A Point–Set–Domain Image Object Matching Method for Airborne Object Localization

Xiaomin Liu[1], Runqi Zhao[1], Jun-Bao Li[2], Jeng-Shyang Pan[3], Huaqi Zhao[1]*

[1] The Heilongjiang Provincial Key Laboratory of Autonomous Intelligence and Information Processing, Jiamusi University, China
[2] Faculty of Computing, Harbin Institute of Technology, China
[3] College of Computer Science and Engineering, Shandong University of Science and Technology, China
xiaominliu@vip.sina.com, jarvilinh@gmail.com, lijunbao_hit@163.com, jspan@ieee.org, zhaohuaqi@126.com

Abstract

Image object localization is an important research direction in the development of intelligent autonomous control systems for unmanned aerial vehicles (UAVs). Major challenges remain, such as cross-view images, large-scale deformation, and multitemporal variation. We propose a point–set–domain matching method to locate objects. First, the property constraints of a point, including sparsity, repeatability, and distinguishability, are combined into a keypoint response used to optimize convolutional neural networks, creating keypoint detector and feature descriptor models. With these models, we can improve the performance of point matching and obtain the corresponding keypoint set accurately. This approach solves the cross-view problem. Second, a spatial transformation model of the corresponding keypoint set is obtained using keypoint-constrained diffeomorphism matching, which can align the spatial location of two images and solve the large-scale deformation problem. Third, an approach combining probability statistics with watershed maximally stable extremal regions is proposed to divide the object image and reference image into several subregions, and then the similarity based on diffeomorphism is employed to localize the object in the UAV image, which solves the multitemporal variation problem. The experimental results show that the proposed method can successfully determine the location of the object in the UAV image.

Keywords: Image object localization, Point matching, Set matching, Region matching, Diffeomorphism

1 Introduction

Drones ones are an important area of technological development. They are intelligent robots with autonomous positioning, operation, and planning capabilities and can work in complex environments. The United States has released the "Unmanned Systems Integrated Roadmap (2017–2042)," which states that the perception and navigation capabilities of unmanned systems need to be improved [1].

Image object localization for intelligent autonomous control of drones has become a major research direction, and its applications have expanded from military to civilianareas. In military applications, drones are used for high resolution reconnaissance [2-3], and network offense and defense [4]. In civilian applications, drones can be used in bridge inspection [5-6] of photovoltaic power plants, and forest fire prevention [7], among other areas.

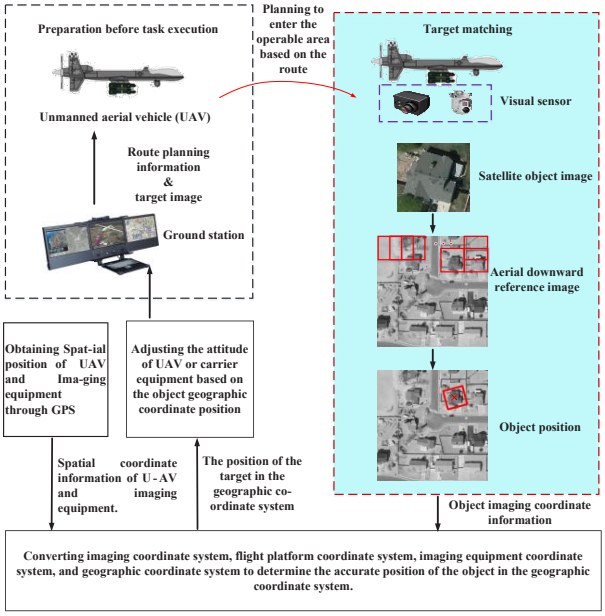

Figure 1. The basic workflow of UAV object localization task

The typical workflow for drone object localization is shown in Figure 1. Given an object localization task, the ground workstation provides the planned route and a satellite object image, and the drone enters the operation area according to the planned route. The location of the object in the drone image is determined using object matching methods, illustrated in the red dashed box in Figure 1. As shown in Figure 2, the collected satellite images are usually overhead images, and the camera carried by the drone usually has a 360° acquisition angle. There are still challenges in completing complex drone

*Corresponding Author: Huaqi Zhao; E-mail: zhaohuaqi@126.com
DOI: https://doi.org/10.70003/160792642025052603003

object matching, as shown in Figure 3. The difficulty lies in the different viewpoints, large-scale deformation, and ground changes between the satellite reference image and the airborne object image in the object localization task.

Existing object matching methods often fail to accurately complete object localization tasks. Therefore, this paper analyzes object matching and localization to develop novel methods. First, local keypoints have better discriminability in cross-view images, and effective point matching methods can solve the problem of cross-view changes in object localization. Second, an effective set matching method is used to determine the deformation relationship between the satellite reference images and the object images to solve the problem of large-scale deformation. Finally, an image matching method based on region division is adopted.

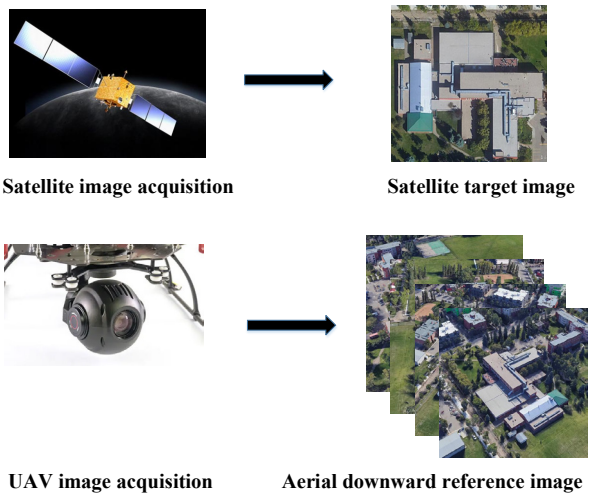

Figure 2. The basic workflow of UAV target localization task

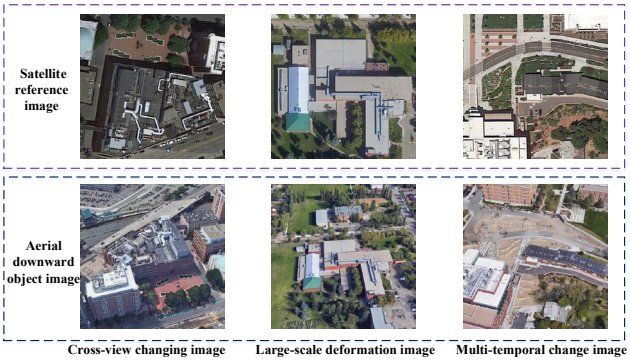

Figure 3. Illustration of satellite image and object image

This article makes four main novel contributions:

1. We propose a three-level image object matching framework based on points, sets, and domains. This framework addresses the challenges of cross-view images, large-scale deformation, and ground angle variation in unmanned aerial vehicle (UAV) image object localization.

