve different image transformations, including affine, homography, nonrigid deformation, and light changes. This dataset was collected by us in order to comprehensively evaluate the image registration results of our LSV.

In all comparative experiments, the size of the images is 800×600, the SIFT feature points are extracted with default parameters, and the nearest neighbor distance ratio (NNDR) threshold for constructing putative matches is 0.9. These settings are consistent with the public dataset SUIRD.

1) Ablation Studies: First, we test the effect of number of common elements (i.e., the cardinality of RLS_i) without considering the spatial structure. For convenience, the ratio of common elements of neighborhood (RCN) can be used to calculate the similarity of the neighborhood; it is defined as

$$t = \frac{|\mathrm{RLS}_i|}{|N_{x_i}^k|}. \tag{9}$$

In order to determine a suitable threshold t, we compare two common prematching strategies: threshold-based matching and NNDR. Fig. 9 illustrates the linear separability of the three prematch methods. From the results, the RCN achieves the best performance at the threshold $t = 0.2$ (second coordinate system), the distance threshold of the SIFT descriptor is almost indivisible (third coordinate system), and the NNDR achieves better results when the threshold is equal to 0.7 (the last coordinate system). Clearly, the RCN manifests the optimal linear separability, and most prematches are outliers when $t < 0.2$.

Second, to determine the contribution of various components to LSV, we test the RCN (we choose the optimal threshold), LSV1 (without VFGSD), and LSV2 (add VFGSD) together. Randomly selected ten image pairs from SUIRD are used for the test, as shown in Fig. 10. Each group of results schematically shows the matching result, and motion field provides the decision correctness of each correspondence in the putative set. From the results, using only RCN produces satisfying results, whereas the LSV did even better; especially for LSV with VFGSD, it minimizes mismatches; this is in line with our expectations.

In order to provide a comprehensive quantitative evaluation, we selected all image pairs from extreme viewpoint changes (not only rotation occurred, but also low overlap, distortion, and scaling are mixed together) in SUIRD and randomly add one image pair from each of the other viewpoint changes

[1][Online]. Available: http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/
[2][Online]. Available: http://www.flickr.com/

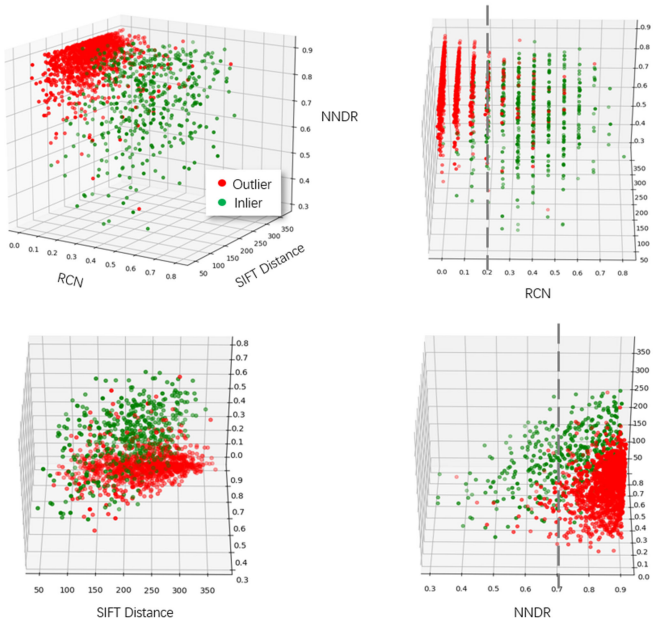


Fig. 9. Schematic illustration of RCN performance. Putative matches are calculated based on the minimum distance of the SIFT descriptor (1243 putative matches from three pairs of image).

TABLE II
RESULTS OF F1 CAUSED BY DIFFERENT NEIGHBORHOOD SIZES

Size	k = 10	k = 15	k = 20	k = 25	k = 30
F1-score	95.04%	97.04%	97.57%	97.24%	96.40%

(horizontal rotation, vertical rotation, scaling, and mixture). The contribution of various components on LSV is shown in Fig. 11. From the boxplot, the vortex field can improve precision without sacrificing recall. It is demonstrated that the effect of vortex field ensures that the inlier structure is not destroyed and removes the neighborhood mismatches at the same time. Additionally, there is a considerable performance amelioration for abnormal values.[3] Finally, we also provide the impact of different neighborhood sizes on LSV, summarized in Table II.

2) Robustness on Viewpoint Changes: Remote sensing images often involve local distortions caused by imaging viewpoint changes, resulting in complex spatial relationships between image pairs. Therefore, this experiment focuses on matching images that are obtained from the same scene in different views. We first conducted quantitatively comparison of normal viewpoint changes (horizontal rotation, vertical rotation, scaling, and mixture) with nine state-of-the-art algorithms. Recall, precision, and F1-score are summarized in Table III. In view of the result of Table III, LSV achieves optimal performance or comparable performance to the state of the art on all four viewpoint changes. GLPM and LMR are very close to our method; they achieve the same performance as ours on three viewpoint changes. SIR and

[3]Abnormal values are defined as value less than $Q_1 - 1.5 \times \text{IQR}$ or greater than $Q_3 + 1.5 \times \text{IQR}$, where Q_3 and Q_1 represent the upper and lower quartiles, respectively, and $\text{IQR} = Q_3 - Q_1$ denote interquartile range.

