Intelligent Matching Method for Heterogeneous Remote Sensing Images Based on Style Transfer

Jiawei Zhao, Dongfang Yang, Yongfei Li, Peng Xiao, and Jinglan Yang

***Abstract*—Intelligent matching of heterogeneous remote sensing images is a common basic problem in the field of intelligent remote sensing image processing. Aiming at the difficulty of matching satellite-aerial remote sensing images, this article proposes an intelligent matching method for heterogeneous remote sensing images based on style transfer. First, based on the idea of image style transfer of a generative adversarial networks, this method improves the conversion effect of the model on heterogeneous images by constructing a new generative network loss function and converts satellite images into aerial images. Then, the advanced deep learning-based matching algorithms D2-Net and LoFTR are used to achieve matching between the generated aerial image and the original aerial image. Finally, this transformation relationship is mapped to the corresponding satellite–aerial image pair to obtain the final matching result. The image style transfer experiments and the matching experiments we carry out under different test datasets show that the smooth cycle-consistent generative adversarial networks proposed in this article can effectively reduce the complexity of the algorithm and improve the quality of image generation. In addition, combining it with deep learning-based feature-matching methods can effectively improve the accuracy and robustness of the matching algorithm. Our code and data can be found at: https://gitee.com/AZQZ/intelligent-matching.**

***Index Terms*—Heterogeneous remote sensing images, image style transfer, intelligent matching.**

I. Introduction

With the rapid development of aerospace technology, remote sensing images have started to be widely used in the fields of environment, transportation, resources, national defense, etc. This has led to research on intelligent processing of remote sensing images, cognitive navigation, intelligent aircraft, target positioning and other fields. How to use platforms such as unmanned aerial vehicle (UAVs) to intelligently match the collected heterologous remote sensing images is a key common problem in the above fields.

Traditional image matching algorithms can be divided into feature-based matching methods, region-based matching methods, and combination methods based on region and feature [1]. The feature-based matching methods such as scale-invariant feature transform (SIFT) [2], speeded-up robust features (SURF) [3], and radiation-variation insensitive feature transform (RIFT) [4] achieve matching by extracting local invariant features such as points [5], lines [6], and surfaces [7] of images. This method has a small amount of computations and good robustness to various changes. However, it does not match well for large changes in image appearance, large angle transformation, and a complex model composed of many parameters. Moreover, feature extraction is very complex and only shallow features can be extracted [8], making it difficult to extract deeper and more expressive features. On the other hand, the region-based matching methods such as sum of squared differences [9], normalized cross-correlation [10], and mutual information [11] usually use the grayscale and phase information of the image. They adjust the parameters of the optimized transformation model according to the preestablished similarity measure, and regard the matching problem as an optimization problem. The principle of this method is simple, but it is computationally intensive and time-consuming, and it is difficult to ensure real-time performance in practical applications. Moreover, most similarity measurement methods have many local minima, and it is difficult to obtain a globally optimal solution [12]. Region-based and feature-based methods have their own advantages and disadvantages. By extracting the reliable matching between two images and finding the corresponding relationship between them, they can deal with rotation, translation, scale difference, and geometric distortion, the similarity measurement can effectively eliminate the nonlinear radiation difference. Based on the above reasons, researchers have combined these two methods to achieve more accurate matching [13], [14]. The above methods have achieved good results in matching homologous images. However, for heterologous images, due to the changing functions and types of image acquisition equipment, as well as the imaging principles and spatial positions of each sensor, the difference between the collected images is greater. This makes it difficult for traditional homologous image matching methods to be directly applied to heterologous images. It is even more difficult to achieve heterologous remote sensing image matching with complex features.

In recent years, machine learning methods, especially deep learning methods, have made significant progress in the field of computer vision. Deep learning can analyze and process Big Data through its deep multilevel structure and automatically learn the characteristics of specific objects from the training data to accurately capture the characteristics of target objects

Manuscript received 4 June 2022; revised 20 July 2022; accepted 6 August 2022. Date of publication 10 August 2022; date of current version 26 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61673017 and in part by the Natural Science Foundation of Shaanxi Province under Grants 2021JQ-702 and 2019JM-434. *(Corresponding author: Dongfang Yang.)*

The authors are with the Department of Automatic Control, Xi'an Research Institute of Hi-tech, Xi'an 710025, China (e-mail: huodazhaojiawei@163.com; yangdf301@163.com; lyfei314@163.com; xiaopeng2dd@163.com; diamond_yang.sjz@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3197748

and understand the contents of images. With the continuous development of deep learning technology, an increasing number of neural networks are used in the field of image processing. The convolutional neural networks (CNNs) proposed by LeCun et al. [15], the fully convolutional networks (FCN) proposed by Shelhamer et al. [16], and the siamese networks proposed by Chopra et al. [17] are commonly used network structures in image matching. The generative adversarial networks (GANs) proposed by Goodfellow et al. [18] play an important role in image generation, style transfer, and other visual images, and a large number of improved networks such as deep convolutional GANs [19], Wasserstein GANs [20], cycle-consistent GANs (CycleGAN) [21], dual-path network-CycleGAN [22], etc. continue to emerge. Li et al. [23] used deep translation to convert optical images into synthetic aperture radar (SAR) images and Song et al. [24] used GAN to convert optical images into maps, reflecting the feasibility and practicality of GAN application to remote sensing image processing.

Based on the above analysis, taking the satellite-aerial remote sensing images as an example, the problem of large differences in their features can be solved in two ways: One is to design a more expressive feature extraction or description network and the other is to reduce the difference between two images based on GANs. This article chooses the second approach and then uses the state-of-the-art deep learning-based matching algorithm to achieve the matching of satellite-aerial remote sensing images.

The main contributions of this article are as follows.

1) A general matching framework is proposed to match satellite images and aerial images with large feature differences. The core idea is to first convert satellite images and aerial images into the same feature domain, and then input them to a feature-matching network to achieve image matching.