2. To address the problem of cross-view matching in UAV image object matching, we propose a point matching method based on keypoint response constraints. This method improves the performance of point matching by incorporating keypoint response into the loss function used to train neural networks, building on existing learning-based methods for point matching.

3. To address the issue of large-scale deformation in UAV image object matching, we propose a diffeomorphism set matching method based on keypoint constraints. Corresponding point sets are used as constraints to generate a space transformation model, and a static velocity field is introduced to effectively improve the performance of set matching.

4. To address the problem of ground change in UAV target matching, we propose a domain matching method based on region partitioning. Probability statistics and watershed maximally stable extremal region (MSER) detection methods are used to partition the image into regions. Furthermore, the diffeomorphism similarity calculation method is used for the first time to determine the position of a satellite object image in an airborne image.

2 Related Work

Object matching is an important research topic in the field of pattern recognition, whose aim is to identify objects of the same type in two or more images using a matching algorithm. In this paper, object matching is divided into a hierarchical matching process from local to global, which effectively determines the position of the object in the airborne image.

2.1 Feature Point Matching Research

In point matching, the keypoint is first obtained by a matching method, and the corresponding point set of the image is determined by calculating the similarity of the keypoints. An early manual point matching algorithm is the Harris corner matching method, which has problems such as a fixed scale, a low pixel positioning accuracy, a tendency to produce many false corner points, and a high computational complexity [8]. Building on this, FAST meets the requirements of real-time positioning systems [9], and the SIFT [10] method improves the accuracy of feature matching. PCA-SIFT [11], SURF [12], SSIFT [13], and other methods have been proposed successively. With the application of mobile devices in many computer vision tasks, binary descriptor operators have also received increasing attention. Liu et al. proposed a new ring sampling binary descriptor operator [14].

In recent years, learning point matching methods have become a research hotspot. DeTone et al. proposed a self-supervised keypoint learning framework, Superpoint [15], which has achieved a strong matching performance. However, due to the limited variety of detected keypoints, the algorithm may fail in certain special applications. Recently, the idea of point matching has also been widely used in Object detection and Real-time Reconstruction [16-18].

In summary, in recent years, effective point matching methods are still based on learning-based point matching methods. However, due to the limited variety of detected

keypoints, the algorithm may fail in certain special applications.

2.2 Set Matching

Geometric transformation is used to spatially align corresponding points and faces in two images, which can eliminate or weaken the problem of large-scale deformation between satellite reference images and airborne object images. Feature matching can be summarized as solving the spatial transformation relationship between two corresponding point sets [19]. Rigid feature matching has been widely applied in various fields, and the most common and influential method is the iterative closest point (ICP) algorithm proposed by Besl et al. and its extension methods [20]. Non-rigid transformation can be defined as a spatial transformation with local geometric deformation and has become an important research direction in image matching, being applied in many complex image processing tasks [21]. There are many non-rigid matching methods that can describe the feature matching process [22]; the most well-known is the thin plate spline (TPS) method. Chui et al. have proposed a robust point matching method [23].

In recent years, the environment of image matching tasks has become increasingly complex, and object matching methods that are applicable to large-scale changes have become a focus of research. Large deformation diffeomorphism metric mapping (LDDMM) has been proven to solve the problem of large-scale deformation [24]. Tan et al. proposed using multi-resolution diffeomorphism matrix projection to solve this problem [25]. Many authors have conducted in-depth analyses on diffeomorphism set matching for various applications. This approach has to some extent solved the problem of large-scale deformation [26-27].

For the large-scale deformation problem addressed in this paper, non-rigid matching methods are a good choice. However, these methods do not perform well when directly applied to the problem of UAV image object localization. Reducing the number of degrees of freedom can solve the problem of large-scale deformation in matching.

2.3 Domain Matching

Regional features possess high invariance and stability, and these features have repeatability in multiple images from different viewpoints, complementing other methods. Template matching is considered to be the simplest method of regional matching, with representative methods such as fast affine template matching (FAST-Match) [28] and MSER [29]. Alim et al. classified multispectral images using morphological contours guided by extremal regions and maximum stable extremal regions, which have high robustness to changes in viewpoint and the same complexity as MSER [30]. Recently, the idea of domain matching has also been widely used in Image Reconstruction [31].

For object matching in complex environments, an improved optimization method is required to divide the satellite reference image and the airborne image into regions and then calculate the similarity between the two images. Domain matching methods based on region division can effectively solve the problem of ground cover change.

3 An Image Object Matching Framework Based on Points, Sets, and Domains

The proposed point–set–domain framework for image object matching is shown in Figure 4. The architecture consists of three main parts: point matching based on keypoint response constraints, set matching based on diffeomorphism, and domain matching based on region partitioning. These are detailed below.

Different imaging methods are used to produce satellite reference images and airborne object images, leading to cross-view problems between them. In point matching, the keypoint response is crucial in determining the keypoints, and is used to represent the probability of a point being a keypoint. Different point matching methods result in keypoint responses with different attributes. In a learning-based point matching method, this paper proposes to use keypoint response to generate the loss for training the network. This constrains the generation of keypoint detectors and feature descriptor operators, and improves the performance of point matching.

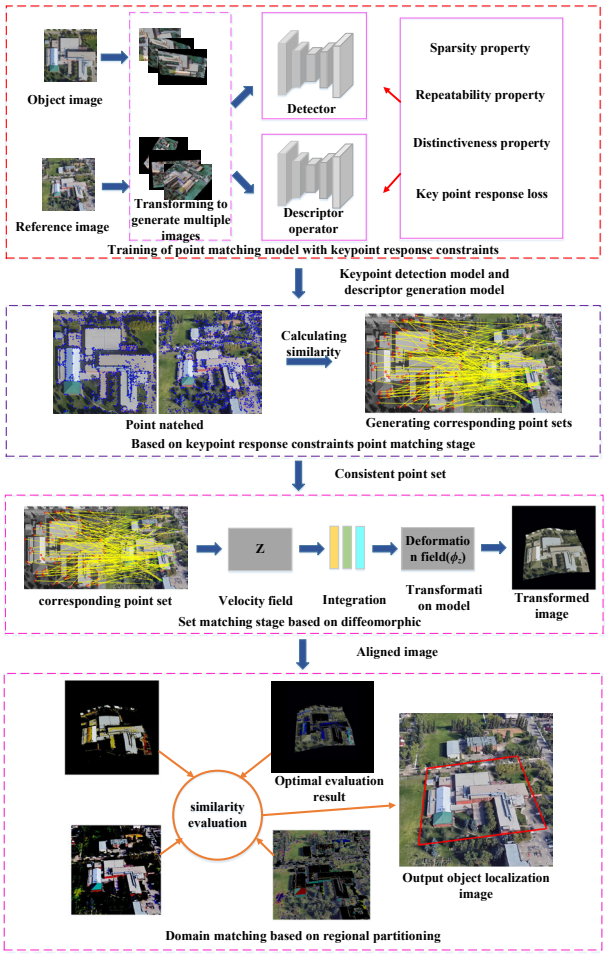

Figure 4. Illustration of object image and reference image

3.1 Point Matching Based on Keypoint Response Constraint

Different imaging methods are used to produce satellite reference images and airborne object images, leading to cross-view problems between them. In point matching, the keypoint response is crucial in determining the keypoints, and is used to represent the probability of a point being a keypoint. Different point matching methods result in keypoint responses with different attributes. In a learning-based point matching method, this paper proposes to use keypoint response to generate the loss for training the network. This constrains the generation of keypoint detectors and feature descriptor operators, and improves the performance of point matching.