TABLE III
QUANTITATIVE COMPARISON ON NORMAL VIEWPOINT CHANGES

Viewpoint	Method	Recall	Precision	F1
HR	**Ours**	**0.99±0.01**	**0.98±0.01**	**0.99±0.01**
	SIR	0.96±0.05	0.99±0.00	0.98±0.02
	GLPM	0.98±0.03	0.99±0.01	0.98±0.01
	LMR	0.97±0.02	0.98±0.02	0.98±0.02
	LPM	1.00±0.00	0.94±0.04	0.96±0.02
	GS	0.75±0.16	0.99±0.01	0.84±0.11
	VFC	0.97±0.02	0.99±0.01	0.98±0.01
	LFGC	0.20±0.16	0.83±0.24	0.31±0.22
	MODS	0.96±0.03	0.99±0.01	0.97±0.02
	KVLD	0.96±0.02	0.99±0.01	0.97±0.01
VR	**Ours**	**0.99±0.01**	**0.98±0.01**	**0.98±0.01**
	SIR	0.94±0.05	0.99±0.00	0.96±0.03
	GLPM	0.98±0.02	0.99±0.01	0.98±0.01
	LMR	0.98±0.02	0.98±0.01	0.98±0.01
	LPM	1.00±0.00	0.94±0.02	0.97±0.01
	GS	0.65±0.11	0.99±0.01	0.78±0.09
	VFC	0.98±0.01	0.99±0.00	0.98±0.00
	LFGC	0.55±0.20	0.99±0.01	0.68±0.20
	MODS	0.65±0.42	0.87±0.33	0.69±0.39
	KVLD	0.95±0.03	1.00±0.00	0.97±0.02
S	**Ours**	**1.00±0.00**	**0.98±0.01**	**0.99±0.00**
	SIR	0.99±0.00	1.00±0.00	0.99±0.00
	GLPM	1.00±0.00	0.99±0.01	0.99±0.00
	LMR	0.99±0.00	0.99±0.01	0.99±0.00
	LPM	1.00±0.00	0.95±0.02	0.97±0.01
	GS	0.96±0.01	0.99±0.01	0.98±0.01
	VFC	0.99±0.00	0.99±0.00	0.99±0.00
	LFGC	0.69±0.25	1.00±0.00	0.78±0.21
	MODS	0.70±0.36	1.00±0.00	0.75±0.37
	KVLD	0.98±0.00	1.00±0.00	0.99±0.00
M	**Ours**	**0.99±0.01**	**0.97±0.02**	**0.98±0.01**
	SIR	0.95±0.05	0.99±0.01	0.97±0.03
	GLPM	0.99±0.01	0.98±0.02	0.98±0.01
	LMR	0.98±0.01	0.97±0.02	0.98±0.01
	LPM	1.00±0.00	0.91±0.05	0.95±0.03
	GS	0.68±0.12	0.99±0.01	0.80±0.08
	VFC	0.98±0.01	0.98±0.01	0.98±0.01
	LFGC	0.35±0.31	0.84±0.29	0.45±0.31
	MODS	0.60±0.40	0.99±0.01	0.65±0.39
	KVLD	0.85±0.30	0.99±0.02	0.87±0.29

Normal viewpoint changes include: **HR** (horizontal rotation), **VR** (vertical rotation), **S** (scaling), and **M** (mixture). The values in the table represent the mean and standard deviation.