2) To reduce the feature differences between satellite images and aerial images, a smoothing loss function is proposed, which can accelerate the convergence of the network, improve the stability of the network model, and improve the quality of image generation.

3) This work proves that style transfer preprocessing methods can improve matching performance. Moreover, it is demonstrated that a better style transfer model can improve the generated images, thus improving the matching results.

The rest of this article is structured as follows. Section II describes the proposed method. Section III presents the experimental results and analysis. Finally, Section IV concludes this article.

II. Methodology

The purpose of matching is to obtain the best transformation matrix from aerial image B to satellite image A. In this section, we propose a novel method for matching heterogenous remote sensing images. The proposed method includes the following two aspects: 1) The idea of style transfer based on GAN is applied to image transfer to reduce the imaging difference between satellite image A and aerial image B. 2) The aerial image A' generated by style transfer is matched with the original aerial image B and then the corresponding transformation relationship is mapped to the original image pair B–A to obtain the final matching result. The basic pipeline of the method is shown in Fig. 1.

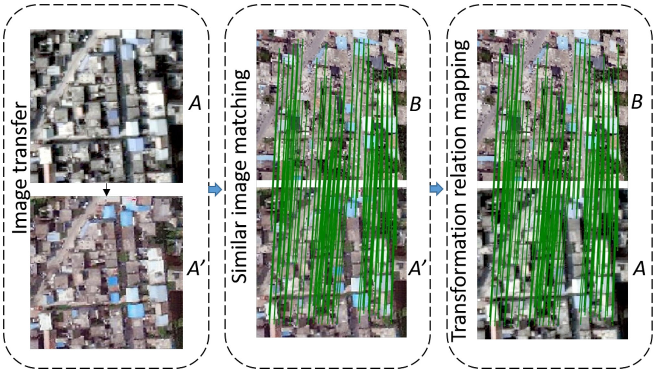


Fig. 1. Pipeline of the proposed method. The green lines are the correspondence of matching point pairs.

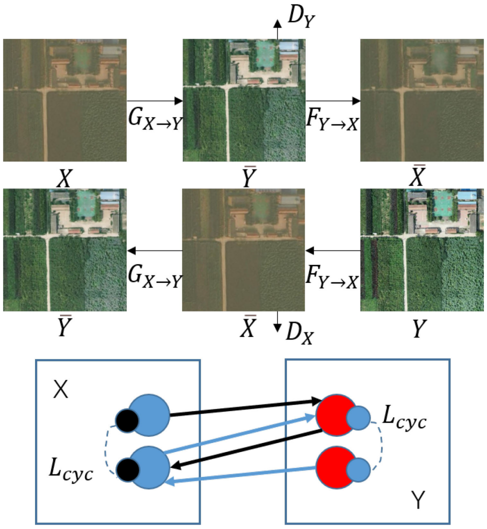


Fig. 2. Structure diagram of the circulating circuit. The G and F represent the generators of domains X and Y, respectively. D_X and D_Y are the discriminators of domains X and Y. With the forward adversarial loss, backward adversarial loss, and cycle-consistent loss, the quality of generated aerial images can be improved during the cycle.

A. Image Transfer Method

This article uses CycleGAN for image transfer. CycleGAN is an improved variant of the GAN. It uses two symmetrical GANs to form a ring network that consists of two generators and two discriminators. The purpose of this network is to realize the mutual conversion between the source domain X and the target domain Y. Its structure diagram is shown in Fig. 2. In the figure, X is the image from the X domain, and Y is the image from the Y domain. By training the model with two mappings $(G_{X\to Y}, F_{Y\to X}), G_{XY}(x)$ is made to infinitely approximate the image of the target domain Y, and then the detailed information of the generated image is further optimized by the discriminator. CycleGAN has a unique cycle consistent adversarial learning capability, so that the input image can still be reconstructed after passing through two generators in sequence. This idea not only fits the style distribution of the target domain image but also retains the content characteristics of the source domain image. This effectively reduces the addition of wrong information and useless information during the conversion process. The introduced cycle consistency constraint prevents the generators (G_{XY}, F_{YX}) from contradicting each other, alleviates the problems of model collapse and gradient disappearance, enhances the overall conversion effect between different domains, realizes the bidirectional conversion of heterogeneous image styles, and makes the conversion model training more stable. It makes it convenient to carry out the matching work between the aerial image and satellite image. However, CycleGAN uses the mean square error (MSE) as the loss function of the generative network, which will cause outliers to obtain higher weights at the expense of other normal samples, thus reducing the performance of the overall model. Therefore, we designed the L1 smooth loss function to replace the MSE to reduce the sensitivity of the CycleGAN model to outliers, accelerate the network convergence, and improve the stability of the network. The resulting model is called smooth cycle-consistent GANs (SCycleGAN). For outliers, when the network gradient update is greater than 1, the smoothing function will reduce the error, reduce the sensitivity to outliers, and improve the robustness of the network. As the error decreases, when the gradient update is less than 1, the property of the sum of squares operation makes the gradient smoother near zero, and the network converges faster. The formula is as follows:

$$\mathrm{SmoothL1}\left(x_i, y_i\right) = \frac{1}{n} \sum_i z_i$$

$$z_i = \begin{cases} 0.5\left(x_i - y_i\right)^2, \left(\left|x_i - y_i\right| < 1\right) \\ \left|x_i - y_i\right| - 0.5, \left(\left|x_i - y_i\right| \geq 1\right) \end{cases}. \tag{1}$$