3.2 Set Matching Based on Diffeomorphsm

In a complex environment, there are large-scale deformations in the satellite reference images and airborne object images. Determining the spatial transformation model is a solution to address this problem. In this paper, we propose to use corresponding keypoint sets for set matching based on diffeomorphism, and employ a keypoint-constrained diffeomorphism set matching method to obtain the spatial transformation relationship between the satellite object images and airborne reference images. This allows the determination of the position of the satellite reference image in the object image. To improve the efficiency of the algorithm, we introduce the theory of static velocity domain, which also solves the problem of large-scale deformation in UAV image object matching tasks.

3.3 Domain Matching Based on Region Partitioning

UAVs typically capture images at a range of different times, leading to occlusion problems in the reference images, such as snow coverage in winter and grass coverage in summer. Using global similarity calculation methods directly can affect the performance of UAV image object localization. In this paper, we propose for the first time an effective region partitioning method for satellite reference images using a combination of probability statistics and watershed segmentation. We also innovate by using a diffeomorphism similarity extremal method to obtain an accurate position for the object. This addresses the problem of terrain changes in UAV image object localization.

4 Principles of the Point-Set-Domain Object Matching Method

In UAV image object localization tasks, there are problems such as cross-view images, large-scale deformation, and changes in terrain. Point matching methods have become an important part of researching cross-view object localization. Set matching can effectively complete the spatial transformation between satellite reference images and airborne object images, and it is necessary to study set matching methods for large-scale deformation problems in airborne object localization tasks. The area division for optimal domain matching is key to solving terrain change problems. This article proposes innovative point matching methods with keypoint response constraints, set matching methods based on diffeomorphism, and domain matching methods based on area division to address cross-view, large-scale deformation, and terrain change problems in UAV object image matching tasks.

4.1 Point Matching with Keypoint Response Constraints

This article adopts the general theoretical framework proposed by Yan et al. for learning keypoint detector and descriptor operators, which only considers properties such as sparsity, repeatability, and distinctiveness. However, these properties cannot be used alone to effectively extract keypoints from cross-view images. To combine the advantages of different keypoint detectors, this article proposes a point matching method based on keypoint response constraints.

The point matching with keypoint response constraints first needs to construct keypoint detectors and feature descriptor operators. According to the theoretical framework in [32], the keypoint detector and feature descriptor operator model can be obtained using the following constraints:

$$\arg\max_{\theta_F} \prod_j \prod_v P_v\left(IP_j, DS_j\right) \tag{1}$$

where $P_v(IP_j, Ds_j)$ represents the probability of satisfying the *vth* property, $v \in \{1, 2, ..., V\}$, and V is the number of desired properties, assuming all properties are independent. Thus, certain keypoint properties can be used to describe the probabilities relevant to the detector and descriptor operators by (1), and the keypoint response values can be obtained through optimization algorithms.

Using the aforementioned properties to constrain convolutional neural networks is theoretically feasible, but in practice, it is difficult to find keypoints that fully satisfy the constraints of sparsity, repeatability, and distinctiveness. To address this issue, this article proposes an innovative point matching method based on keypoint response constraints. Assuming that a certain existing keypoint detector (e.g., SIFT, Superpoint) is used to obtain the keypoint response image O, the keypoint detector and feature descriptor operator model can be optimized by jointly using the keypoint response O and the attribute optimization formula $P_{fusion} = \arg\max_{\theta_F} \prod_j \prod_v P_v\left(IP_j, DS_j\right)$. Therefore, this article constructs a convolutional neural network loss function as follows:

$$\begin{cases} \mathcal{L}_t = \mathcal{L}_p(\mathcal{X}, Y) + \mathcal{L}_d(\mathcal{X}, Y) + \mathcal{L}_o(\mathcal{X}, O) \\ \text{s.t. } \operatorname{argmax} P_{\text{fision}} \end{cases} \tag{2}$$

where Y is the original image, $\mathcal{X}$ is the convolution image,

O is the keypoint response image, $\mathcal{L}_p$ represents the cross-entropy calculation of the fully convolutional neural network, and $\mathcal{L}_d$ is the descriptor loss. Full details of the method of calculation are given in [33]. The innovation of this formula lies in the introduction of the keypoint response loss $\mathcal{L}_o$, which represents the normalized gray space cross-entropy calculation. Here, $x_{hw} \in \mathcal{X}$, and the calculation formula is as follows:

$$\mathcal{L}_o(\mathcal{L}, O) = \frac{1}{H_c W_c} \sum_{h=1, w=1}^{H_c, W_c} l_o\left(x_{hw}; o_{hw}\right) \tag{3}$$

where h and w represent the coordinate positions, and $l_o(x_{hw}; o_{hw})$ is defined as follows:

$$l_o\left(x_{hw}; o_{hw}\right) = -\log\left(\frac{\exp\left(x_{hwo}\right)}{\sum_{k=1}^{K} \exp\left(x_{hwk}\right)}\right) \tag{4}$$

where K is the number of pixels after convolution.

Based on the above theory, the detector F and feature descriptor operator D an be obtained. The inner product operation is used to compute the similarity between two features D_1 and D_2 when calculating the corresponding key-point set. The formula is as follows:

$$\text{sim} = D_1 \bullet D_2 \tag{5}$$

Where • denotes the inner product. As shown in Figure 5, the corresponding keypoint set between the satellite ref-erence image and the airborne object image can be determined using (5).

