Feature Matching and Position Matching Between Optical and SAR With Local Deep Feature Descriptor

Yun Liao, Yide Di, Hao Zhou, Anran Li, Junhui Liu, Mingyu Lu, and Qing Duan

***Abstract*—Image matching between the optical and synthetic aperture radar (SAR) is one of the most fundamental problems for earth observation. In recent years, many researchers have used hand-made descriptors with their expertise to find matches between optical and SAR images. However, due to the large nonlinear radiation difference between optical images and SAR images, the image matching becomes very difficult. To deal with the problems, the article proposes an efficient feature matching and position matching algorithm (MatchosNet) based on local deep feature descriptor. First, A new dataset is presented by collecting a large number of corresponding SAR images and optical images. Then a deep convolutional network with dense blocks and cross stage partial networks is designed to generate deep feature descriptors. Next, the hard L2 loss function and ARCpatch loss function are designed to improve matching effect. In addition, on the basis of feature matching, the two-dimensional (2-D) Gaussian function voting algorithm is designed to further match the position of optical images and SAR images of different sizes. Finally, a large number of quantitative experiments show that MatchosNet has a excellent matching effect in feature matching and position matching. The code will be released at: https://github.com/LiaoYun0x0/Feature-Matching-and-Position-Matching-between-Optical-and-SAR.**

***Index Terms*—Deep learning, feature descriptor, feature matching, image matching, optical images, position matching, synthetic aperture radar (SAR) images.**

I. Introduction

IN EARTH observations, optical and synthetic aperture radar (SAR) images can be compared and analyzed to obtain more valuable information by complementation. In recent years, image segmentation [1], image classification [2], multimodal manifold learning [3], and feature matching [4], [5] have been widely used in jointly processing and analyzing SAR and optical data. And in the fields of image registration [6], image fusion [7], and change detection [8], feature matching between the SAR images and optical images becomes very significant. However, as shown in Fig. 1, due to the imaging mechanism of the optical and SAR images are very different, it is difficult to match the feature between the optical images and SAR images. Speckle noise [9] that affects the performance of the features widely exists in SAR images and makes them difficult to be recognized. Also, the distance-dependence along the range axis and the characteristics of radar signal wavelengths [10] result in the geometric distortion in the SAR images.

Fig. 1. Example comparing the different size of optical and SAR images.

Image matching methods can be divided into three categories: area-based descriptors matching methods, handcrafted feature descriptors matching methods, and learning-based feature descriptors matching methods. Area-based methods [11]–[14] can directly match the images at the pixel level through the appropriate patch similarity measurement. However, appearance changing, lighting changing, and image distortion can mislead similarity measurement and match searching. Therefore, these methods are usually only applicable in the following cases: scaling, local deformation, and small rotation.

Experts and scholars deduce and design hand-crafted feature descriptors by existing knowledge widely used in visual applications. For nonlinear brightness changes, due to the diversity of gradient statistics around feature points, SIFT feature points are not reliable in the calculation of the main direction, which will produce fewer correct matching points and more wrong matching points, resulting in false registration or registration failure. In recent decades, many handcrafted feature descriptors matching methods [15]–[20] have emerged, but due to nonlinear radiometric difference, it is very difficult to extract sufficient number of highly repetitive features from optical and SAR images [21], [22].

Compared to the handcrafted descriptors, the learning-based feature descriptors can discover more valuable information hidden in the data. Learning-based feature descriptors also have better performance and feature description capability.

Manuscript received September 11, 2021; revised October 8, 2021, November 24, 2021, and December 5, 2021; accepted December 6, 2021. Date of publication December 13, 2021; date of current version December 31, 2021. This work was supported in part by the Open Foundation of Key Laboratory in Software Engineering of Yunnan Province under Grant 2020SE307, in part by the Scientific Research Fund of Yunnan Provincial Education Department under Grant 2021J0007, and in part by the National Natural Science Foundation of China under Grant 61976124. *(Corresponding author: Qing Duan.)*

Yun Liao, Junhui Liu, and Qing Duan are with the National Pilot School of Software, Yunnan University, Kunming, Yunnan 650106, China (e-mail: 676295641@qq.com; hanks@ynu.edu.cn; qduan@ynu.edu.cn).

Yide Di and Mingyu Lu are with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China (e-mail: ghostdyd@126.com; lumingyu@dlmu.edu.cn).

Hao Zhou and Anran Li are with the Yunnan Lanyi Network Technology Co, Kunming 650000, China (e-mail: 1083480050@qq.com; 1165071324@qq.com).

Digital Object Identifier 10.1109/JSTARS.2021.3134676

In many kinds of images, feature descriptors based on deep learning [23]–[30] achieve better results in image patch comparison than traditional descriptors. However, the learning-based feature descriptors face many difficulties too. For example, the deep learning method usually extracts a large number of features from the images, which often contain noise and outliers.

The above methods are all devoted to solve the problem of feature matching for images of the same size, but in many cases in real life, there are not enough data that fully meet the requirements. In this article, we work on solving the feature matching problems for optical and SAR images of different sizes, and based on this, the position matching is further implemented. In order to better solve the problems between optical images and SAR images, a novel and automatic method—MatchosNet is proposed. A large number of experiments are conducted to demonstrate that MatchosNet has very excellent effects in processing the feature matching and position matching between the optical and SAR images. Code will be publicly available.

In summary, our main contributions are as follows.

1) A complex deep convolutional neural network consisting of multiple dense convolutional blocks and cross stage partial networks is designed to generate deep feature descriptors. The network achieves feature reuse in the channel dimension and obtains better performance with fewer parameters. In addition, a new dataset is proposed by collecting a large number of corresponding SAR images and optical images.
2) In the model training, the ARCpatch loss function and hard L2 loss function are designed. The ARCpatch loss function uses classification strategy to maximize the distance between positive and negative samples. The hard L2 loss function uses the strategy of actively mining the "hardest samples." Finally, the two different loss functions are assigned with the most suitable weights to form an effective composite loss function.
3) The trained feature descriptors are used to achieve feature matching between optical images and SAR images. In addition, a 2-D Gaussian function voting algorithm is designed to achieve position matching of SAR images and optical images with different sizes.