The forward adversarial loss, backward adversarial loss, cycle-consistent loss, and objective function of SCycleGAN are consistent with CycleGAN as shown below:

$$L_{\mathrm{GAN}}\left(G, D_Y, X, Y\right) = E_{y \sim p_{\mathrm{data}}(y)}\left[\log D_Y(y)\right] + E_{x \sim p_{\mathrm{data}}(x)}\left[\log\left(1 - D_Y(G(x))\right]\right. \tag{2}$$

$$L_{\mathrm{GAN}}\left(F, D_X, X, Y\right) = E_{x \sim p_{\mathrm{daa}}(x)}\left[\log D_X(x)\right] + E_{y \sim p_{\mathrm{data}}(y)}\left[\log\left(1 - D_X(F(y))\right]\right. \tag{3}$$

$$L_{\mathrm{cyc}}(G, F) = E_{x \sim p_{\mathrm{data}}(x)}\left[\|F(G(x)) - x\|_1\right] + E_{y \sim p_{\mathrm{dta}}(y)}\left[\|G(F(y)) - y\|_1\right] \tag{4}$$

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} \left[L_{\mathrm{GAN}}\left(G, D_Y, X, Y\right) + L_{\mathrm{GAN}}\left(F, D_X, X, Y\right) + \lambda L_{\mathrm{cyc}}(G, F)\right]. \tag{5}$$

In the above equations, *X* and *Y* are the real images in the source and target domains, $G(x)$ and $F(y)$ are the generated images, $p_{\mathrm{data}}(x)$ and $p_{\mathrm{data}}(y)$ are the distributions of the real images in the target domain, $\sim$ is the obedience relation, E is the expectation function, and λ controls the relative importance of the two objectives.

1) Forward Adversarial Loss: We aim to convert *X* to *Y*. Therefore, the objective is to learn the mapping from *X* to *Y* and set this mapping on $G(x)$, which corresponds to the generator in the GAN. For generated images, we also need discriminator D_Y to determine whether it is a real image, to compose an adversarial generative network. Generator $G(x)$ aims to minimize object function against an adversary discriminator D_Y, which tries to maximize the objective. We express the objective as (2).

2) Backward Adversarial Loss: The backward adversarial loss has the same form to the forward adversarial loss as (3).

3) Cycle-Consistent Loss: In practice, it is difficult to train the whole network by using adversarial loss alone. The reason is that the mapping $G_{X \rightarrow Y}$ can completely make *x* in domain *X* the same image in domain *Y*, which invalidates the loss. Therefore, the significance of cycle-consistent loss is to assume another mapping $F_{Y \rightarrow X}$, which can convert *x* in domain X to image *y* in domain Y. We call $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ forward cycle-consistency and $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ is backward cycle-consistency. We express the objective as (4).

4) Objective Function: By combining forward adversarial loss, backward adversarial loss, and cycle-consistent loss, objective function can be obtained as (5).

B. Image Matching Method

Images A' and *B* are similar in texture and imaging features. Therefore, they are less difficult to match. The problem of difference between images is solved. Then, this article chooses two more advanced network models, D2-Net [25] and LoFTR [26], to achieve matching.

1) D2-Net: Traditional keypoint detection first generates feature descriptors and then uses some postprocessing methods to find keypoints according to these feature descriptors. Because descriptors are obtained from larger image regions and detection points are obtained from low-level information of small regions (e.g., corner points, etc.), this results in unstable detection results. D2-Net, on the other hand, performs key feature detection directly from feature descriptors. In other words, the feature detection module is highly coupled with the description module, and thus these descriptors are well suited for matching. The pipeline of the matching method is shown in Fig. 3.

a) Network architecture: D2-Net selects the classic VGG16 architecture and improves it [27]. To make the feature points abstract enough and obtain high localization accuracy, the feature network discards the last convolutional layer of VGG16 and selects the last convolutional layer $(\mathrm{Conv}\ 4\_3)$ in the middle layer 4 as the feature map for key point extraction. The network architecture is shown in Fig. 3. The feature map is the output of the original image after multilayer convolution and pooling of the CNN network, and the resolution generally decreases. To maintain the resolution of the feature map, the sliding step size of the last pooling layer window is changed from 2 pixels to 1 pixel, and the pooling method also replaces max pooling with average pooling. The three convolutions $(\mathrm{Conv}\ 4\_1 \text{ to } \mathrm{Conv}\ 4\_3)$ in the fourth layer use dilated convolutions with a dilation rate of 2, which can expand the receptive field, improve the generalization

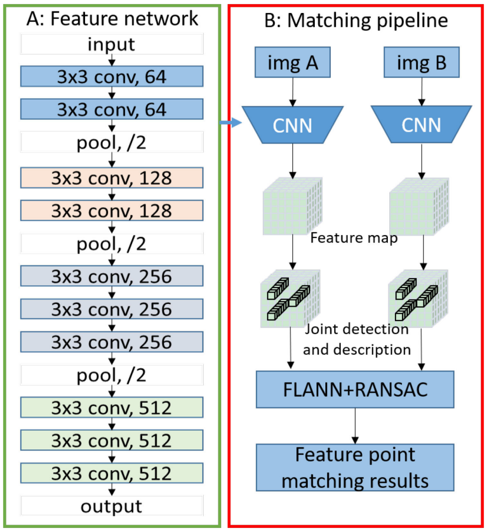


Fig. 3. Matching algorithm pipeline of D2-Net.

ability of feature expression, and facilitate the invariant expression of heterologous image features. With the improved VGG16 architecture, the output feature map is expanded from 1/8 to 1/4 of the original image, and the localization accuracy is doubled.

b) Matching strategy: The features extracted by the feature network are often too dense, and most of the features are not significant enough. Hence, D2-Net proposes to detect key features directly from the feature descriptor. For the D2 feature map, if the location (i, j) is a keypoint, the final detection result of this pixel location from the channel dimension should take the value corresponding to the channel with the largest detector response value so that the channel is selected. In addition, from the spatial dimension, the location must be a local maximum on the 2-D feature map of the channel defined as follows:

$$k = \arg\max D_{ij}^{t}, D_{ij}^{k} \text{ is a local max in } D^{k} \tag{6}$$

where (i, j) is the detection pixel, (i, j) is the feature value of the kth layer, and D_{ij}^{k} is the feature value at the pixel (i, j) on the feature map.