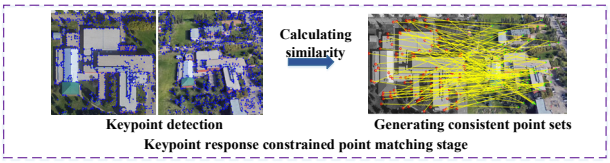

Figure 5. Illustration of the point matching result

4.2 Set Matching Based on Diffeomorphsm

Based on the multi-scale kernel maximum mean discrepancy diffeomorphism projection proposed by Pai et al. [34], this paper investigates a keypoint-constrained diffeomorphism set matching method for addressing large-scale deformation in UAV image object localization. Based on the definition of diffeomorphism [29], we optimize the spatial transformation model using the following formula:

$$\begin{aligned} \underset{\varphi}{\text{argmin}}(E(R,T)) = \underset{\varphi}{\text{argmin}}\,\text{Diss}(R, T \circ \varphi) \\ + \rho\,\text{Reg}(\varphi) \end{aligned} \tag{6}$$

where ρ is the constraint that controls the regularization freedom. The diffeomorphism φ transforms the satellite reference image to the airborne object image on board at $t = 1$. The novelty of our approach lies in the use of a static velocity field to fit the spatial transformation model φ under the constraint of the corresponding keypoint set. The definition of the velocity field V can be extended to the family of velocity fields V_m and can be described as follows:

$$v_m = \sum_{i_m}^{N_m} K_m\left(x_{i_m}, x\right) a_{i_m}^m \tag{7}$$

where the value function (6) is described in a multi-scale reproducing kernel framework as follows:

$$v(x) = \underset{v_m}{\text{argmin}} E\left(I_1, I_2\left(\exp\left(v_m\right)\right)\right) \tag{8}$$

The static velocity field $v(x, t)$ defined as a constant. By parameterizing it with the *SVF* method, high computational speed can be achieved.

Therefore, applying group theory, we consider the velocity field as a member of the Lie algebra, obtain a member of the Lie group space by exponentiation, define a linear combination of basis functions through the discrete parameters of the velocity field, and express it as follows:

$$v(x) = \sum \alpha_i \rho_i(x) \tag{9}$$

where $p_i(x)$ is the basis function and α_i is the coefficient. Finally, by optimizing (9), the projection φ can be obtained, where x is the position vector for the corresponding keypoint set, which means that the determination of φ is not performed on the entire image but on the corresponding set of keypoints. This enhances the performance of diffeomorphism set matching in UAV image object localization. As shown in Figure 6, this lays the foundation for subsequent object matching similarity calculations.

Figure 6. Image object matching based on point-, set-, domain-

4.3 Domain Matching Method Based on Region Division

This paper investigates the detection method of probability statistical MSER, which builds on the traditional MSER method. According to the description in [29], the definition of MSER initially comes from edge confidence. It is a truncation algorithm that minimizes the confidence interval estimation around the mean, and the confidence interval of the half-bandwidth of the mean is estimated by the following formula:

$$CI = \frac{z_{\sigma/2} S(n)}{\sqrt{n}} \tag{10}$$

where $z_{\sigma/2}$ is the standard normal distribution with a standard deviation of $\sigma/2$, $S()$ represents the sampling estimate of the standard deviation, and n is the number of sampling points for the estimate.

To improve the accuracy of regional localization, this paper innovatively combines probability statistical region detection with watershed-based methods to obtain the maximum stable extreme value region detection formula as follows:

$$q(i) = \frac{CI_{i+\Delta} - CI_{i-\Delta}}{CI_i} \tag{11}$$

Where CI represents the confidence interval of the maximum stable extreme value region, and the local minimum of the calculated $q(i)$ value is the detection of the maximum stable region. The region division result is shown in Figure 7.

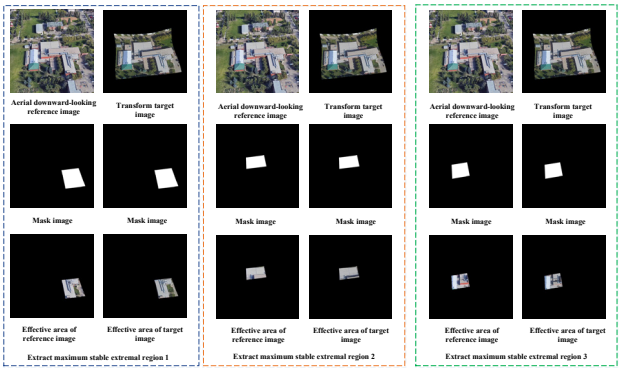

Figure 7. Illustration of the region matching based on region division

In traditional similarity calculations for object matching, the SSD method is commonly used. However, this method does not take into account nonlinear changes in image grayscale. This paper proposes an innovative approach that uses diffeomorphism for the regional similarity calculation, which solves the problem of terrain changes in airborne object matching.

Diffeomorphism is currently mainly used for image shape matching and detecting changes in object positions. However, in the color space of regions, when the spatial relationship between two regions is determined, the color information should also have a shape relationship. Therefore, this paper proposes a similarity calculation method based on color diffeomorphism for regional matching.

Given a satellite reference image R and an airborne object image T, let the color space domain be $\Omega \in C^d$, where $d = 3$ represents the number of channels in the color space. The color information of all point sets in the effective region forms a vector set. To better use the diffeomorphism method to accurately project the effective region of the satellite reference image onto the effective region of the airborne object image, two effective regions R_e and T_e are given for the satellite reference image R and the airborne object image T, where $e = 1, ..., k$, k represents the number of effective regions, and the color space domain is $\Omega \in C^d$. Then, the color space transformation $\varphi_c: \Omega \times \square \to \Omega$ can align the two color spaces. The similarity evaluation $Diss(.,.)$ is defined as follows:

$$\mathrm{Diss}(C_r, C_t \circ \varphi_c) = \frac{1}{2} \| C_r - C_t \circ \varphi_c \|^2 \tag{12}$$

Where C_r and C_t are the corresponding color information in the airborne object image and the satellite reference image, respectively, and the similarity evaluation of the effective region is calculated through $Diss$. As shown in Figure 7, after performing the region division using the maximum stable extreme value region detection method, multiple region division masks can be obtained. The effective regions of the reference image and the transformed object image can be obtained through the mask image, and the similarity evaluation calculation can be completed in this effective region. Meanwhile, the optimal value of $Diss$ is searched to determine whether the object matching is accurate.

5 Steps for Point-Set-Domain Based Image Object Matching Method

As outlined in Algorithm 1, the point–set–domain image object matching method first selects partial satellite object images and airborne images. The Superpoint point matching method is used to obtain the keypoint response images O of the satellite object image and the airborne object image. Then, the keypoint response is used as the loss optimization formula (2), and the keypoint detection model and feature descriptor operator are obtained using convolutional neural networks.

Next, the keypoint detector and feature descriptor operator are used to detect the keypoints in the given satellite reference image and the airborne down-looking reference image. The similarity between the keypoints is calculated using (5) to obtain a corresponding point set. For the corresponding point set, (9) is used to determine the transformation model φ of the satellite object image and airborne object image, and φ is used to align the spatial positions of the satellite reference image and the airborne object image.

Finally, (11) is used to divide the transformed satellite reference image into several regions, and (12) is used to calculate the similarity of the corresponding regions. Through the point–set object matching method, the location of the object image can be obtained as shown in Figure 8.