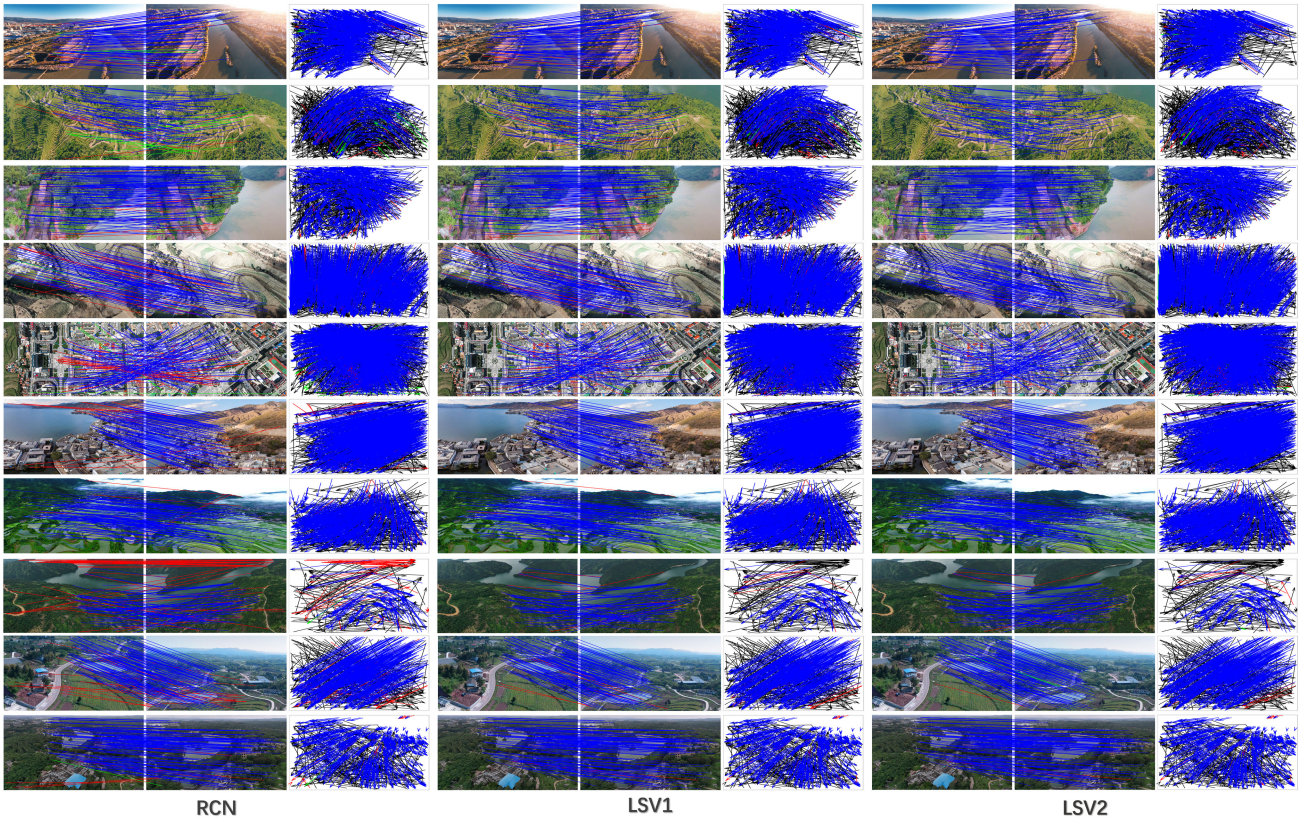


Fig. 10. Feature matching results of our method (including RCN, LSV1, and LSV2 on ten image pairs involving different types of transformations (blue = true positive, black = true negative, green = false negative, and red = false positive). For visual convenience of image pairs, at most 200 randomly selected matches are drawn, and we do not show the true negatives.

GS have advantages in precision, while recall is significantly behind other methods. Due to the limitation of the parameterized model (eight-point algorithm), the performance of LFGC deteriorates drastically under large viewpoint changes.

In fact, LSV is designed to solve the tougher mismatch removal task (e.g., extreme scenarios). Quantitative comparison in extreme viewpoint changes is shown in Fig. 12. As expected, we observe that LSV does have appreciable advantage in terms of F1-score compared with other methods in extreme viewpoint changes; LSV can maintain a 95.28% precision at 99.26% recall. The precision of LMR (the second best method) is generally close to LSV, but only 95.35% recall. Furthermore, LMR needs relatively more neighborhood points to build the structure descriptor and performs poorly when only fewer feature points are given. GLPM is an improvement on LPM, which uses a small putative set with a low threshold (e.g., NNDR threshold) to guide the matching on a large putative set and construct a relatively stable neighboring preservation. GLPM has higher precision than LPM. LSV has no advantage in precision because other methods usually exchange recall for precision, especially for SIR. In general, LSV comprehensive performance is even better.

3) Robustness on Low Inlier Ratio: Inlier ratio can be used to represent the difficulty of removing mismatches. To evaluate robustness of LSV in exceptionally difficult circumstances, five low-inlier-ratio (22.41% average inlier ratio) challenging image pairs involving different types of transformations (including homography, epipolar geometry, nonrigid deformation, and extreme light changes) were selected from ODB. The representative matching results are shown in Fig. 13 . For convenience, the results are summarized in Table IV.

GLPM, LPM, and SIR usually have high recall or precision, but not simultaneously; the gap between them and LSV is further widened. These algorithms rely on low-threshold prematching. However, low-threshold prematching is usually unreliable on low-inlier-ratio scenarios. In the previous experiment, GS had the worst performance and achieved the third best result in low-inlier-ratio image pairs. The major cause is that its performance is related to the number of feature points (GS is a graph matching method), and the number of pairs of feature points in this experiment is only 442–691. LFGC has achieved satisfactory results under low-degree-of-freedom deformation. Table IV shows that LSV is the exclusive algorithm, where recall, precision, and F1-score all exceed 80%.