The rest of this article is organized as follows: Section II introduces the related work of Image matching. Section III describes our method in detail. The qualitative and quantitative experiments are described in Section IV. The conclusions and the future directions are drawn in Section V.

II. Related Work

With the development of computer vision technology and deep learning technology, more and more methods have been proposed to jointly process and analyze SAR and optical data. In 2019, Tochon *et al.* [1] presented a novel methodology for the hierarchical representation and segmentation of multimodal images. Hong *et al.* [2] proposed a general MDL framework to deal with the pixel-level RS image classification tasks in 2021. The framework consists of two subnetworks: Ex-Net and Fu-Net and they were used to extract features and fuse features, respectively. In 2021, Pournemat *et al.* [3] proposed a model to simultaneously learn the underlying low-dimensional manifold in each modality, and locally align these manifolds across different modalities. In general, feature matching is the most widely used method in optical image and SAR image processing and analysis. The methods can be categorized as area-based descriptors matching methods, handcrafted feature descriptors matching methods, and learning-based feature descriptors matching methods.

Area-Based Descriptors Matching Methods: In recent years, some area-based descriptors have been proposed for image matching between optical and SAR images. Area-based descriptors are mainly divided into two types: phase congruency (PC)-based descriptors [31] and local self-similarity (LSS)-based descriptors [32]. For descriptors based on phase consistency, Ye *et al.* [22] proposed the histogram of orientated phase congruency (HOPC). They extended the phase congruency model to generate the direction representation, and then designed the HOPCncc which could well solve the complex nonlinear radiation difference. Fan *et al.* [33] proposed the phase congruency structural descriptor (PCSD). They designed a Harris (Und-Harris) feature extraction method based on uniform nonlinear diffusion to reduce the adverse effect of speckle noise on feature extraction and proposed a PCSD that was constructed in a grouping manner on PC structure images. For descriptors based on LSS, Ye *et al.* [21] developed the dense local self-similarity (DLSS) in 2017. They proposed DLSS and defined a similarity measure DLCS by integrating multiple small LSS descriptors, and then used template matching strategy for image matching detection. Xiong *et al.* [12] presented RLSS for optical-to-SAR image template matching in 2020. The RLSS described the local shape properties of the images in a discriminable manner, and it could be integrated into a dense sampling grid to obtain the DRLSS descriptor, thereby further improving the discriminability. Gao *et al.* [34] introduced a novel discrete cosine transform-based feature (DCTF) descriptor in 2021. It preserved local structure more compactly in the frequency domain by utilizing the mathematical properties of the discrete cosine transform (DCT).

Handcrafted Feature Descriptors Matching Methods: Dong *et al.* [17] proposed DSP-SIFT descriptor by integrating pooling gradient directions of different domain sizes on the basis of SIFT. Aguilera *et al.* [35] proposed the edge-oriented histogram (EOH) that represents the image features by the edge point distribution between far-infrared and visible images. Nunes *et al.* [36] developed the multispectral feature descriptor (MFD). They used the log-Gabor filter to get the image data at different frequencies of the electromagnetic spectrum. Fu *et al.* [37] proposed the directional response maps (DMs) and the directional response binary maps (DBMs). They used DMS and DBMS to capture the common structure and texture attributes of multispectral images, and combined the corresponding normalized feature vectors to obtain the histogram of the directional image. Qian *et al.* [38] presented the extraction of phase consistency feature points based on low-contrast nonsuppressed SAR-Harris multiscale space in 2020. Jiang *et al.* [39] proposed a simple yet efficient method termed LAF for both rigid and nonrigid feature matching of remote sensing images and apply it to the image registration task in 2021. Handcrafted feature descriptors have made great contributions to the image matching. However,

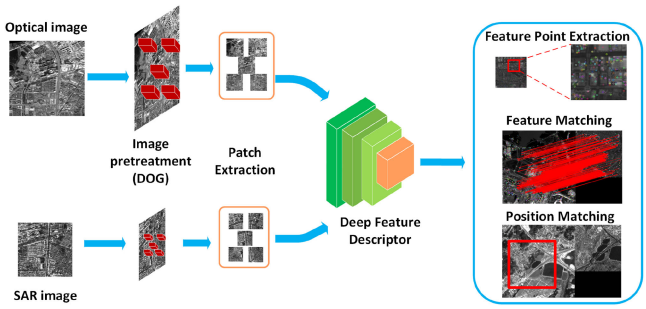


Fig. 2. Procedure of the MatchosNet.

due to the nonlinear radiometric difference, handcrafted feature descriptors cannot optimally account for all variations in the appearance of the image and extract sufficient number of highly repetitive features from optical and SAR images.

Learning-Based Feature Descriptors Matching Methods: Han *et al.* [23] proposed a patch matching system called MatchNet in 2015. The twin network structure and measurement network adopted by MatchNet are typical practices in matching algorithms, which lay a foundation for the following research work. Balntas *et al.* [40] implemented TFeat in 2016. TFeat utilizes three sets of training samples, which can obtain better descriptors and faster learning speed. Mishchuk *et al.* [41] proposed the HardNet in 2017. The loss function for learning a local image descriptor maximizes the distance between the closest positive and closest negative examples in a batch. Du *et al.* [42] proposed FM-CycleGAN for enforcing the feature matching consistency of unsupervised-image-synthesis by introducing feature matching loss to CycleGAN in 2021. Ma *et al.* [4] introduced a new method based on image transfer and local feature for multispectral image matching in 2021. A new regularized conditional generative adversarial network (GAN) was proposed for image transfer to preprocess the multispectral images.

Although great progress has been made in the field of feature matching, there is still very large room for improvement. Therefore, MatchosNet is proposed to further improve the effect of feature matching and expand the ability of position matching. Unlike the above research, MatchosNet not only generates deep convolution descriptors for optical images and SAR images, but also applies them to the position matching. The method designs a special network with dense blocks and cross stage partial networks to generate deep feature descriptors and a compound loss function composed of the ARCpatch loss function and hard L2 loss function to get a better match. The experiment uses the SEN1-2 dataset [43], SARptical [44] dataset, and an image dataset composed of the corresponding optical images and SAR images, which are collected by us to objectively show the superiority of feature matching and position matching of MatchosNet.