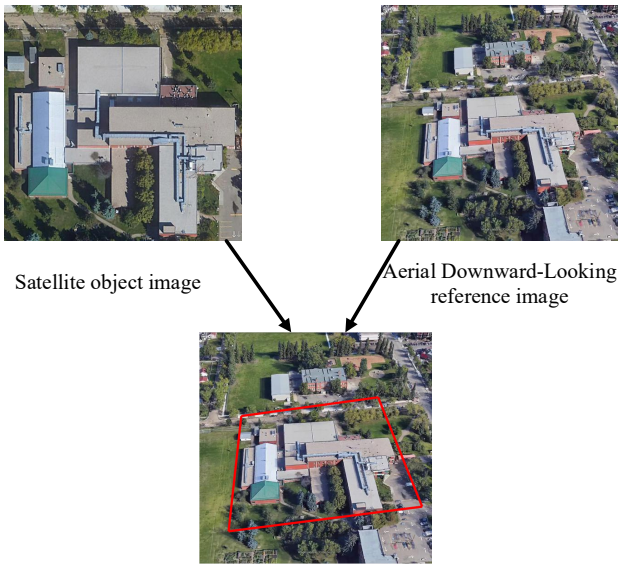

Figure 8. Illustration of the region matching

6 Experimental and Analytical Study on Object Matching of Tertiary Images in Point Cloud Domain

6.1 Standard Dataset

Zheng et al. provided a multi-view and multi-source image dataset, University-1652, based on image geolocation [36]. This dataset differs significantly from early geolocation datasets, which were mostly based on image pairs in which the reference and object images were from different platforms such as cameras and satellites. In contrast, the object and reference images in the University-1652 dataset are from satellite and UAV platforms.

In this study, satellite and UAV images were used, as shown in Figure 9. The experimental dataset includes image pairs with different viewpoints, large-scale deformations, and ground cover changes.

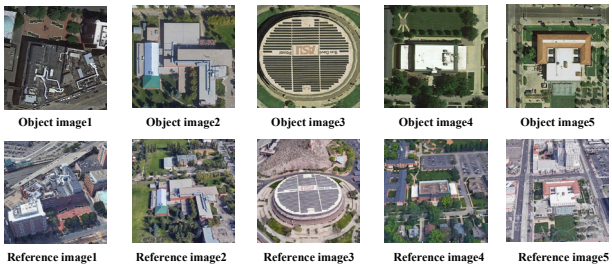

Figure 9. Illustration of the satellite object and the UAV reference image

6.2 Evaluation Metrics

Based on the object matching idea, this paper divides evaluation metrics into four categories: point matching, set matching, domain matching, and object matching metrics. Based on these four categories, the performance of the proposed algorithm was experimentally verified.

1. Point matching metrics. Based on the existing evaluation methods for point matching [38-40], this paper uses five evaluation metrics to evaluate point matching performance [41], according to the five key feature attributes summarized in [40]: repeatability rate, recall rate, accuracy rate, quantization rate, and running time:

Algorithm 1. Point–Set–Domain object matching algorithm

Input: Satellite object image S,
airborne down-looking reference image T
Output: Position of the object in the airborne image
1. Assuming the traditional keypoint detection is KDF, the key response $O_s = KDF(s)$, $O_T = KDF(T)$
2. **for** $iter\_i = 1$ to N **do**
3. compute the loss of function with eq. (1), obtain the keypoint detector K and feature descriptor operator model M.
4. compute the feature descriptors of keypoints in the object image and the reference image $OM = H(O)$, $RM = M(R)$
5. computing $v(x) = argminE(OM, RM(exp(v_x))$ to obtain the transformation model y.
6. **end for**
Return: the airborne object image and mark the position of the object.

A. The Repeatability rate *RPR* is defined as:

$$RPR = \frac{CKN}{KN} \tag{13}$$

where *CKN* represents the number of corresponding keypoints detected by the keypoint detector, and *KN* represents the total number of detected keypoints.

B. The Recall rate *RR* is defined as:

$$RR = \frac{DTMPN}{DTMPN + UDTMPN} \tag{14}$$

where *DTMPN* is the number of corresponding keypoints correctly matched by the feature matching algorithm, and *UDTMPN* is the total number of corresponding keypoints that were not correctly matched or not detected by the algorithm.

C. The Accuracy rate (*AR*) is defined as:

$$AR = \frac{DTMPN}{CKN} \tag{15}$$

where *DTMPN* is the number of corresponding keypoints correctly matched by the feature matching algorithm, and *CKN* represents the total number of corresponding keypoints detected.

D. The Quantization rate (*QR*) is defined as:

$$QR = \frac{KN}{IPN} \tag{16}$$

where *KN* represents the total number of detected keypoints, and *IPN* represents the total number of pixels in the image.

E. The runtime (*RT*) is defined as:

$$RT = T_{CKD} + T_{CKDM} \tag{17}$$

where *TCKD* represents the keypoint detection time, and *TCKDM* represents the corresponding keypoint detection time.

2. Set matching evaluation metrics. This paper uses the root mean square error (RMSE) as the evaluation criterion for set matching, and tests the intra-class and inter-class matching performance of the algorithm using the decidability index.

Given a set of points in the satellite reference image and the object image, denoted by *TI* and *RI*, respectively, the RMSE is defined as:

$$RMSE(TI', RI) = \sqrt{\frac{1}{NM}\sum_{x=1}^{M}\sum_{y=1}^{N}(TI'(x,y) - RI(x,y))^2} \tag{18}$$

where $TI' = A \times TI$, A is the transformation matrix determined by set matching, N and M represent the height and width of the image, and x and y represent the position of the pixels in the image, respectively.

3. Domain matching evaluation metric. In this paper, the optimal evaluation of the matched image is carried out through the domain matching process, and the target position is determined based on the optimal evaluation. The region root mean square error (RRMSE) is defined similarly to (18) and can be considered as the RMSE within the region.

To effectively evaluate the discrimination between inter-class and intra-class matching in an algorithm, this paper proposes a decision index to evaluate the performance of the algorithm matching, defined as follows:

$$RMSE - DI = \left| \frac{|m_g - m_1|}{\sqrt{(s_g + s_i)/2}} \right| \tag{19}$$

where $m_g(m_i)$ and $S_g(s_i)$ represent the mean and standard deviation of the intra-class and inter-class RRMSE, respectively. A larger value indicates stronger discrimination between intra-class and inter-class, and better matching performance.

4. Object matching evaluation index. This paper aims to apply the point–set–domain object matching method to UAV image object localization, so the receiver operating characteristic *ROC* curve is used to verify the performance of the object matching method [41]. The curve is a plot of the true positive rate against the false positive rate. The area under the curve *ROC* represents the performance of the object matching method, with a larger indicating better performance.