B. Results on Image Registration

In this section, we focus on image registration according to the feature matching results and follow the same evaluation in [1] and [53]: the root mean square error (RMSE), maximum error (MAE), and median error (MEE). The evaluation criteria are

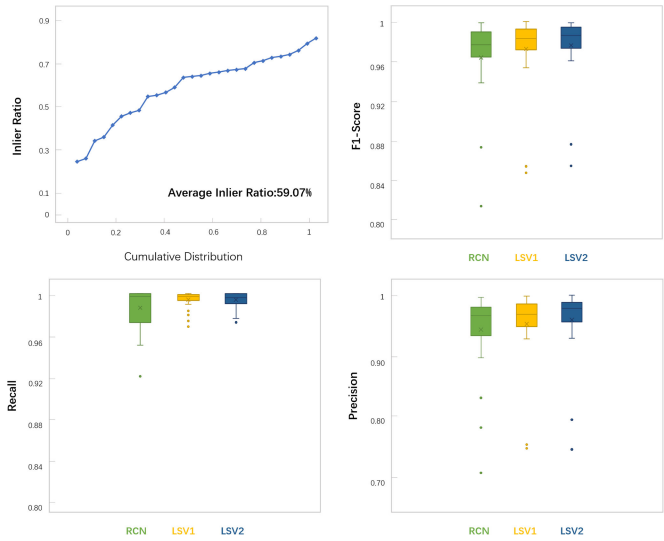


Fig. 11. Quantitative comparisons of contribution of various components. The middle line of the box is the median of the data, the upper and lower limits of the box are the upper and lower quartiles of the data, respectively, and × denotes average. The mean of F1-score of three methods is 96.41%, 97.23%, and 97.57%. The mean of recall of three methods is 98.71%, 99.44%, and 99.44%. The mean of precision of three methods is 94.38%, 95.27%, and 95.92%.

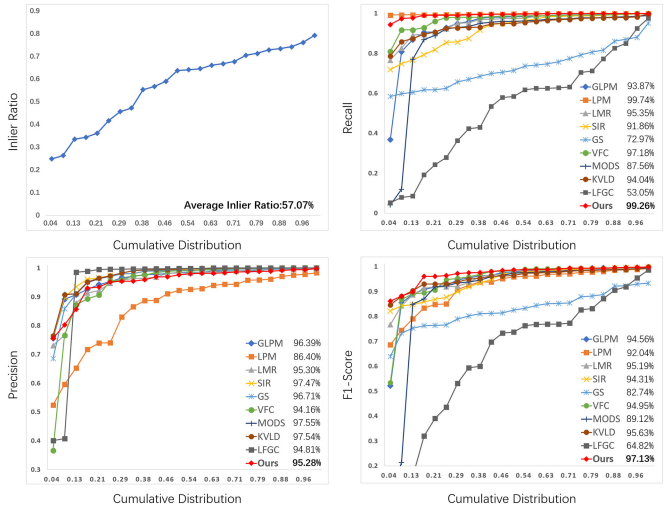


Fig. 12. Quantitative comparisons of ten methods on extreme viewpoint changes. Initial inlier ratio, recall, precision, and F1-score with respect to the cumulative distribution.

defined as follows:

$$\begin{aligned}\text{RMSE} &= \sqrt{\frac{1}{M}\sum_{n=1}^{M}(S_n - R_n)^2} \\ \text{MAE} &= \max\left\{\sqrt{(S_n - R_n)^2}\right\}_{n=1}^{M} \\ \text{MEE} &= \text{median}\left\{\sqrt{(S_n - R_n)^2}\right\}_{n=1}^{M}\end{aligned} \tag{10}$$

where S_n and R_n denote the corresponding landmarks and M denotes the number of selected landmarks. $\max(\cdot)$ and $\text{median}(\cdot)$ return the maximal and median of a set, respectively.

TABLE IV
QUANTITATIVE COMPARISON WITH SIR [32], GLPM [31], LMR [28], LPM [30], GS [29], VFC [49], LFGC [26], MODS [51], AND KVLD [50] ON LOW-INLIER-RATIO SCENARIOS

Method	Recall	Precision	F1
Ours	**88.13%±0.06**	**81.52%±0.07**	**84.48%±0.05**
SIR	64.17%±0.24	97.81%±0.01	74.98%±0.18
GLPM	65.99%±0.19	91.54%±0.03	74.72%±0.14
LMR	79.86%±0.06	76.92%±0.11	78.11%±0.08
LPM	97.73%±0.02	50.73%±0.15	65.69%±0.12
GS	85.38%±0.07	79.85%±0.08	81.94%±0.03
VFC	57.40%±0.25	93.63%±0.05	68.68%±0.20
LFGC	74.92%±0.18	97.02%±0.03	83.52%±0.13
MODS	20.14%±0.16	94.46%±0.09	30.54%±0.22
KVLD	18.76%±0.28	39.90%±0.49	23.85%±0.33

The values in the table represent the mean and standard deviation.

Suppose that a set of inlier $Q = \{(X_i, Y_i)\}_{i=1}^{n}$ is calculated from the putative matches S, where $X_i = (x_i, y_i)$ and $Y_i = (x_i', y_i')$ represent 2-D coordinates. The following two transformation models are used for image registration.

1) Projective Transformation (Homography): In this model, the objective function is defined as

$$\begin{pmatrix} x_i' \\ y_i' \\ 1 \end{pmatrix} = H \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix}. \tag{11}$$

The direct linear transformation algorithm [54] are used to estimate a homography matrix H; (11) can be written as

$$\begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} \text{vec}(H) = 0 \tag{12}$$

where

$$A_i = \begin{pmatrix} -x_i & -y_i & -1 & 0 & 0 & 0 & x_i x_i' & y_i x_i' & x_i' \\ 0 & 0 & 0 & -x_i & -y_i & -1 & x_i y_i' & y_i y_i' & y_i' \end{pmatrix},$$

$i = 1, 2, \ldots, n$, and $\text{vec}(\cdot)$ is the matrix vec operator. First, calculate the least squares solution of H using the singular value decomposition algorithm [55]. Finally, sample pixel from the sensed image I^s based on bilinear interpolation to obtain the transformed image I^t.