III. Methodology

A. The Procedure of the Proposed Method

As shown in Fig. 2, the method proposed in this project mainly consists of five sections. These are image pretreatment, image patch extraction, deep feature descriptor generation, feature matching, and position matching.

First, the image is preprocessed to generate the center point of image patch. Since the optical image is larger than the SAR image and contains more content, the optical image and SAR image cannot be directly input into the model for feature detection. The differential Gaussian operator (DoG) [45] is used to extract feature points from optical and SAR images, and then image patches of the same size are extracted with these feature points as the center points. The DoG function is described as

$$\begin{aligned} \mathrm{D\,o\,G} &\triangleq G_{\sigma_1} - G_{\sigma_2} \\ &= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma_1}e^{-\left(x^2+y^2\right)/2\sigma_1^2} - \frac{1}{\sigma_2}e^{-\left(x^2+y^2\right)/2\sigma_2^2}\right) \end{aligned} \tag{1}$$

where G_{σ_1} and G_{σ_2} represent the Gaussian filtering of two images, respectively.

During image pretreatment, we detect the DOG value for all pixel points in the image. If the DOG value of a pixel is the maximum or minimum value of all adjacent pixel points, it can be considered as a feature point.

During image patch extraction, we extract the surrounding area based on the detected feature points of optical images and SAR images, and reconstruct them into 64×64 pixel patches. These reconstructed patches can be used as training data of deep convolutional neural network to solve the problem of size difference between optical images and SAR images.

Next, a effective deep convolutional neural network is designed by referring to DenseNet and CSPNet. The deep convolutional neural network can achieve better results and generate fewer parameters by improving the utilization of feature. The architecture will be described in Section III-B. A compound loss function composed of the hard L2 loss function and ARCpatch loss function is designed to get a better match. The loss function will be described in Section III-C. The network model is trained to generate descriptors for the feature points of the optical and SAR images.

When performing feature matching, the feature points are matched by the feature descriptors. The horizontal offset, vertical offset, total offset, and the correct number of the matching points are further calculated by different methods.

Finally, the position matching of SAR images and optical images is implemented. The 2-D Gaussian function voting algorithm is designed to further match the position of optical images and SAR images of different sizes. The voting algorithm will be described in detail in Section III-D. The horizontal offset, vertical offset, total offset, and the correct number of the position matching are calculated to analysis results.

B. Architecture of the Proposed Framework

The traditional convolutional neural network [46] simply connects the upper and lower layers, and the output of the ith layer is the input of the $(i+1)$th layer. Its function can be defined as

$$X_i = H_i\left(X_{i-1}\right) \tag{2}$$

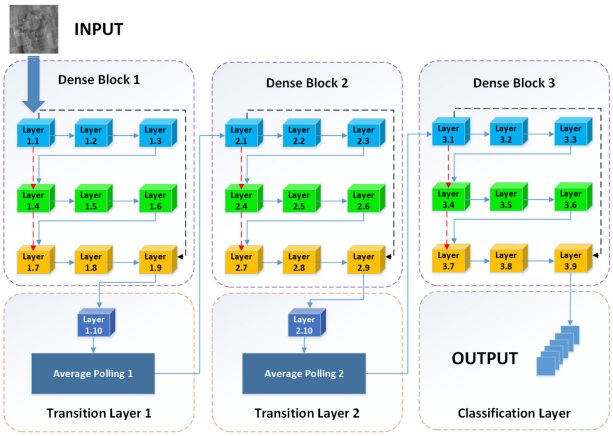


Fig. 3. Architecture of the deep neutral network.

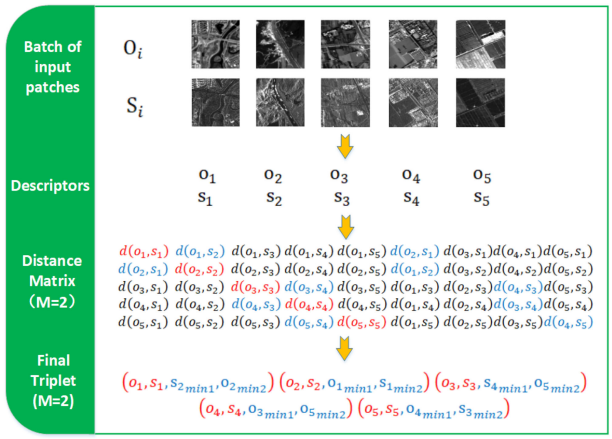


Fig. 4. Sampling procedure of the Hard L2 method.

where X_i represents the output of the ith layer, $H_i()$ denotes a composite function of operations such as Batch Normalization, ReLU, pooling, and convolution.

He *et al.* [47] proposed ResNet and add a skip-connection to bypass the nonlinear transformations. The function of the ResNet can be presented as

$$X_i = H_i\left(X_{i-1}\right) + X_{i-1}. \tag{3}$$

Huang *et al.* [48] proposed DenseNet and introduced direct connections from any layer to all subsequent layers. The function of the DenseNet can be presented as

$$X_i = H_i\left(\left[X_0, X_1, \ldots, X_{i-1}\right]\right). \tag{4}$$

Compared with the traditional convolutional network and ResNet, DenseNet can achieve better results because it makes more effective use of features, enhances the transmission of features, reduces the gradient disappearance, and reduces the number of parameters.

Wang *et al.* [49] proposed CSPNet and introduced transition layer to remove computational bottlenecks and strengthen learning ability of the convolutional network. The function of the transition layer of the CSPDenseNet can be presented as

$$\begin{aligned} \mathbf{x_k} &= \mathbf{w_k}^{*}\left[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x_{k-1}}\right] \\ \mathbf{x_T} &= \mathbf{w}_{\mathbf{T}}^{*}\left[\mathbf{x}_0, , \mathbf{x}_1, \ldots, \mathbf{x_k}\right] \\ \mathbf{x_U} &= \mathbf{w}_{\mathbf{U}}^{*}\left[\mathbf{x}_0, \mathbf{x}_{\mathrm{T}}\right]. \end{aligned} \tag{5}$$

CSPDenseNet retains the advantages of the feature reuse of DenseNet, while preventing excessive repetitive gradient information by truncating the gradient flow.