To make (6) differentiable, define

$$\alpha_{ij}^{k} = \frac{\exp\left(D_{ij}^{k}\right)}{\sum_{(i',j')\in N(i,j)} \exp\left(D_{i'j'}^{k}\right)} \tag{7}$$

$$\beta_{ij}^{k} = D_{ij}^{k} \Big/ \max_{t} D_{ij}^{t} \tag{8}$$

$$\gamma_{ij} = \max_{k} \left(\alpha_{ij}^{k} \beta_{ij}^{k}\right) \tag{9}$$

$$s_{ij} = \gamma_{ij} \Big/ \sum_{(i',j')} \gamma_{i'j'} \tag{10}$$

where α_{ij}^{k} is the spatial response score, $N(i, j)$ is the set of nine neighborhoods of pixel (i, j), β_{ij}^{k} is the channel response weight, γ_{ij} is the score of each pixel being a keypoint, and s_{ij} is the score of each pixel that is a keypoint after normalization.

After detecting and extracting the key features of the image pairs, to obtain more accurate keypoint locations, the subpixel-level localization accuracy is obtained by using the feature map local interpolation encryption method with the SIFT algorithm, while the descriptors are obtained by bilinear interpolation in the neighborhood. Finally, the combination of fast library for approximate nearest neighbors and random sample consensus (RANSAC) constraints is used to eliminate the false matching points to obtain the final matching result.

c) Loss function: D2-Net adopts the triplet margin ranking function as the loss function. This is because in the feature detection process, it is desired that the feature points can adapt to the effects of different ambient light intensities and geometric differences. At the same time, it is desired that the feature vectors be as unique as possible in order to find homonymous image points. To address this problem, the triplet margin ranking loss function enhances the uniqueness of related descriptors by penalizing any irrelevant descriptors that lead to false matches. In addition, to achieve the repeatability of the detection features, the detection scores are added to the loss function, as shown in (11).

$$L\left(I_1, I_2\right) = \sum_{c\in C} \frac{s_c^{(1)} s_c^{(2)}}{\sum_{q\in C} s_q^{(1)} s_q^{(2)}} m(p(\mathrm{c}), n(\mathrm{c})) \tag{11}$$

where $s_c^{(1)}$ and $s_c^{(2)}$ are the feature detection scores obtained at two points on image I_1, I_2, respectively, C is the set of all point-to-point correspondences on image I_1, I_2, and $p(\mathrm{c})$ and $n(\mathrm{c})$ are the positive pair distance and negative pair distance of the corresponding points, respectively.

The above loss function generates a weighted average of the margin terms m based on the detection scores of all matches. Therefore, to minimize the loss, the most relevant correspondences with lower margin terms will obtain higher relative scores and let correspondences with higher relative scores obtain similar descriptors different from the rest of the features, thereby improving the robustness of the matching.

2) LOFTR: LoFTR is an end-to-end feature-matching scheme that does not rely on keypoint detection. It uses self-attention and cross-attention mechanisms to build feature-by-feature coarse matching directly on low-resolution feature maps and then optimizes on high-resolution feature maps to find subpixel-level fine matching. Compared with traditional methods, LoFTR uses a transformer to construct features with a global receptive field based on two images, which can construct accurate matches in low-texture areas. The pipeline of the matching method is shown in Fig. 4.

a) Matching strategy: LoFTR adopts a relatively simple network architecture, which uses a combination of ResNet-18 and FPN to extract coarse-level feature maps at 1/8 of the original image dimension and fine-level feature maps at 1/2 of the original image dimension. Then, the coarse-level feature maps extracted by the feature network are summed with their respective positional encodings and input to the transformer for coarse-level feature extraction. The transformer consists of several alternating self-attention and cross-attention layers, with the self-attention layer causing each point to focus on the

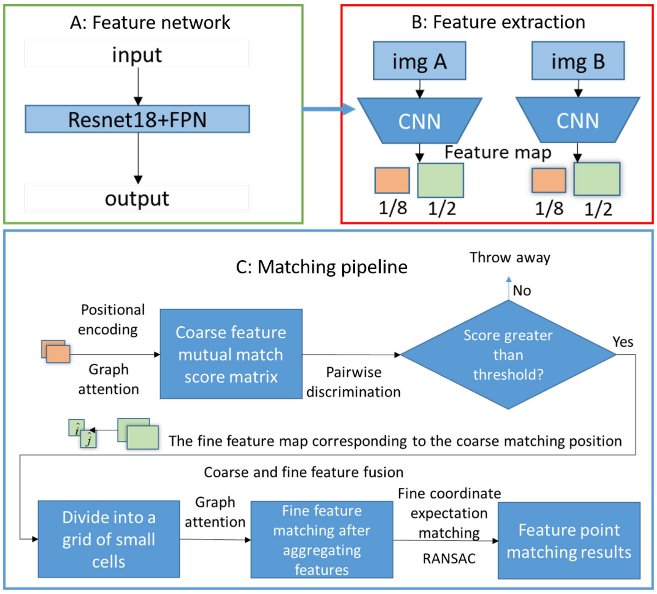

A: Feature network
input
Resnet18+FPN
output
B: Feature extraction
img A
img B
CNN
CNN
Feature map
1/8
1/2
1/8
1/2
C: Matching pipeline
Throw away
No
Positional encoding
Graph attention
Coarse feature mutual match score matrix
Pairwise discrimination
Score greater than threshold?
Yes
i
j
The fine feature map corresponding to the coarse matching position
Coarse and fine feature fusion
Divide into a grid of small cells
Graph attention
Fine feature matching after aggregating features
Fine coordinate expectation matching
RANSAC
Feature point matching results

Fig. 4. Matching algorithm pipeline of LoFTR.