2) Thin-Plate Spline (TPS) Transformation [56]: In this model, the transformation coefficient $\theta_{(n+3)\times 2}$ is found by solving the linear system

$$\theta = \begin{pmatrix} \mathcal{K} & X' \\ X'^T & \mathbf{O}_{3\times 3} \end{pmatrix}^{-1} \begin{pmatrix} Y \\ \mathbf{O}_{3\times 2} \end{pmatrix} \tag{13}$$

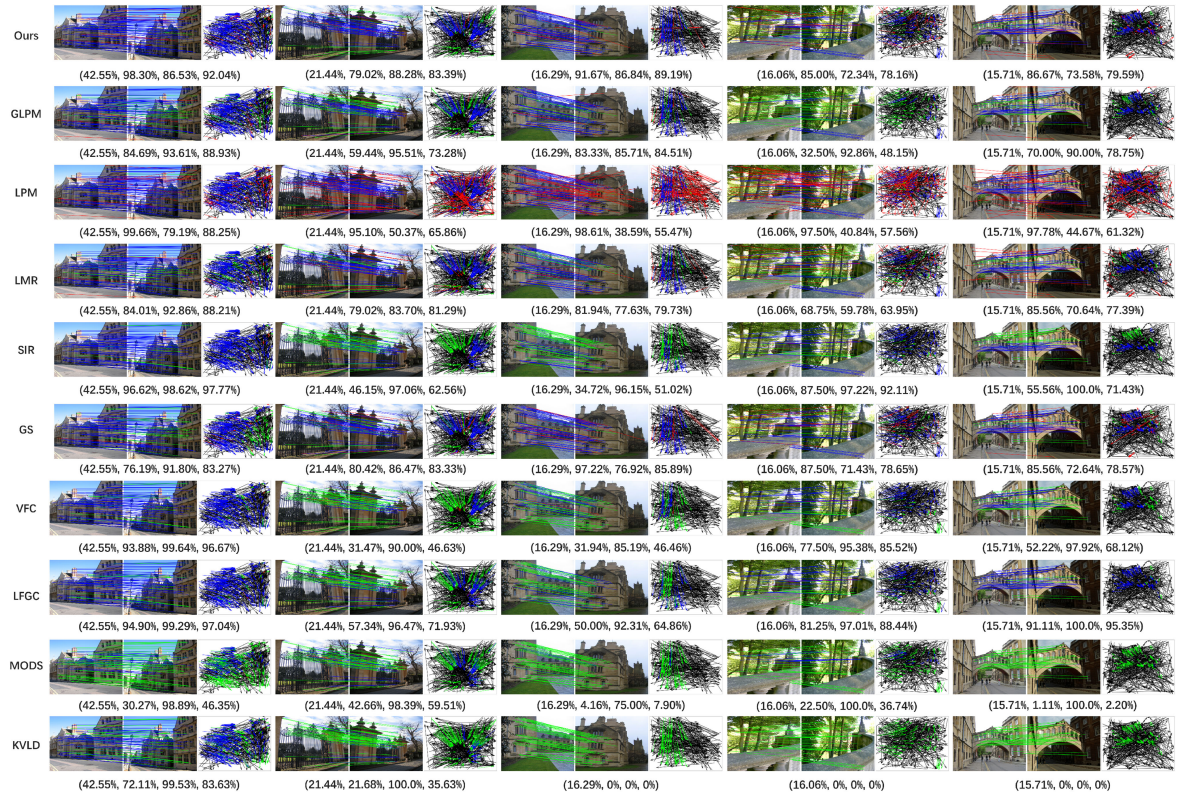


Fig. 13. Feature matching results on low inlier ratio scenarios (blue = true positive, black = true negative, green = false negative, and red = false positive). For each group of results, the first value is the initial inlier ratio, while the rest three values are the recall, precision, and F1-score, i.e., (inlier ratio, recall, precision, and F1-score). For visibility, in the image pairs, at most 200 randomly selected matches are shown, and we do not show the true negatives.