6.3 Experimental Setup

We conducted experiments on the University-1652 dataset to verify the performance of the algorithm using 600 cross-view images, comprising 100 satellite reference images and 500 airborne object images. Each satellite object image corresponds to 5 different conditions of airborne object images for the same object. Similarly, 600 large-scale deformation datasets and ground cover change datasets were also selected. The experimental framework evaluated the effects of point matching, set matching, and domain matching on cross-view, large-scale deformation, and ground cover change object matching.

6.4 Experimental Results and Analysis

6.4.1 Results and Analysis of Point Matching Experiment

This section validates the point matching method on the cross-view dataset and provides experimental comparisons and analysis. Table 1 shows the experimental comparisons and analysis of the point matching methods, which include SIFT [10], FAST [9], ORB [43], TILDE [44], Superpoint [15], Pop-net [32], and KPop-net (keypoint response constraints Pop-net).

From the RPR and QR values, it can be seen that each point matching method can detect enough keypoints for feature point matching. In the cross-view dataset, the Kpop-net method shows relatively high performance in terms of RR and AR. Especially from the AR value, it can be seen that SIFT, FAST, and TILDE do not perform well in the task of cross-view matching of satellite-guided airborne objects. ORB, as a manual point matching method, has high accuracy and efficiency, and therefore is widely used in multiple application fields. Superpoint and Pop-net are the most representative point matching methods, with Superpoint having the highest accuracy and Pop-net being slightly worse. Kpop-net is the method studied in this paper, and it is more effective than the Pop-net method, as shown in the table. From the runtime RT, it can be seen that ORB and Superpoint have better efficiency, while the efficiency of other algorithms is relatively high. Since Kpop-net introduces complex attribute constraints, it leads to a higher runtime, but accuracy and time are usually inversely proportional, and to ensure accuracy, efficiency is often sacrificed.

Table 1. Comparison of point matching on cross-view dataset

Method	RPR	RR	AR	QR	RT
SIFT [10]	0.747	0.0008	0.075	0.010	0.12
FAST [9]	0.716	0.0029	0.076	0.006	3.17
TILDE [41]	0.298	0.0002	0.005	0.001	5.07
ORB [42]	0.357	0.0211	0.196	0.001	0.03
Superpoint [15]	0.522	0.0189	0.200	0.002	0.37
Pop-net [32]	0.571	0.0079	0.167	0.003	0.79
KPop-net	**0.508**	**0.0134**	**0.187**	**0.002**	**0.72**

6.4.2 Results and Analysis of Set Matching Experiment

This article presents an experimental comparison and analysis of the set matching method in a dataset with large-scale morphological changes. The results of the airborne object matching experiment are shown in Table 2. From the RMSE-DI, it can be seen that the SIFT [10] and ORB [43] methods for airborne object matching performance are inferior. Corresponding to the high performance of the Superpoint [44], the set matching results are higher, and it has the highest matching effect among all algorithms. From the results of DEMONS [45], LDDMM [25], ICP [46], and CPD [47], it can be seen that dense diffeomorphism has a positive effect on airborne object matching, but it still cannot surpass the Superpoint and KDM set matching methods. KDM uses a set matching method based on keypoint-constrained diffeomorphism, which can better fit the spatial transformation relationship between satellite reference images and airborne object images. From Table 2, it can be seen that due to the use of diffeomorphism in the KDM method, its performance is higher than that of the Kpop-net object matching method, which also shows that the set matching studied in this article effectively improves intra-class and inter-class distinguishability.

Table 2. Comparison of set matching on large-scale difference dataset

Matching method	RMSE-DI
SIFT [10]	0.4815
ORB [43]	0.3977
Superpoint [15]	0.3525
DEMMONS [45]	0.5218
LDDMM [25]	0.1044
ICP [46]	0.3180
CPD [47]	0.0407
KPop-net	**0.2339**
KDM	**0.2439**

6.4.3 Results and Analysis of Domain Matching Experiments

This section presents an experimental comparison and analysis of domain matching methods on a dataset featuring ground cover change. The compared domain matching methods are shown in Table 3. Among them, BBS [48] and FAST-Match [28] are earlier domain matching methods, TBMR [49] and MSER [29] are region segmentation methods, and KDMM (keypoint-constrained diffeomorphism matching based on MSER) is the domain matching method studied in this paper, which is based on effective region segmentation. This method uses joint probability statistics and watershed maximum stable extreme region methods to perform region segmentation on satellite reference images and airborne object images, and then completes the domain matching process.

As shown in Table 3, the traditional BBS and FAST-Match methods use a sliding window to extract airborne object image blocks and perform template matching with satellite reference images. They do not have better performance for the satellite image guided airborne object matching task. From the RRMSE-DI, it can be seen that these two algorithms do not have higher intra-class and inter-class discrimination. Since TBMR and MSER use region detection methods to perform region segmentation on reference images, these two methods have higher intra-class and inter-class discrimination. KDMM is an improved algorithm based on MSER, and it has the highest performance, which demonstrates that domain matching methods with effective region segmentation can improve the performance of object matching.

Table 3. Comparison of set matching on large-scale difference dataset

Matching method	RMSE-DI
BBS [48]	0.0352
FAST-Match [28]	0.1203
TBMR [49]	0.0254
MSER [35]	0.3762
KDMM	**0.3977**

6.4.4 Results and Analysis of Object Matching Experiments

To verify the performance of the object matching method, we used ROC curves for analysis on a validation dataset that includes viewpoint changes, large-scale morphological changes, and ground cover changes. Existing domain matching methods were also compared. As shown in Figure 10, on the validation dataset, the traditional method, SIFT [10], ORB [43], DEMONS [45] have the relative lower performance. LDDMM [25] has the higher performance them the other traditional methods, which illustrates its scale invariance. For the feature matching based on deep learning method, Fast-RCNN [50] have the lowest performance, Strong-CNN [37], OriCNN [51], VIGOR [42], CVM-net [52] have a better performance, and our proposed the most performance. The AUC of Kpop-net, KDM, KDMM is 0.447, 0.577 and 0.679. The KDMM object matching method using improved MSER for region segmentation has the highest performance. Therefore, fusing point matching, set matching, and domain matching to complete the object matching task is of great research significance. It can better solve the problems of cross-view images, large-scale deformation, and ground cover changes in airborne object positioning tasks. Compared with other object matching methods, using the point–set–domain object matching method for object localization has significant advantages.

7 Conclusion

This article introduces the basic architecture of a three-level image object matching method for UAV target localization. Innovation is proposed for the three stages in the architecture—point matching, set matching, and domain matching—based on keypoint response constraints, diffeomorphism, and region partition, respectively. In point matching, the method considers the sparsity, repeatability, and distinguishability of keypoints, and makes novel use of the Superpoint keypoint response to construct the loss function of a convolutional neural network, which

effectively improves the point matching performance and solves the viewpoint difference problem. In set matching, the method uses the diffeomorphism set matching method based on corresponding keypoint set constraints to determine the spatial transformation relationship between two images, and improves the algorithm's efficiency through a static velocity field, effectively solving the problem of large-scale deformation. In domain matching, the method innovatively proposes a maximum stable extreme region detection method combining probability statistics and watershed, uses this method to divide the two images into multiple sub-regions, and uses a similarity calculation method based on diffeomorphism to determine the location of the target in the UAV image. The three methods are experimentally validated against existing point, set, domain, and object matching methods. Compared with existing object matching methods, the point–set–domain image object matching method improves performance by an average of 12%.