association of all points around it and the cross-attention layer causing the point to focus on the association with all points on the other image. In the coarse-level matching process, the matching score matrix of all positions is first calculated using the product approach, and then the optimal match is calculated, either by the optimal transmission algorithm or the dual-softmax method. Then, some outlier matching point pairs are filtered out by the mutual nearest neighbor algorithm. In the fine-level matching process, the coarse-level matching result $(\tilde{i}, \tilde{j})$ is obtained by coarse-level matching, which is mapped to the corresponding fine-level feature map position $(\hat{i}, \hat{j})$, and two sets of local windows of size $\omega \times \omega$ are cropped (equivalent to cropping out the features at $\omega \times \omega$ positions), inputting them to the transformer of fine-level feature extraction to extract matching features and generating two transformed local feature maps $\hat{F}^A_{tr}(\hat{i})$ and $\hat{F}^B_{tr}(\hat{j})$ centered on $\hat{i}$ and $\hat{j}$, respectively. Then, the matching probability (i.e., similarity) between the central feature of $\hat{F}^A_{tr}(\hat{i})$ and all the features in $\hat{F}^B_{tr}(\hat{j})$ is calculated. Then, the probability distribution is calculated to determine the matching point position of subpixel accuracy in $\hat{F}^B_{tr}(\hat{j})$. Finally, the final matching results are obtained using the RANSAC constraint method.

b) Loss function: The loss function of LoFTR includes the negative log-likelihood loss of the coarse-level matching correlation score matrix as well as the loss of the fine-level matching coordinates. In the coarse-level matching layer, LoFTR uses a negative log-likelihood loss to supervise the dense confidence matrix obtained by the differentiable matching layer. That is, the negative log-likelihood loss for those that can establish ground-truth matches is minimized. Since the differentiable matching layer ensures that the gradients are efficiently passed back to all features, there is no need for error-matching supervision. In the fine-level matching layer, LoFTR calculates the variance $\sigma^2(\hat{i})$ of the heatmap generated for each point $\hat{i}$ to measure its uncertainty to improve the accuracy of the fine-level matching position. The formula is as follows:

$$L_c = -\frac{1}{M_c^{\text{gt}}} \sum_{(\tilde{i},\tilde{j})\in M_c^{\text{gt}}} \log P_c(\tilde{i},\tilde{j}) \tag{12}$$

$$L_f = -\frac{1}{|M_f|} \sum_{(\hat{i},\hat{j}')\in M_f} \frac{1}{\sigma^2(\hat{i})} \left\| \hat{j}' - \hat{j}'_{gt} \right\| \tag{13}$$

$$L = L_c + L_f \tag{14}$$

where L is the total loss, L_c is the coarse-level loss, and L_f is the fine-level loss. M_c^{gt} is the ground-truth coarse-level matching, defined as the mutual nearest neighbors of two sets of 1/8-resolution grids, P_c is the confidence matrix returned by the optimal transport layer or dual-softmax operator, M_f is the final fine-level matching, $\hat{j}'$ is the final fine-level matching position, and $\hat{j}'_{gt}$ is the ground-truth of the matching position corresponding to position $\hat{i}$.

III. Experimental Verification

The experiments were conducted on a hardware platform equipped with NVIDIA Quadro P4000 GPUs using Ubuntu 20.04 OS and is based on the open-source deep learning framework PyTorch, version 1.11.0, and CUDA version 10.2.

A. Original Image Data Introduction

The original image data are a set of satellite–aerial image pairs, and the image size is 5896×17204. Both satellite and aerial images are aligned to the EPSG:32649 GCS. Among them, the aerial image is taken by UAV, the shooting angle adopts the method of looking down, and the two-dimensional orthophoto image is generated in real time by using the Dajiang Zhitu software, with a spatial resolution of 0.25 m. The satellite image is sourced from the online map Google Satellite in GIS, acquired by the QGIS software and sampled to a spatial resolution of 0.5 m. The image data selected in this article include rural and urban areas, which have the advantages of high spatial resolution and rich types of features. The original image data are shown in Fig. 5.

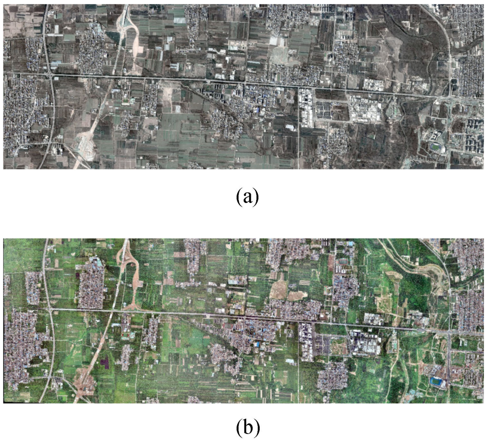
B. Evaluation Criterion

Here, the number of correct matching points (NCM), matching success rate (SR), and matching end point error (EPE) are used to evaluate the performance. The position of a feature point matched by the algorithm on the target image is (x_i, y_i), the corresponding feature point position on the reference image is (x'_i, y'_i), and its corresponding position after the ground-truth transformation of the homography matrix is $(\hat{x}_i, \hat{y}_i)$. Then, the judgment equations and EPE of NCM are as follows:

$$\text{Corr}(x) : \sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \leq \varepsilon \tag{15}$$

$$\text{EPE} = \frac{1}{n} \sum_{j=1}^{n} \sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{16}$$

NCM is the number of all matching points on the whole image that satisfy (15), which can reflect the robustness of the


(a)

(b)

Fig. 5. Original image data. (a) Satellite image. (b) Aerial image. There still exists small position offset (several pixels) between pairs.