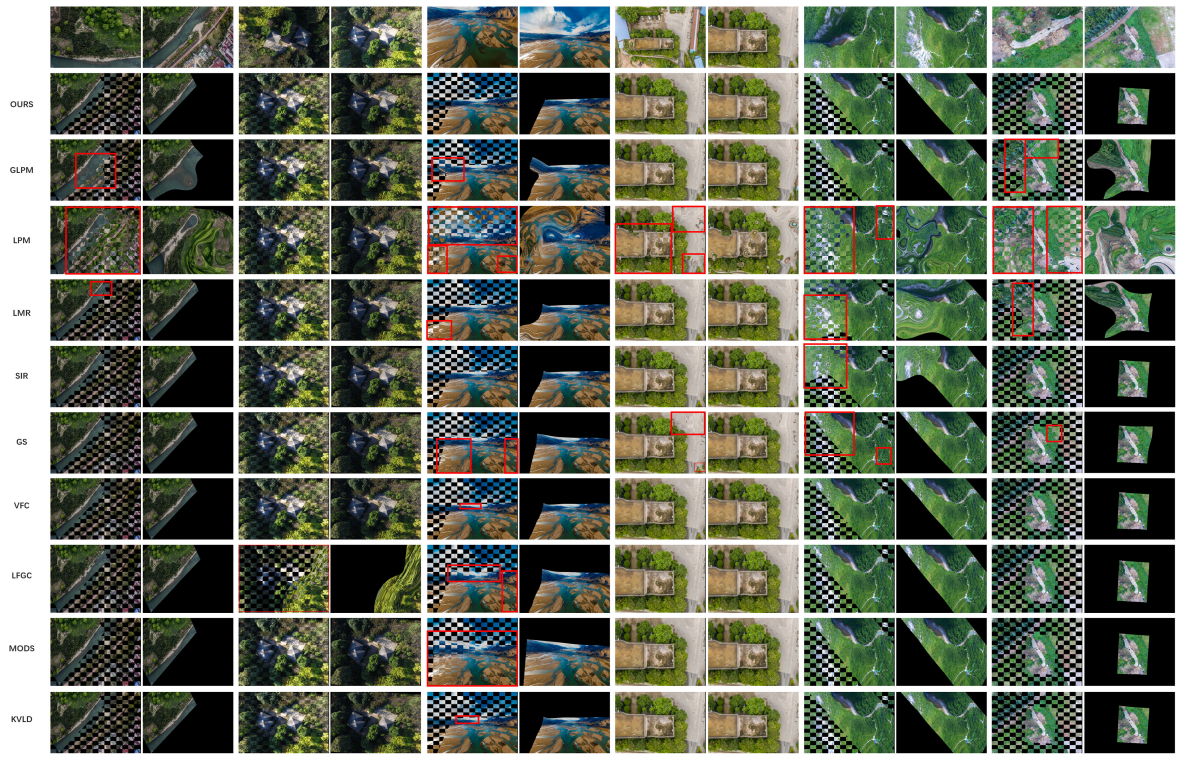


Fig. 14. Representative image registrations of ten methods on six UAV image pairs. The first row is sensed (left image) and reference (right image) images, respectively. Red rectangles indicate the misalignments.

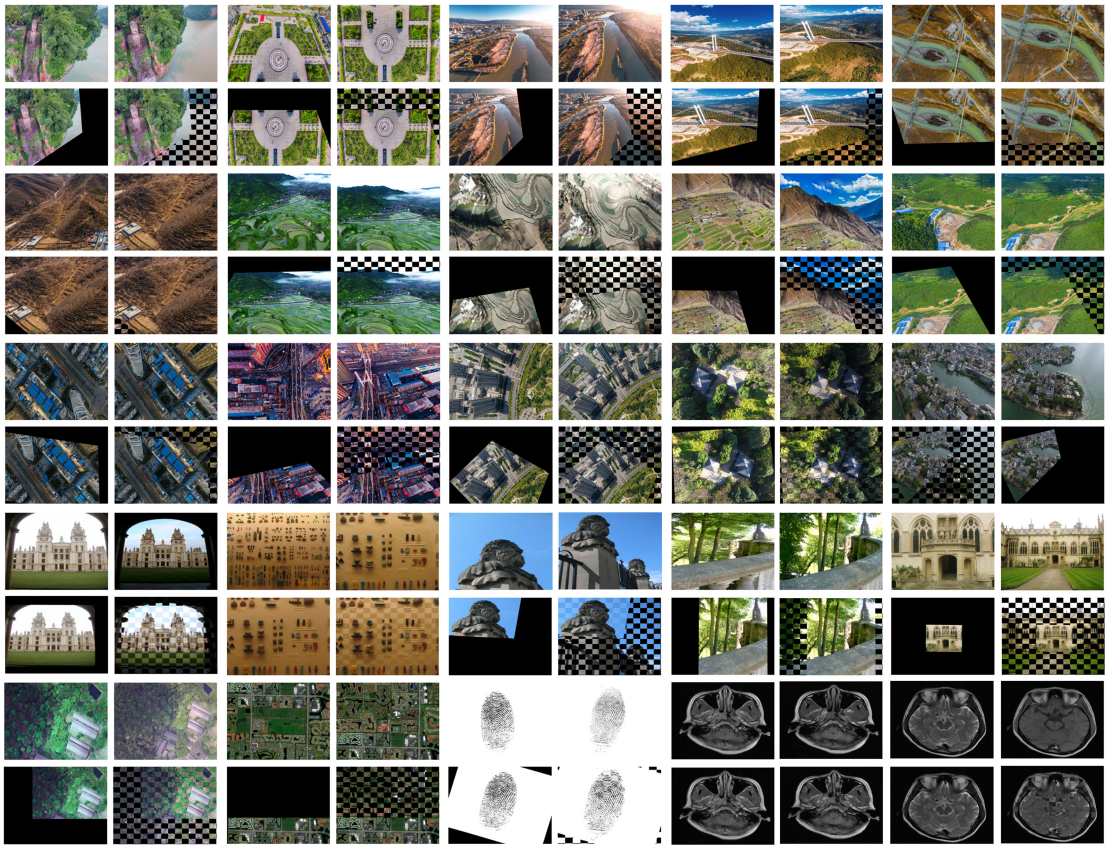


Fig. 15. Representative image registration results of LSV on three datasets. In each category, the first row is sensed (left image) and reference (right image) images, respectively. The first to third categories (i.e., landmark, terrain, and building) are selected from the SUIRD, the fourth category is selected from OBD, and the last category is from MTIRD.