Because there are so many features and parameters in optical image and SAR image feature matching, it is very necessary to design a model with strong learning ability. As shown in Fig. 3, to make more effective use of features, enhance the transmission of features and prevent excessive repetitive gradient information, a dense convolutional network is designed. The deep convolutional neural network consists of three dense blocks and two transition layers. As shown in Table I, MatchosNet accepts the data of size $64 \times 64 \times 1$ and output the final result of size $256 \times 256 \times 1$. Each dense block contains nine layers, including six convolution layers, and three connection layers. Each transition layer contains a convolution layer and an average pooling layer, receiving data of $h \times w \times c$ and exporting data of $\frac{h}{2} \times \frac{w}{2} \times \frac{c}{4}$. The classification layer contains a special convolution layer whose convolution kernel size is 8×8, which can convert the data of size $8 \times 8 \times 21$ into the output data of size $256 \times 256 \times 1$. Compared with other methods, this network has better feature transfer effect and can generate nice deep convolution descriptors. Detailed experiments and comparison results are described in Section IV.

C. Model Training and Loss

Next, two loss functions are designed to achieve back propagation from matching to visual descriptors, thus optimizing the whole deep learning model.

1) The Hard l2 Loss Function: Recently, Tian *et al.* [24] and Mishchuk *et al.* [41] proposed two kinds of excellent loss functions and they both required the minimum matching distance between each row and each column to the ground truth. We learned the ideas of their methods and designed the Hard L2 method. As shown in Fig. 4, for each positive sample, 2n-1 negative samples are generated, and L2 distance [24] is used to select the first M negative samples with the smallest distance to the ground truth to optimize the model and obtain powerful feature descriptors.

According to the L2 distance formula, $d(o_i, s_j) = \sqrt{2 - 2o_i s_j}$, $i = 1 \cdots n, j = 1 \cdots n$ of size $n \times n$ is calculated, where o_i and s_j denote the optical descriptors and SAR descriptors, respectively. $s_{j_{\min}}$ and $o_{k_{\min}}$ are the first M closest nonmatching descriptors to o_i and s_i, respectively, where $j \neq i$ and $k \neq i$. Then, the triplets are formed: $(o_i, s_i, s_{j_{\min 1}}), (o_i, s_i, o_{k_{\min 2}}), \cdots (o_i, s_i, s_{j_{\min \mathrm{M}}})$ from the descriptors.

The goal is to minimize the distance between the matching descriptor and the first M closest nonmatching descriptors. These n distances are fed into the margin loss

$$\begin{aligned} L_{\text{hardl2}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \max\left(0, \left(1 + d\left(o_i, s_j\right)\right.\right. \\ \left.\left. - \min\left(d\left(o_i, S_{j_{\min}}\right), d\left(O_{k_{\min}}, s_i\right)\right)\right)\right). \end{aligned} \tag{6}$$

TABLE I
DESCRIPTION ON DEEP CONVOLUTIONAL NETWORKS STAGE

Block	Layer	Layer Description	Output Tensor(h*w*n)
INPUT	Input	Input image patch	64*64*1
Dense Block 1	Layer 1.1	Conv (3*3),stride(2),padding(1)	32*32*16
	Layer 1.2	Conv (1*1),stride(1)	32*32*48
	Layer 1.3	Conv (3*3),stride(1),padding(1)	32*32*12
	Layer 1.4	Connect Layer1.1 and Layer 1.3	32*32*28
	Layer 1.5	Conv (1*1),stride(1)	32*32*48
	Layer 1.6	Conv (3*3),stride(1),padding(1)	32*32*12
	Layer 1.7	Connect Layer1.4 and Layer 1.6	32*32*40
	Layer 1.8	Conv (1*1),stride(1)	32*32*20
	Layer 1.9	Connect Layer1.1 and Layer 1.8	32*32*36
Transition Layer 1	Layer 1.10	Conv (1*1),stride(1)	32*32*18
	Average Polling 1	kernel=2, stride=2, padding=0	16*16*9
Dense Block 2	Layer 2.1	Input from Transition Layer 1	16*16*9
	Layer 2.2	Conv (1*1),stride(1)	16*16*48
	Layer 2.3	Conv (3*3),stride(1),padding(1)	16*16*12
	Layer 2.4	Connect Layer2.1 and Layer 2.3	16*16*21
	Layer 2.5	Conv (1*1),stride(1)	16*16*48
	Layer 2.6	Conv (3*3),stride(1),padding(1)	16*16*12
	Layer 2.7	Connect Layer2.4 and Layer 2.6	16*16*33
	Layer 2.8	Conv (1*1),stride(1)	16*16*16
	Layer 2.9	Connect Layer2.1 and Layer 2.8	16*16*25
Transition Layer 2	Layer 2.10	Conv (1*1),stride(1)	16*16*12
	Average Polling 2	kernel=2, stride=2, padding=0	8*8*6
Dense Block 3	Layer 3.1	Input from Transition Layer 2	8*8*6
	Layer 3.2	Conv (1*1),stride(1)	8*8*48
	Layer 3.3	Conv (3*3),stride(1),padding(1)	8*8*12
	Layer 3.4	Connect Layer3.1 and Layer 3.3	8*8*18
	Layer 3.5	Conv (1*1),stride(1)	8*8*48
	Layer 3.6	Conv (3*3),stride(1),padding(1)	8*8*12
	Layer 3.7	Connect Layer3.4 and Layer 3.6	8*8*30
	Layer 3.8	Conv (1*1),stride(1)	8*8*15
	Layer 3.9	Connect Layer3.1 and Layer 3.8	8*8*21
Classification Layer	Output	Conv (8*8),stride(1)	256*256*1

2) The ArcPatch Loss Function: A large part of previous classification problems used Soft max as the loss layer of the network. Experiments show that Soft max considers whether the samples can be correctly classified, while there is a large optimization space in the problem of expanding the interclass distance between dissimilar samples and reducing the interclass distance between similar samples.

Deng *et al.* [50] proposed the Angular Margin Loss in the ArcFace. ArcFace is more "compact" in convergence compared to other losses, which compresses the same class into a tighter space. It is more dense than other losses, making the features learned by the network have a more pronounced angular distribution.