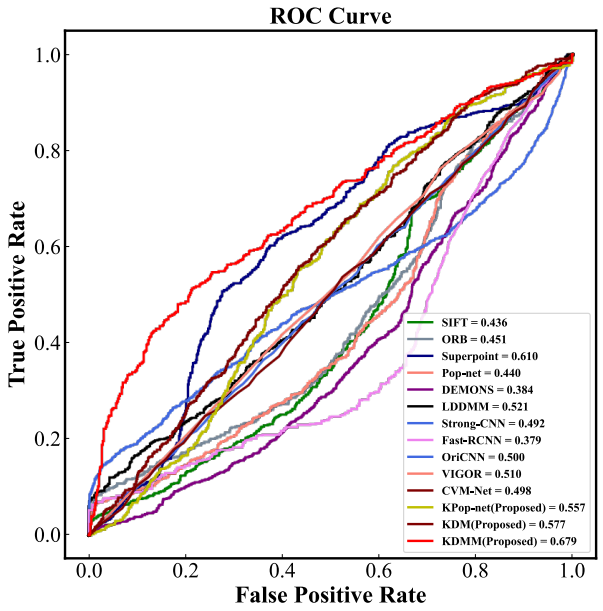

Figure 10. ROC curves of object localization on University1652

Acknowledgements

This work is supported by the National Natural Science Foundation of China (6227010741), the Natural Science Foundation of Heilongjiang (LH2022F052), Doctoral Project of Jiamusi University (JMSUBZ2022-13), Heilongjiang Provincial Department of Education basic research funds basic research project (No. 2023-KYYWF-0580), Jiamusi University "East Pole" Academic Team (DJXSTD202417), and Heilongjiang Provincial Key Research and Development Program (Innovation Base (JD24A014)).

References

[1] J. Lai, C. Yuan, P. Lyu, J. Liu, H. He, Unmanned System Visual/LiDAR Perception and Navigation Technology Independent of GNSS, *Navigation Positioning and Timing*, Vol. 8, No. 3, pp. 1-14, April, 2021.

[2] S. Chen, *Design and Simulation of Hitting Process of Reconnaissance and Combat Integrated Drones on Time-sensitive Targets*, Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2019.

[3] S. Yang, H. Cheng, T. Li, H. Zhao, UAV Reconnaissance Images Accurate Targeting Method Based on Image Registration, *Infrared Technology*, Vol. 39, No. 6, Article No. 529, June, 2017.

[4] X. Chen, G. Li, L. Zhao, Research on UCAV Game Strategy of Cooperative Air Combat Task, *Fire Control and Command Control*, Vol. 43, No. 11, pp. 17-23, November, 2018.

[5] P. Zhu, *Research on UAV Autonomous Navigation Technology for Bridge Detection*, Master's Thesis, Chongqing University, Chongqing, China, 2017.

[6] Z. Lou, *Research on UAV Visual Positioning Algorithm in Autonomous Inspection of Photovoltaic Farm*, Master's Thesis, Zhejiang University, Zhejiang, China, 2019.

[7] Z. Jiao, *Research on UAV system for Forest Fire Prevention*, Master's Thesis, Xi'an University of Technology, Xi'an, China, 2019.

[8] H. Chris, M. Stephens, A combined corner and edge detector, *Alvey vision conference*, Manchester, UK, 1988, pp. 147-151.

[9] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision*, Graz, Austria, 2006, pp. 430-443.

[10] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, Vol. 60, No. 2, pp. 91-110, November, 2004

[11] Y. Ke, R. Sukthankar, PCA-SIFT: A more distinctive representation for local image descriptors, *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004*, Washington, DC, USA, 2004, pp. 1-8.

[12] B. Herbert, T. Tuytelaars, L. V. Gool, Surf: Speeded up robust features, In: A. Leonardis, H. Bischof, A. Pinz (Eds.), *Lecture notes in computer science*, Vol. 3951, Springer, Berlin, Heidelberg, 2006, pp. 404-417.

[13] L. Liu, F. Peng, Y. Tian, Y. Xu, K. Zhao, Fast image matching for localization in deep-sea based on the simplified SIFT (scale invariant feature transform) algorithm, *Proc. SPIE 6795, Second International Conference on Space Information Technology*, Vol. 6795, pp. 705-711, November, 2007.

[14] H. Liu, Q. Zhang, B. Fan, Z. Wang, J. Han, Features combined binary descriptor based on voted ring-sampling pattern, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30, No. 10, pp. 3675-3687, October, 2020.

[15] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, Salt Lake City, UT, USA, 2018, pp. 224-236.

[16] Y. Liu, Y. Zhang, Y. Wang, Y. Zhang, J. Tian, Z. Shi, J. Fan, Z. He, Sap-detr: Bridging the gap between salient points and queries-based transformer detector for fast model convergency, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 15539-15547.

[17] G. Zhang, Z. Luo, Z. Tian, J. Zhang, X. Zhang, S.

Lu, Towards Efficient Use of Multi-Scale Features in Transformer-Based Object Detectors, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 6206-6216.

[18] Y. Xie, J. Xing, G. Liu, J. Lan, Y. Dong, Real-time Reconstruction of Unstructured Scenes Based on Binocular Vision Depth, *Journal of Internet Technology*, Vol. 20, No. 5, pp. 1611-1623, September, 2019.

[19] Y. Ma, *Registration, Recognition and Labeling in 3D Point Clouds*, Ph. D. Thesis, University of Defense Technology, Hunan, China, 2018.

[20] P. J. Besl, N. D. McKay, A method for registration of 3-D shapes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 2, pp. 239-256, February, 1992.

[21] B. Zitova, J. Flusser, Image registration methods: a survey, *Image and Vision Computing*, Vol. 21, No. 11, pp. 977-1000, October, 2003.

[22] W. Lian, L. Zhang, M.-H. Yang, An efficient globally optimal algorithm for asymmetric point matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 7, pp. 1281-1293, July, 2017.

[23] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, *Computer Vision and Image Understanding*, Vol. 89, No. 2-3, pp. 114-141, February-March, 2003.

[24] J. Zhong, A. Qiu, Multi-manifold diffeomorphic metric mapping for aligning cortical hemispheric surfaces, *NeuroImage*, Vol. 49, No. 1, pp. 355-365, January, 2010

[25] M. Tan, A. Qiu, Multiscale frame-based kernels for large deformation diffeomorphic metric mapping, *IEEE Transactions on Medical Imaging*, Vol. 37, No. 10, pp. 2344-2355, October, 2018.