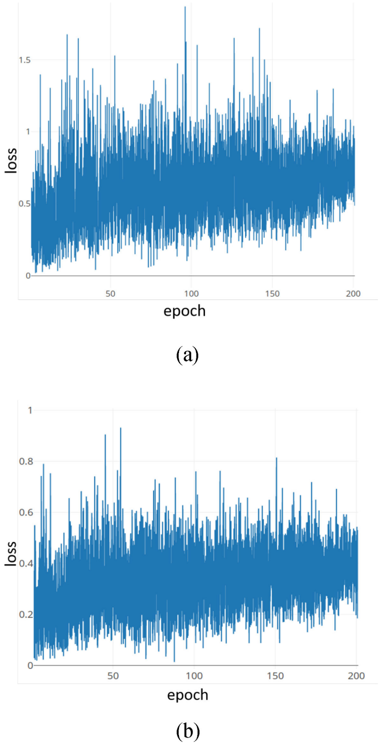
TABLE I
COMPARISON OF PERFORMANCE INDICES OF DIFFERENT MODELS

Model	LPIPS	PSNR (dB)	pHash
Original images	0.639	13.820	14.150
CycleGAN	0.549	14.333	13.305
SCycleGAN	0.453	14.354	13.100

matching algorithm for matching on different image pairs. n is the number of correct matching points in the matching process, and EPE reflects the accuracy of the algorithm's matching results on different image pairs. SR is expressed as the percentage of NCM in the number of total matching points (NTP) given by the algorithm, which can reflect the matching point pair success rate of the algorithm matching on different image pairs.

C. Experiment Preparation

Image style transfer network training is performed before the matching experiment. To construct a style transfer image dataset, the original image data are cropped by randomly sampling the central point. All image data are cropped into satellite–aerial image pairs with a size of 256×256. The satellite–aerial images correspond one to one, with a total of 2800 pairs of images. They are divided into two groups containing a training set of 2600 pairs and a test set of 200 pairs. To verify the effectiveness of the proposed model, CycleGAN and SCycleGAN were trained and tested using the constructed datasets. The training and testing results are shown in Figs. 6 and 7. At the same time, we calculated the similarity of Learning perceptual image patch similarity (LPIPS) [28], peak signal-to-noise ratio (PSNR) [29], and perceptual hash (pHash) [30] of satellite images, generated aerial images and original aerial images, recorded the average values of all image pairs in the test set, and quantitatively analyzed the image conversion effect. The results are shown in Table I, which are the corresponding numerical comparison results under the three models (original images, CycleGAN, SCycleGAN).



(a)

(b)

Fig. 6. Comparison of training curves before and after generating network loss function improvement. (a) CycleGAN. (b) SCycleGAN.

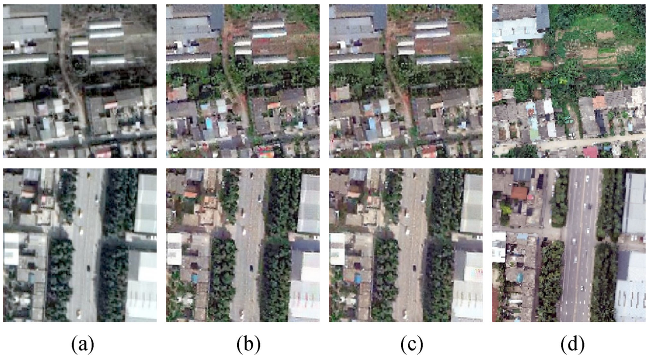

(a) (b) (c) (d)

Fig. 7. Different methods for mapping satellite image aerial image. (a) Input. (b) CycleGAN. (c) SCycleGAN (ours). (d) Ground-truth.

The experimental results show that the training time (CycleGAN: 2730 min, SCycleGAN: 1860 min) can be shortened by using the proposed training method, and the loss fluctuation range of the generative network of SCycleGAN is significantly better than that of CycleGAN. However, due to the complex characteristics of remote sensing images and the different shooting times and angles of satellite images and aerial images, there are landform changes and distortions, resulting in large fluctuations in the training curve. This reflects the game process between the generator and the discriminator. From Fig. 7, it can be seen that the style transfer network model does not change the scale, perspective, and target morphology of the original image but only makes modal changes. Table I shows that CycleGAN and SCycleGAN can improve the similarity of the original

TABLE II
COMPARISON RESULTS OF THREE TEST SETS UNDER DIFFERENT MODELS

Model	Test set	NCM	NTP	SR (%)	EPE
D2-Net	SA1	71	203	34.50	10.75
D2-Net	SA2	90	203	44.20	3.12
D2-Net	SA3	98	203	48.28	2.84
LoFTR	SA1	44	138	31.88	12.14
LoFTR	SA2	50	159	31.33	4.39
LoFTR	SA3	52	160	32.50	3.68

satellite images and aerial images to a certain extent, and the generated images converted by SCycleGAN are more similar to the original aerial images in terms of structure and color. Overall, the aerial images generated by the proposed model are of high quality, similar to the original aerial image, and the structure, color, and details of the image are characterized completely. There are no large area distortions, artifacts, distortions, or other phenomena, and the conversion effect is better.

D. Matching Experimental Results and Analysis

To verify the feasibility of using the image conversion mechanism as a preprocessing method for heterologous matching, we selected two more advanced network models, D2-Net and LoFTR, to evaluate the matching performance of the training models when applied to a new dataset by evaluating the test dataset. The production method of the test dataset is similar to the method of constructing the style transfer image dataset in Section III-C. It needs to crop out satellite–aerial image pairs with a size of 960×540. Since the matching transformation model between satellite and aerial images can be expressed by projection transformation, the matching test dataset is generated by applying random projection transformation based on four-point disturbance to the obtained image pairs [31], which contain 1000 pairs of satellite–aerial image pairs and corresponding homography ground-truth. The image size is 256×256, and the dataset is called SA1. Additionally, we use the style transfer models trained by CycleGAN and SCycleGAN to perform style transfer on the satellite images in SA1 and form new test datasets with the generated aerial images, the original aerial images, and the corresponding homography ground-truth, which are called SA2 and SA3, respectively. The experiment counts the average values of NCM, SR, and EPE for each dataset test under different network models. The accuracy threshold is 5 pixels, and the results are shown in Table II. The visualization results are shown in Figs. 8 and 9.