where $\mathcal{K}_{n\times n}$ is a radial basis kernel with each entry computed by $\mathcal{K}_{ij} = \|Y_i - Y_j\|^2 \log\|Y_i - Y_j\|$, $X' = (1, X)$ is the $n \times 3$ homogeneous coordinate. Finally, sample pixel from the sensed image I^s based on bilinear interpolation to obtain the transformed image I^t. The bicubic interpolation [57] is used to improve the smoothness and precision of I^t.

TPS transformation is a more extensive transformation model than homography, but also more sensitive to set Q, that is, if there are a few false correspondences in Q or insufficient true correspondence, it will result in poor registration. However, this property can be used to intuitively reflect the robustness of the algorithm. Fig. 14 shows the registration results of four different types of UAV images using TPS transformation. From the registration results, clearly, a very small proportion of falsely matched can degrade the registration result . VFC and RVLD secure a very close performance to our LSV on the low-altitude remote sensing images. However, in the second pair of images, although most areas are aligned correctly, there is a marked deviation at the junction of the sky and the mountains. Such texture areas, which relatively lack control feature points, are more likely to suffer poor registration results. LPM is clearly unable to satisfactorily align the images since the poor precision results in a lot of image distortion (unsmooth control points can cause intolerable image distortion). SIR proposes a strict local dissimilarity measure, which results in a higher precision, so it has a more ideal registration results using TPS. Our method is superior to the other nine methods. The quantitative comparison results are summarized in Table V.

TABLE V
QUANTITATIVE COMPARISON OF IMAGE REGISTRATION

Method	RMSE	MAE	MEE
Ours	**1.57 ± 0.59**	**2.11 ± 0.79**	**1.61 ± 0.69**
SIR	5.18 ± 7.78	10.99 ± 17.10	5.12 ± 6.54
GLPM	5.96 ± 9.55	12.50 ± 20.44	6.61 ± 9.88
LMR	24.64 ± 50.98	46.51 ± 95.58	40.94 ± 86.25
LPM	28.08 ± 36.78	68.67 ± 89.95	26.96 ± 54.66
GS	5.94 ± 5.61	13.20 ± 12.32	5.80 ± 5.69
VFC	1.97 ± 0.87	4.20 ± 2.47	2.13 ± 0.62
LFGC	42.63 ± 84.13	82.70 ± 160.86	52.31 ± 107.41
MODS	8.88 ± 14.88	19.04 ± 31.33	10.89 ± 19.79
KVLD	2.21 ± 1.29	4.75 ± 3.48	2.14 ± 0.61

The values in the table represent the mean and standard deviation.

Finally, extensive image registration results of our method (including low-altitude image registration, satellite remote sensing image registration, hyperspectral and visible image registration, fingerprint image registration, and multimodal MR image registration) are shown in Fig. 15. Note that fingerprint image pair and multimodal MR image pairs use TPS transformation, and the others use projective transformation.

IV. Conclusion

In this article, we have introduced a novel LSV descriptor for image matching/registration, and it can simultaneously guarantee local structure invariant of feature points in different deformations, severe outliers, various rotations, scaling changes, as well as their extreme mixtures. Experimental results show that our method gives the most stable performance and outperforms the nine state-of-the-art methods on image matching/registration accuracy. Meanwhile, our method is more simple and can be easily implemented, and only requires fewer images as the training set compared with the other methods.

Acknowledgment

The authors would like to thank David G. Lowe, Su Zhang, Jiayi Ma, Shuicheng Yan, Xiangru Li, and Hairong Liu for providing their implementation source codes and experimental datasets, which facilitate the comparison experiments greatly.

References

[1] Z. Yang, Y. Yang, K. Yang, and Z.-Q. 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.

[2] R. Gonzalez and P. Wintz, *Digital Image Processing*, 2nd ed. Reading, MA, USA: Addison-Wesley, 1987.

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

[4] 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.

[5] 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.

[6] B. D. De Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Isgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, 2019.

[7] X. Tong *et al.*, "Image registration with fourier-based image correlation: A comprehensive review of developments and applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 4062–4081, Oct. 2019.

[8] R. Michel and E. Rignot, "Flow of Glaciar Moreno, Argentina, from repeat-pass shuttle imaging radar images: Comparison of the phase correlation method with radar interferometry," *J. Glaciol.*, vol. 45, no. 149, pp. 93–100, 1999.

[9] S. S. Young and R. G. Driggers, "Superresolution image reconstruction from a sequence of aliased imagery," *Appl. Opt.*, vol. 45, no. 21, pp. 5073–5085, 2006.