ArcFace is a loss function used for face recognition and it maximizes the classification boundary in the Angle space and has a very good effect in dealing with the classification problem. However, ArcFace loss function is not applicable in this project due to the great difference between face matching problem and key point feature matching problem. ArcFace maximizes the classification boundary, but the feature matching problem that we are dealing with does not have any classification information.

Therefore, as shown in Fig. 5, a new loss function called ARCpatch is designed. Different from ArcFace method, ARCpatch does not have a center vector matrix and cannot form an accurate number of categories. A special classification method is designed according to the sample matching situation of feature

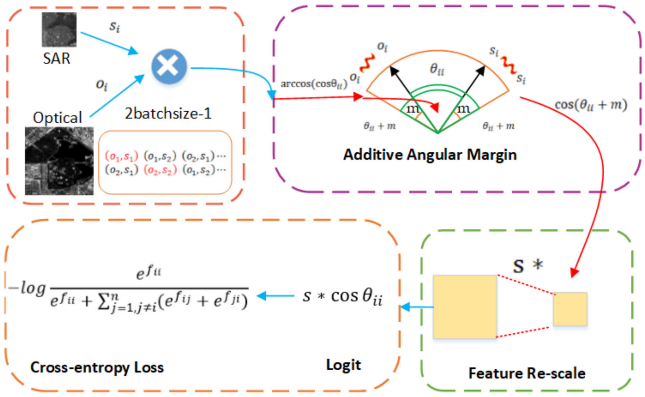


Fig. 5. ARCpatch loss function schematic.

matching problem. For each batch of samples, 2batchsize-1 categories are generated to calculate the loss, which include a positive sample matching category and 2batchsize-2 negative sample matching categories. The loss for the *i*th sample can be presented as

$$-\log \frac{e^{f_{ii}}}{e^{f_{ii}} + \sum_{j=1, j\neq i}^{n} \left(e^{f_{ij}} + e^{f_{ji}}\right)}. \tag{7}$$

ARCpatch algorithm maximizes the distance between the positive samples and the negative samples of feature points in

Fig. 6. Example of position matching.

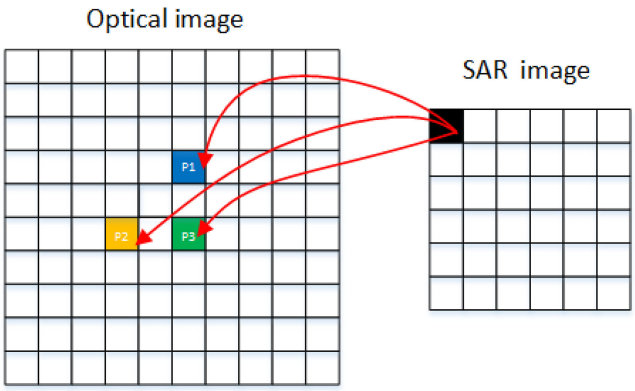
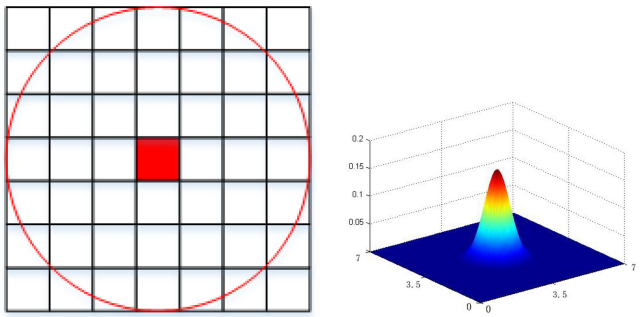


Fig. 7. Candidate position coordinates.



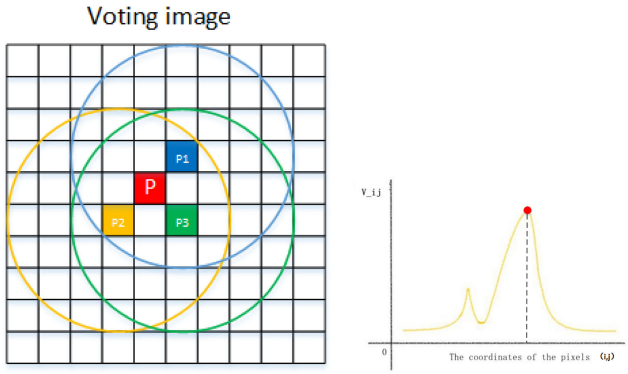
(a) (b)

Fig. 8. Gaussian weighted template and the distribution. (a) Gaussian weighted template (b) The distribution of Gaussian weights.

angular space and its overall formula can be presented as

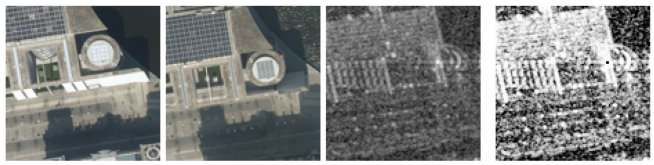
$$L_{\text{ARCpatch}} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(\cos(\theta_{ii}+m))}}{e^{s(\cos(\theta_{ii}+m))}+\sum_{j=1,j\neq i}^{n}\left(e^{s(\cos\theta_{ij})}+e^{s(\cos\theta_{ji})}\right)} \tag{8}$$

where n represents the value of batch size, $\cos\theta_{ii}$ represents the distance between the positive samples, $\cos\theta_{ij}$ and $\cos\theta_{ji}$ represent the distance between the negative samples. we add an additive angular margin penalty m between o_i and s_i to simultaneously enhance the compactness between the positive samples and the negative samples.



(a) (b)

Fig. 9. Voting strategy and the function. (a) Voting strategy image (b) The graph of the function.



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

Fig. 10. The normalization operation on the SAR image.