It can be seen from the figure that the image style transfer preprocessing method extracts more uniform features than the direct matching algorithm, and the number of matching points between the converted images increases significantly. Combined with Table II, it can be seen that the NCM after the image style transfer preprocessing method is more than that after the direct matching algorithm. This shows that this method can improve the robustness of the matching algorithm. From the results of SR and EPE, the image style transfer preprocessing method can improve the matching accuracy and the image matching point pair success rate. In addition, accuracies after preprocessing using SCycleGAN are higher than those of CycleGAN, which further proves that the image style transfer method proposed in this article can improve the matching effect between satellite images and aerial images.

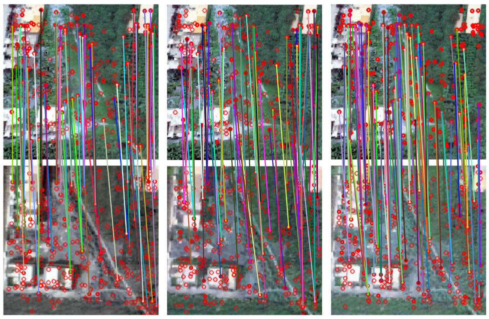


Fig. 8. Matching results of D2-Net under different test sets. From left to right, SA1 (the aerial image and satellite image), SA2 (the aerial image and CycleGAN-generated aerial image), and the matching effect of SA3 (aerial image and SCycleGAN-generated aerial image) are visually displayed. The lines of different colors are the corresponding relationship of matching point pairs, and the red points are the feature points detected by the algorithm.

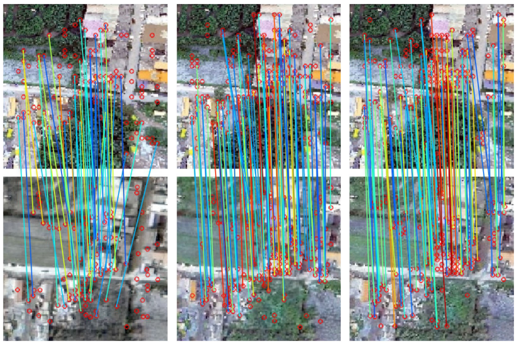


Fig. 9. Matching results of LoFTR under different test sets. From left to right, SA1 (the aerial image and satellite image), SA2 (the aerial image and CycleGAN-generated aerial image), and the matching effect of SA3 (aerial image and SCycleGAN-generated aerial image) are visually displayed.

In addition, we use three classical algorithms, SIFT, SURF, and oriented fast and rotated brief (ORB) [32], for matching experiments, respectively. The number of correct matching points is very small, and the image matching fails. The visualization result examples are shown in Figs. 10–12.

The experimental results show that the style transfer method combined with the classical matching method cannot achieve the matching of satellite images and aerial images. The reason is that style transfer only reduces the differences between images, the transferred images themselves are still heterogeneous images, and there are still differences in resolution and geometric distortion between images. This phenomenon proves the necessity of using deep learning methods to achieve matching in the method of this article.

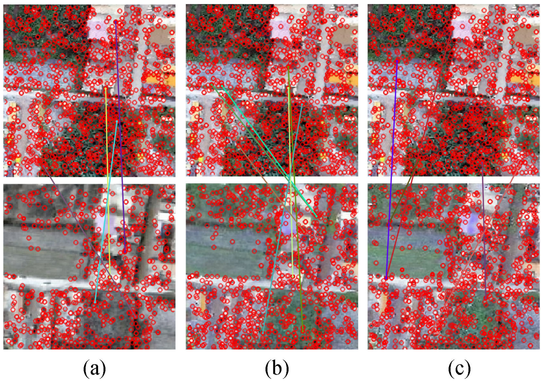


Fig. 10. Matching results of SIFT under different test sets. From left to right, SA1 (the aerial image and satellite image), SA2 (the aerial image and CycleGAN-generated aerial image), and the matching effect of SA3 (aerial image and SCycleGAN-generated aerial image) are visually displayed.

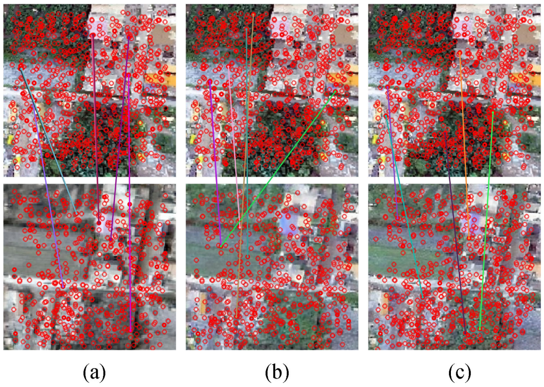


Fig. 11. Matching results of SURF under different test sets. From left to right, SA1 (the aerial image and satellite image), SA2 (the aerial image and CycleGAN-generated aerial image), and the matching effect of SA3 (aerial image and SCycleGAN-generated aerial image) are visually displayed.

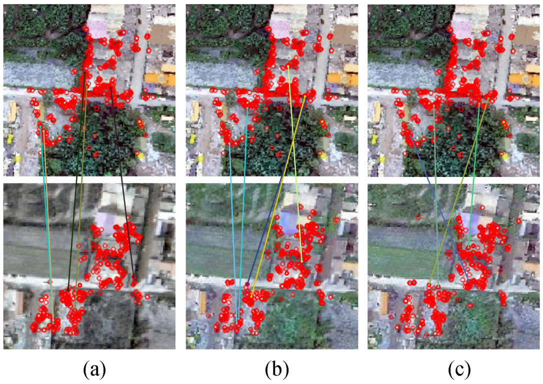


Fig. 12. Matching results of ORB under different test sets. From left to right, SA1 (the aerial image and satellite image), SA2 (the aerial image and CycleGAN-generated aerial image), and the matching effect of SA3 (aerial image and SCycleGAN-generated aerial image) are visually displayed.