[26] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, T. Mansi, Learning a probabilistic model for diffeomorphic registration, *IEEE Transactions on Medical Imaging*, Vol. 38, No. 9, pp. 2165-2176, September, 2019.

[27] M. Hernandez, Band-limited Stokes Large Deformation Diffeomorphic Metric Mapping, *IEEE Journal of Biomedical and Health Informatics*, Vol. 23, No. 1, pp. 362-373, January, 2019.

[28] S. Korman, D. Reichman, G. Tsur, S. Avidan, Fast-match: Fast affine template matching, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, 2013, pp. 2331-2338.

[29] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, *Image and Vision Computing*, Vol. 22, No. 10, pp. 761-767, September, 2004.

[30] A. Samat, C. Persello, S. Liu, E. Li, Z. Miao, J. Abuduwaili, Classification of VHR multispectral images using extratrees and maximally stable extremal region-guided morphological profile, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 11, No. 9, pp. 3179-3195, September, 2018.

[31] K. Ning, Z. Su, Z. Zhang, G. Kim, An Image Reconstruction Algorithm Based on Frequency Domain for Deep Subcooling of Melt Drops, *Journal of Internet Technology*, Vol. 22, No. 6, pp. 1273-1285, November, 2021.

[32] P. Yan, Y. Tan, Y. Tai, D. Wu, H. Luo, X. Hao, Unsupervised learning framework for interest point detection and description via properties optimization, *Pattern Recognition*, Vol. 112, Article No. 107808, April, 2021.

[33] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, *Computer Vision--ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, 2016, pp. 499-515.

[34] A. Pai, S. Sommer, L. Sørensen, S. Darkner, J. Sporring, M. Nielsen, Kernel bundle diffeomorphic image registration using stationary velocity fields and wendland basis functions, *IEEE Transactions on Medical Imaging*, Vol. 35, No. 6, pp. 1369-1380, June, 2016.

[35] W. W. Franklin, *The theoretical foundation of the MSER algorithm*, Ph.D. Thesis, University of Virginia, Charlottesville, VA, 2009.

[36] Z. Zheng, Y. Wei, Y. Yang, University-1652: A multi-view multi-source benchmark for drone-based geo-localization, *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA, USA, 2020, pp. 1395-1403.

[37] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 10, pp. 1615-1630, October, 2005.

[38] Z. Zhu, K. Davari, Comparison of local visual feature detectors and descriptors for the registration of 3D building scenes, *Journal of Computing in Civil Engineering*, Vol. 29, No. 5, Article No. 04014071, September, 2015.

[39] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A Comparison of Affine Region Detectors, *International journal of computer vision*, Vol. 65, No. 1-2, pp. 43-72, November, 2005.

[40] T. Tuytelaars, K. Mikolajczyk, Local Invariant Feature Detectors: A Survey, *Foundations and Trends in Computer Graphics and Vision*, Vol. 3, No. 3, pp. 177-280, January, 2008.

[41] X. Liu, J. Li, J. Pan, Feature Point Matching Based on Distinct Wavelength Phase Congruency and Log-Gabor Filters in Infrared and Visible Images, *Sensors*, Vol. 19, No. 19, Article No. 4244, October, 2019.

[42] S. Zhu, T. Yang, C. Chen, VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, 2021, pp. 3640-3649.

[43] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, *IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 6-13.

[44] Y. Verdie, K. Yi, P. Fua, V. Lepetit, Tilde: A temporally invariant learned detector, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, MA, USA, 2015, pp. 5279-5288.

[45] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Symmetric log-domain diffeomorphic registration: A demons-based approach, in: D. Metaxas, L. Axel, G. Fichtinger, G. Székely (Eds.), *Medical Image Computing and Computer Assisted Intervention*, Vol. 5241, Springer, Berlin, Heidelberg, 2008, pp. 754-761.

[46] B. B. Avants, C. L. Epstein, M. Grossman, J. C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, *Medical Image Analysis*, Vol. 12, No. 1, pp. 26-41, February, 2008.

[47] A. Myronenko, X. Song, Point Set Registration: Coherent Point Drift, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 12, pp. 2262-2275, December, 2010.

[48] S. Oron, T. Dekel, T. Xue, W. T. Freeman, S. Avidan, Best-Buddies Similarity - Robust Template Matching using Mutual Nearest Neighbors, *IEEE transactions on pattern*

analysis and machine intelligence, Vol. 40, No. 8, pp. 1799-1813, August, 2018.
[49] Y. Xu, P. Monasse, T. Géraud, L. Najman, Tree-Based Morse Regions: A Topological Approach to Local Feature Detection, *IEEE Transactions on Image Processing*, Vol. 23, No. 12, pp. 5612-5625, December, 2014.
[50] R. Girshick, Fast r-cnn, *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 2015, pp. 1440-1448.
[51] L. Liu, H. Li, Lending orientation to neural networks for cross-view geo-localization, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, USA, 2019, pp. 5624-5633.
[52] S. Hu, M. Feng, R. M. H. Nguyen, G. H. Lee, Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, 2018, pp. 7258-7267.

Biographies

Xiaomin Liu received Bachelor's Degree from school of Information and Electronic Technology, Jiamusi University, in 2003, Master's Degree from the school of computer science, Heilongjiang University, in 2009, and Doctor's Degree from School of Electronics and Information Engineering, Harbin of Institution Technology, in 2022. Her research focuses on image process, pattern recognition, image matching and artificial intelligence.

Runqi Zhao received his B.S. degree from University of Electronic Science and Technology of China, Zhongshan Institute, in 2021. He is currently studying for a Master's degree in the School of Information and Electronic Technology, Jiamusi University, majoring in computer application technology, and researching in image processing and artificial intelligence.

Jun-Bao Li received Bachelor Degree from Instrumentation Science and Engineering speciality, Harbin Institution of Technology, in 2002, and Master's Degree from the Mechanics and Mechanics Foundation speciality, in 2004. Finally, he received Doctor's Degree from Instrumentation Science and Engineering speciality. He is the tenure professor in School of Electronics and Information Engineering.

Jeng-Shyang Pan received the B. S. degree in Electronic Engineering from the National Taiwan University of Science and Technology in 1986, the M. S. degree in Communication Engineering from the National Chiao Tung University, Taiwan in 1988, and the Ph.D. degree in Electrical Engineering from the University of Edinburgh, U.K. in 1996.

Huaqi Zhao received Bachelor's Degree from school of Industrial Automation, Harbin Engineering University, in 1999, Master's Degree from the school of Mechanical and Electrical Engineering, Northeast Forestry University, in 2005, and Doctor's Degree from school of Mechanical and Electrical Engineering, Northeast Forestry University, in 2009. His research focuses on pattern recognition and artificial intelligence.