Algorithm 1:The Learning Algorithm for MatchosNet.

for sampled minibatch $\{img_i^{opt}, img_i^{sar}\}_{i=1}^{N}$ **do**
  for i $\in \{1, \ldots \mathrm{N}\}$ **do**
    $opt_i = f(img_i^{opt})$
    $\mathrm{sar}_i = f(img_i^{sar})$
  end for
  define $\mathrm{d}(\mathrm{opt_i}, \mathrm{sar_j}) = \sqrt{2 - 2\mathrm{opt_i sar_j^T}}$
  for i $\in \{1, \ldots, \mathrm{n}\}$ and j $\in \{1, \ldots, \mathrm{n}\}$ **do**
    $\cos(\theta_{\mathrm{i,j}}) = \mathrm{opt_i\, sar_j^T} / (\| \mathrm{opt}_i \| * \| \mathrm{sar}_j \|)$
    $s_{i,i} = s * \cos(\theta_{ii} + m)$
    $s_{i,j_{i\neq j}} = s * \cos(\theta_{ij})$
    $d_i^{\mathrm{pos}} = \mathrm{d}(\mathrm{opt_i}, \mathrm{sar_i})$
    $\{d_{i,k}^{\mathrm{neg}} \mid k = 1, \ldots, m\} =$
    $\mathrm{top\_}\, m_{\min}\{\{\mathrm{d}(\mathrm{opt_i}, \mathrm{sar_j}), \mathrm{d}(\mathrm{opt_j}, \mathrm{sar_i})\} \mid i \neq j\}$
  end for
  define $l_{\text{ArcPatch}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} - \log(\frac{\exp(s_{i,i})}{\exp(s_{i,i})+\sum_{\mathrm{j=1,i\neq j}}^{\mathrm{n}}(\exp(s_{i,j})+\exp(s_{j,i}))})$
  define
  $l_{\text{hardl}_2} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\sum_{\mathrm{j=1}}^{\mathrm{m}}\max(0, (1 + d_i^{pos} - \mathrm{d_{i,j}^{neg}}))$
  $\mathrm{L} = \lambda_1 * l_{\text{hardl}_2} + \lambda_2 * l_{\text{ArcPatch}}$
end for

3) The Compound Loss Function: Both of the two loss functions calculate the error between the training set and the label through angle, so they can be combined into a composite loss function to further improve the training effect. As introduced in Algorithm 1, the two loss functions are assigned by the most

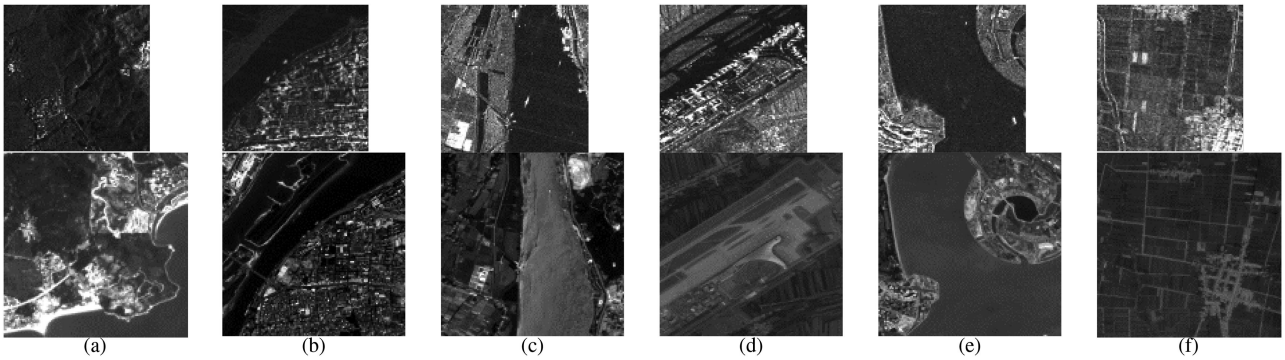


Fig. 11. The corresponding dataset of optical images and SAR images which were arranged by ourselves. (a) Port, (b) Urban area, (c) River system, (d) Airport, (e) Island (f) Plain. The images above are SAR images, and the images below are the corresponding optical images.

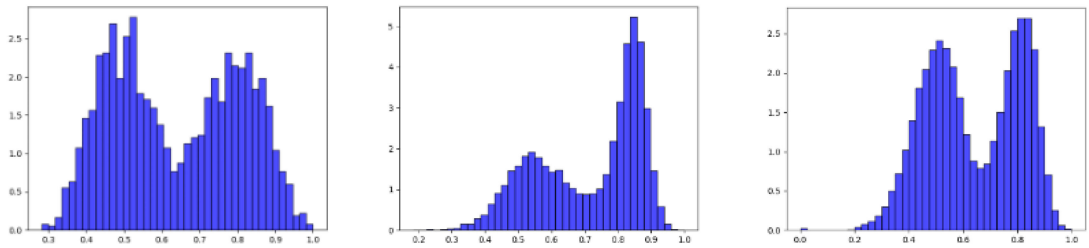


Fig. 12. Detection of positive and negative sample distribution of the three datasets. (a) SEN1-2 dataset (b) SARptical dataset (c) Our proposed dataset.

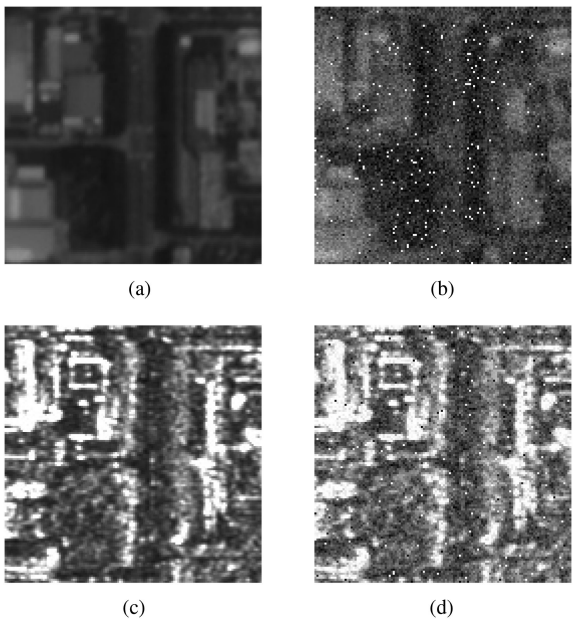


Fig. 13. Images are processed by Gaussian noise and salt-and-pepper noise. (a) Optical (b) Noise-Optical (c) SAR (d) Noise-SAR.

appropriate weights and combined to design the compound loss function for MatchosNet. The loss function can be presented as

$$\text{Loss} = \lambda_1 L_{\text{Hard L2}} + \lambda_2 L_{\text{ARCpatch}}. \tag{9}$$

Through a large number of experiments, it can be found that it is most effective to increase the distance difference first and then increase the Angle difference in the training process. So, the loss function sets $\lambda_1 = 1$, $\lambda_2 = \frac{i^2}{500}$, and i represents the number of epochs. The compound loss function has a more distinct margin than both ARCpatch-Loss and Hard L2-Loss.