IV. Conclusion

In this article, a satellite-aerial remote sensing image matching method based on style transfer is proposed to solve the problems of large differences between satellite images and aerial images, such as different imaging principles and resolutions. The experimental results show that the preprocessing of heterologous images through the style transfer method to reduce the difference between images, combined with the feature-matching method based on deep learning, can effectively improve the accuracy and robustness of the matching algorithm. This establishes the feasibility of using the image style transfer idea for heterogenous image matching. In addition, this article improves the loss function of the original CycleGAN generation network, effectively reduces the complexity of the algorithm, improves the quality of image generation, and provides an effective reference for solving the matching problem of heterogeneous images and the processing of heterogeneous image data.

For future works, we plan to apply this work to other types of heterogenous remote sensing image matching, such as optical and infrared images, optical, and SAR images, and further to serve as a building block for change detection or fusion of remote sensing images.

Acknowledgment

The authors would like to thank the reviewers for their valuable comments and suggestions.

References

[1] R. Feng, H. Shen, J. Bai, and X. Li, “Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions,” *IEEE Trans. Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 120–142, Dec. 2021.

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

[3] H. Bay, T. Tuytelaars, and L. V. Gool, “SURF: Speeded up robust features,” in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[4] J. Li, Q. Hu, and M. Ai, “RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform,” *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 3296–3310, Dec. 2019.

[5] J. Li, Q. Hu, and M. Ai, “Accurate point matching based on multi-objective genetic algorithm for multi-sensor satellite imagery,” *Appl. Math. Comput.*, vol. 236, pp. 546–564, Jun. 2014.

[6] S. Suri and P. Reinartz, “Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 939–949, Feb. 2010.

[7] M. Hasan, M. R. Pickering, and X. Jia, “Robust automatic registration of multimodal satellite images using CCRE with partial volume interpolation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 4050–4061, Oct. 2012.

[8] I. Rocco, R. Arandjelovic, and J. Sivic, “Convolutional neural network architecture for geometric matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2553–2567, Nov. 2019.

[9] C. F. Olson, “Maximum-likelihood template matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 52–57.

[10] J. Inglada and A. Giros, “On the possibility of automatic multisensor image registration,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2104–2120, Oct. 2004.

[11] H. M. Chen, M. K. Arora, and P. K. Varshney, “Mutual information based image registration for remote sensing data,” *Int. J. Remote Sens.*, vol. 24, no. 18, pp. 3701–3706, Dec. 2019.

[12] H. R. Boveiri, R. Khayami, R. Javidan, and A. Mehdizadeh, “Medical image registration using deep neural networks: A comprehensive review,” *Comput. Elect. Eng.*, vol. 87, pp. 1–24, Jan. 2020.

[13] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, “A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.

[14] M. Zhang, Z. Wang, R. Bai, and H. Jia, “A coarse-to-fine optical and SAR remote sensing image registration algorithm,” *J. Geo-Inf. Sci.*, vol. 22, no. 11, pp. 2238–2246, Nov. 2020.

[15] 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.

[16] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[17] S. Chopra, R. Hadsell, and Y. Lecun, “Learning a similarity metric discriminatively, with application to face verification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546.

[18] I. Goodfellow et al., “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Nov. 2020.

[19] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2016, *arXiv:1511.06434v2*.

[20] A. Martin, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *Proc. Int. Conf. Mach. Learn*., 2017, pp. 214–223.

[21] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[22] M. Yang and J. He, “Image style transfer based on DPN-CycleGAN,” in *Proc. IEEE 4th Int. Conf. Pattern Recognit. Artif. Intell.*, 2021, pp. 141–145.

[23] X. Li, Z. Du, Y. Huang, and Z. Tan, “A deep translation (GAN) based change detection network for optical and SAR remote sensing images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 14–34, Jul. 2021.

[24] J. Song, J. Li, H. Chen, and J. Wu, “MapGen-GAN: A fast translator for remote sensing image to map via unsupervised adversarial learning,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2341–2357, Jan. 2021.

[25] M. Dusmanu et al., “D2-Net: A trainable CNN for joint description and detection of local features,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8084–8093.

[26] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, “LoFTR: Detector-free local feature matching with transformers,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8918–8927.

[27] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

[28] R. Zhang, P. Isola, A. Alexei, E. Schechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[29] Y. Wang, J. Li, Y. Lu, Y. Fu, and Q. Jiang, “Image quality evaluation based on image weighted separating block peak signal to noise ratio,” in *Proc. IEEE Int. Conf. Neural Netw. Signal Process*., 2003, pp. 994–997.

[30] G. Ding and C. Zhu, “Perceptual hash algorithm for integrity authentication of remote sensing image,” *J. Southeast Univ.*, vol. 44, no. 4, pp. 723–727, Jul. 2014.

[31] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Deep image homography estimation,” 2016, *arXiv:1606.03798v1*.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

Jiawei Zhao received the B.S. degree in automatic control, in 2020, from the Xi’an Research Institute of Hi-tech, Xi’an, China, where he is currently working toward the M.S. degree.

His research interests include remote sensing image processing, remote sensing application, and artificial intelligence learning.

Dongfang Yang received the B.S., M.S., and Ph.D. degrees from the Xi’an Research Institute of Hi-tech, Xi’an, China, in 2006, 2009, and 2013, respectively, all in automatic control.

He is an Associate Professor with the Xi’an Research Institute of Hi-tech. His research interests include autonomous aerial vehicles, modern navigation technology, and nonlinear filtering.

Yongfei Li received the B.S., M.S., and Ph.D. degrees from the Xi’an Research Institute of Hi-tech, Xi’an, China, in 2014, 2016, and 2021, respectively, all in automatic control.

His research interests include SLAM, vision-based localization, and navigation.

Peng Xiao received the B.S. degree in automatic control, from the Xi’an Research Institute of Hi-tech, Xi’an, China, in 2020. Where he is currently working toward the M.S. degree.

His research interests include remote sensing image processing, remote sensing application, and artificial intelligence learning.

Jinglan Yang received the B.S. degree in automatic control from the Hebei University of Science and Technology, Hebei, China, in 2020. She is currently working toward the M.S. degree in automatic control from the Xi’an Research Institute of Hi-tech, Xi’an, China.

Her research interests include SLAM and vision-based localization.