[10] X. Tong *et al.*, "An improved phase correlation method based on 2-D plane fitting and the maximum kernel density estimator," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1953–1957, Sep. 2015.

[11] X. Tong *et al.*, "A novel subpixel phase correlation method using singular value decomposition and unified random sample consensus," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4143–4156, Aug. 2015.

[12] J. Ma, J. Zhao, Y. Ma, and J. Tian, "Non-rigid visible and infrared face registration via regularized gaussian fields criterion," *Pattern Recognit.*, vol. 48, no. 3, pp. 772–784, 2015.

[13] 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.

[14] 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.

[15] H. Chui and A. Rangarajan, "New algorithm for non-rigid point matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, vol. 2, pp. 44–51.

[16] 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.

[17] 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.

[18] Y. Yang, S. H. Ong, and K. W. C. Foong, "A robust global and local mixture distance based non-rigid point set registration," *Pattern Recognit.*, vol. 48, no. 1, pp. 156–173, 2015.

[19] J. Ma, W. Qiu, J. Zhao, Y. Ma, A. L. Yuille, and Z. Tu, "Robust L2E estimation of transformation for non-rigid registration," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1115–1129, Mar. 2015.

[20] J. Ma, J. Zhao, and A. L. Yuille, "Non-rigid point set registration by preserving global and local structures," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 53–64, Jan. 2016.

[21] Z. Zhou *et al.*, "Accurate and robust non-rigid point set registration using student's-t mixture model with prior probability modeling," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 8742.

[22] H.-B. Qu, J.-Q. Wang, B. Li, and M. Yu, "Probabilistic model for robust affine and non-rigid point set matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 371–384, Feb. 2017.

[23] 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.

[24] J. Zhou *et al.*, "Robust variational Bayesian point set registration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9905–9914.

[25] 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.

[26] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2666–2674.

[27] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4181–4190.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

[36] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 23–79, 2021.

[37] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 467–483.

[38] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, 2016.
[39] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5695–5703.
[40] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3279–3286.
[41] Y. Zheng and D. Doermann, "Robust point matching for nonrigid shapes by preserving local neighborhood structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 643–649, Apr. 2006.
[42] J. E. Bresenham, "Algorithm for computer control of a digital plotter," *IBM Syst. J.*, vol. 4, no. 1, pp. 25–30, 1965.
[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[44] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 539–546.
[45] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit*., 2015, pp. 84–92.
[46] 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.
[47] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.
[48] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.
[49] 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.
[50] Z. Liu and R. Marlet, "Virtual line descriptor and semi-local graph matching method for reliable feature correspondence," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 16. 1–16.11.
[51] D. Mishkin, J. Matas, and M. Perdoch, "MODS: Fast and robust method for two-view matching," *Comput. Vis. Image Understanding*, vol. 141, pp. 81–93, 2015.
[52] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[53] Y. Liu, X. Gong, J. Chen, S. Chen, and Y. Yang, "Rotation-invariant siamese network for low-altitude remote-sensing image registration," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 2020, vol. 13, pp. 5746–5758.
[54] Y. Abdel-Aziz, H. Karara, and M. Hauck, "Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry," *Photogrammetric Eng. Remote Sens.*, vol. 81, no. 2, pp. 103–107, 2015.
[55] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. New York, NY, USA: Springer, 1971, pp. 134–151.
[56] 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.
[57] A. C. Gallagher, "Detection of linear and cubic interpolation in JPEG compressed images," in *Proc. 2nd Can. Conf. Comput. Robot Vis.*, 2005, pp. 65–72.

Jiaxuan Chen received the bachelor's degree in information management and system from Northwest Normal University, Lanzhou, China, in 2018. He is currently working toward the master's degree with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

His current research interests include computer vision and deep learning.

Shuang Chen received the bachelor's degree in information management and system from Xihua Normal University, Chengdu, China, in 2019. She is currently working toward the master's degree with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

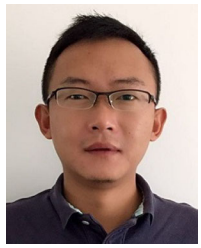
Her current research interests include computer vision and remote sensing image processing.

Yuyan Liu received the bachelor's degree in network engineering from the Shanghai University of Applied Technology, Shanghai, China, in 2018. She is currently working toward the master's degree with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

Her current research interests include computer vision and remote sensing image processing.

Xiaoxian Chen received the bachelor's degree in business administration from Guangdong Pharmaceutical University, Guangzhou, China, in 2018. He is currently working toward the master's degree with the Laboratory of Computer Networks and Information Security, China Agricultural University, Beijing, China.

His current research interests include computer vision and deep learning.



Yang Yang (Member, IEEE) received the master's degree from Waseda University, Tokyo, Japan, in 2007, and the Ph.D. degree from the 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 interests include computer vision, remote sensing, geography information systems, and medical imaging.

Yungang Zhang received the master's degree in electrical engineering from Yunnan Normal University, Kunming, China, in 2005, and the Ph.D. degree in computer science from the University of Liverpool, Liverpool, U.K., in 2014.

He is currently an Associate Professor with the School of Information Science and Technology, Yunnan Normal University. His research interests include machine learning and image analysis.