D. Position Matching Algorithm

In real life, due to equipment and technology, SAR images are very limited. In contrast, optical images are much easier to obtain. Therefore, it is particularly important to locate the specific position of the smaller SAR image on the larger optical image. As shown in Fig. 6, on the basis of feature matching, a 2-D Gaussian function voting algorithm is designed to achieve position matching between SAR images and optical images.

The matching of each pair of feature points can obtain a coordinate of the pixel in the upper left corner of a SAR image on the optical image. As shown in Fig. 7, due to each set of images has many different feature matching points, it is possible to obtain multiple candidate position coordinates.

The voting algorithm of position matching is designed by the 2-D Gaussian distribution. Since the random vectors X and Y in this experiment are not correlated, the function sets $\rho = 0$. The 2-D Gaussian function can be shown as

$$f(\mathrm{x},\mathrm{y}) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(x-\mu_2)^2}{{\sigma_2}^2}\right)\right]. \tag{10}$$

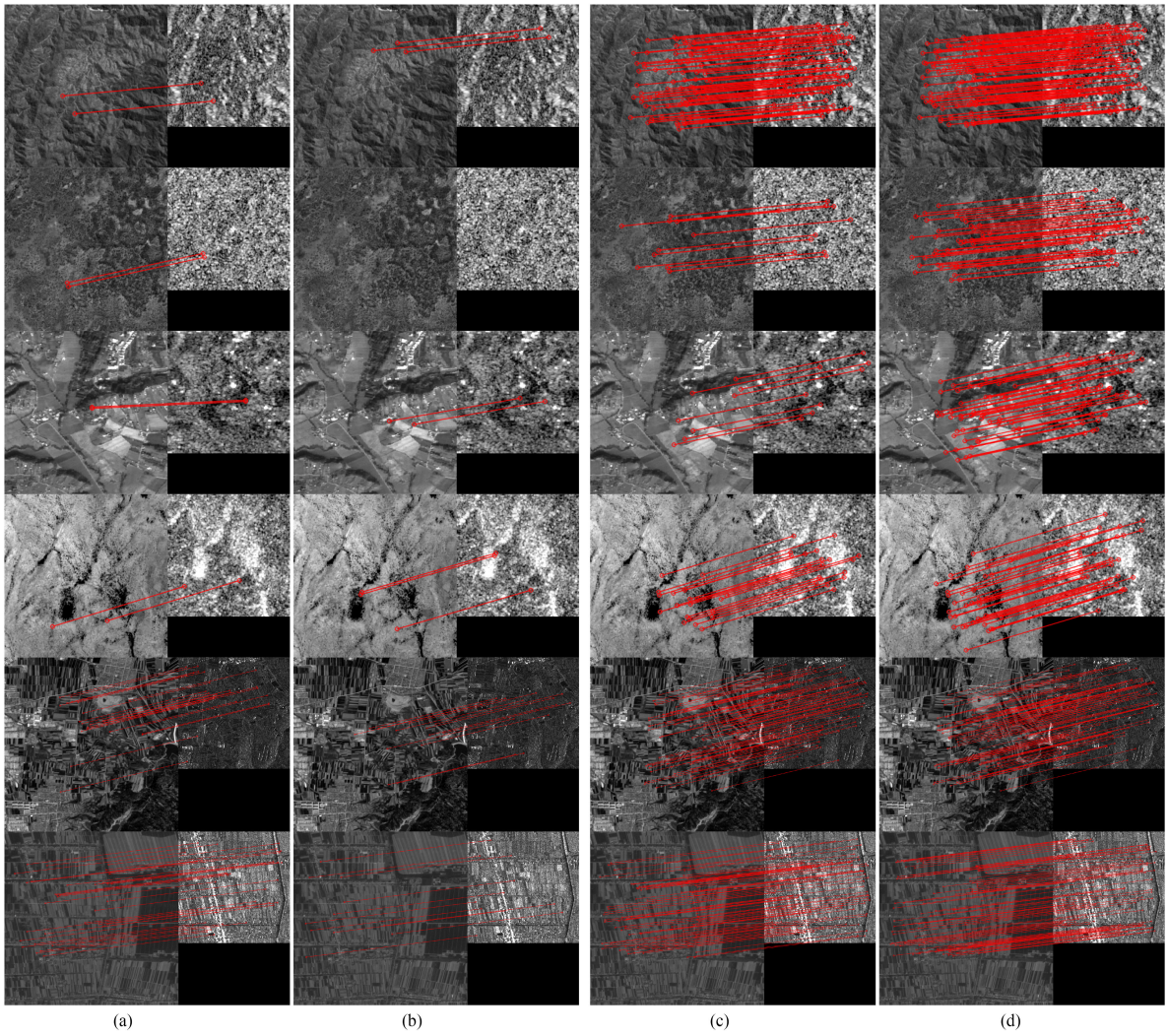


Fig. 14. Feature matching graphs generated by different methods. (a) TFeat (b) MatchNet (c) HardNet (d) MatchosNet.

As illustrated in Fig. 8(a), a Gaussian weight template is designed with the size of 7×7, and its 3-D distribution is shown in Fig. 8(b). The function sets $\mu_1 = 3.5, \mu_2 = 3.5, \sigma_1 = 7$, and $\sigma_2 = 7$. The weight of each position can be expressed as

$$W_{ij} = f(i,j) \quad f(i,j) \sim N\left(\mu_1 = 3.5, \mu_2 = 3.5, \sigma_1 = 7\sigma_2 = 7\right). \tag{11}$$

As shown in Fig. 9(a), each candidate position can assign certain weights to the pixels of the optical image through the weight template, and the final voting value can be obtained after multiple rounds of accumulation of these weights. The formula can be expressed as follows:

$$V_{ij} = \sum w_{ij}. \tag{12}$$

The distribution of the function can be roughly expressed in Fig. 9(b). Finally, the coordinates of the position is selected by the maximum V value, which is the final result of position matching between the SAR image and optical image.

IV. Experiments and Analysis

A. Data Set

The experiment uses three different datasets, which are SEN1-2 dataset [43], SARptical dataset [44], and the dataset of corresponding optical images and SAR images, which are arranged by us.

SEN1-2 dataset [43] was proposed by Schmitt *et al.* in 2017. SEN1-2 compared 282,384 corresponding image blocks collected from all parts of the globe and all weathers seasons. In the experiment of this article, the summer and winter parts of SEN1-2 dataset were used, with 48 158 images and 60 104 images, respectively.

SARptical dataset [44] was proposed by Wang *et al.* in 2018. The dataset consists of over 10,000 pairs of corresponding SAR and optical image patches extracted from TerraSAR-X high resolution spotlight images and aerial UltraCAM optical images.

To improve the matching effect, the normalization operation are performed on the SAR image. The normalization operation can increase the detail of the image, especially the part that is too bright or too dark. Fig. 10 shows the normalization operation of

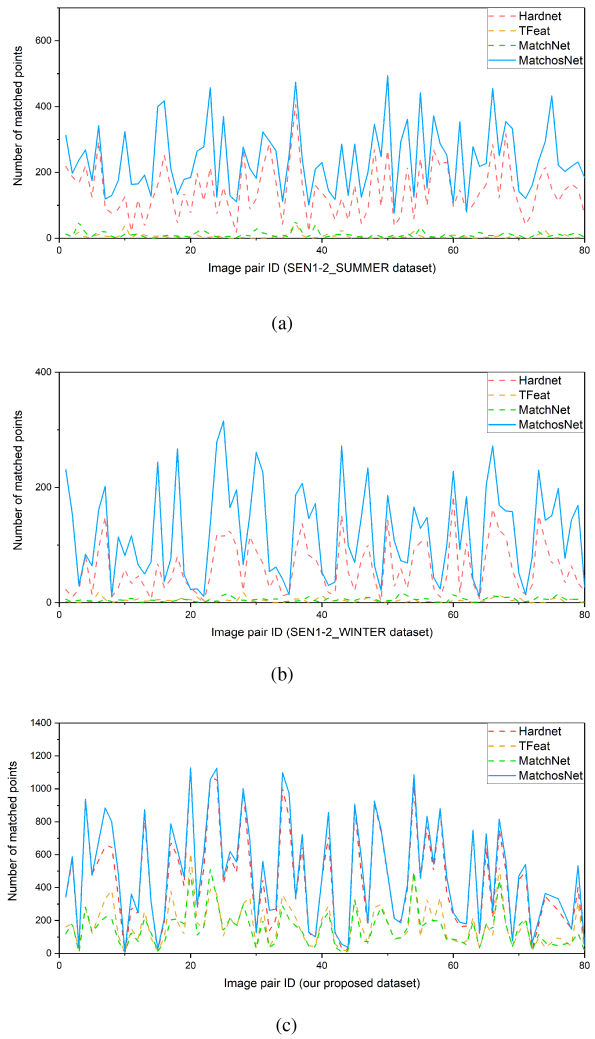


Fig. 15. Acquired number of matched points of each image pair for four different methods from three different datasets.

the SAR image. (a) and (b) are two optical images from different angles under the same key point, (c) represents the SAR image without normalization, and (d) represents the SAR image with normalization.

As shown in Fig. 11, a large number of corresponding optical and SAR images of China are collected. Since the acquisition of SAR images is much more difficult than that of optical images in real life, the SAR images are randomly cut into 512*512 and the optical images are cut into 800*800, including the content of SAR images.

The datasets are divided into six categories: port, urban area, river system, airport, island, and plains. The entire dataset has a total of 192 000 images, of which 96 000 are optical images and 96 000 are SAR images. There are six categories in the dataset, and each category has 16 000 optical images and 16 000 SAR images, respectively.

In the training set of the above three datasets, the experiment marked the specific position of the corresponding SAR image on the optical image (the coordinate of the pixel in the upper left corner), so as to use deep learning to learn the feature and position later. In the test set, nothing is done to the position coordinates.

To justify our proposed dataset, the three datasets are validated by the L2 distance function. The positive and negative distance of L2 represents the distribution of positive and negative samples. In Fig. 12, the horizontal coordinate represents the L2 distance and the vertical coordinate represents the multiplicity of the average [50]. The higher multiplicity of positive and negative distance indicates the better distribution of positive and negative samples of the images. It can be seen that the dataset collected by us has the same good distribution of positive and negative samples as the other two publicly available authoritative datasets.

In addition, a large number of images are randomly selected from different datasets and processed with Gaussian noise and salt-and-pepper noise. As shown in Fig. 13, the optical image and SAR image with the addition of two kinds of noise are obviously more challenging in feature matching and position matching.

B. Baseline

The effectiveness of MatchosNet is compared with three effective methods.

1) Mishchuk *et al.* [41] proposed the HardNet in 2017. The loss they proposed in this article was better than complex regularization methods, maximizing the distance between the closest positive and closest negative examples in a batch. The HardNet model worked well for both shallow and deep convolutional network architectures.
2) Balntas *et al.* [40] implemented TFeat and proposed to utilize triplets of training samples, together with in-triplet mining of hard negatives. Experiments showed that this method obtained excellent results compared to other methods, with lower complexity of the network structure of the model and without the typical computational overhead associated with mining negation. The authors also examined different loss functions associated with triplets and they found that Margin ranking loss worked best. Therefore, we used Margin ranking loss as the loss function of the TFeat in this article, which made the comparative experiments more objective and reasonable.
3) MatchNet was suggested by Han *et al.* [23]. A new approach using deep network architecture based on patch matching was proposed to significantly improve the results, using fewer descriptors than other methods. It had been experimentally proven that MatchNet was highly competitive compared to other methods of the same type. According to the authors of this article, MatchNet worked best when it did not use a full connection layer. Therefore, we used MatchNet without full connection layer in the comparison experiment to make the comparative experiment objective and fair.
4) Lowe [51] proposed SIFT. It was a traditional algorithm that did not rely on deep learning and could be used to achieve reliable matching between different views of objects or scenes. Sift algorithm sought the extreme point in the spatial scale, and extracted its location, scale and

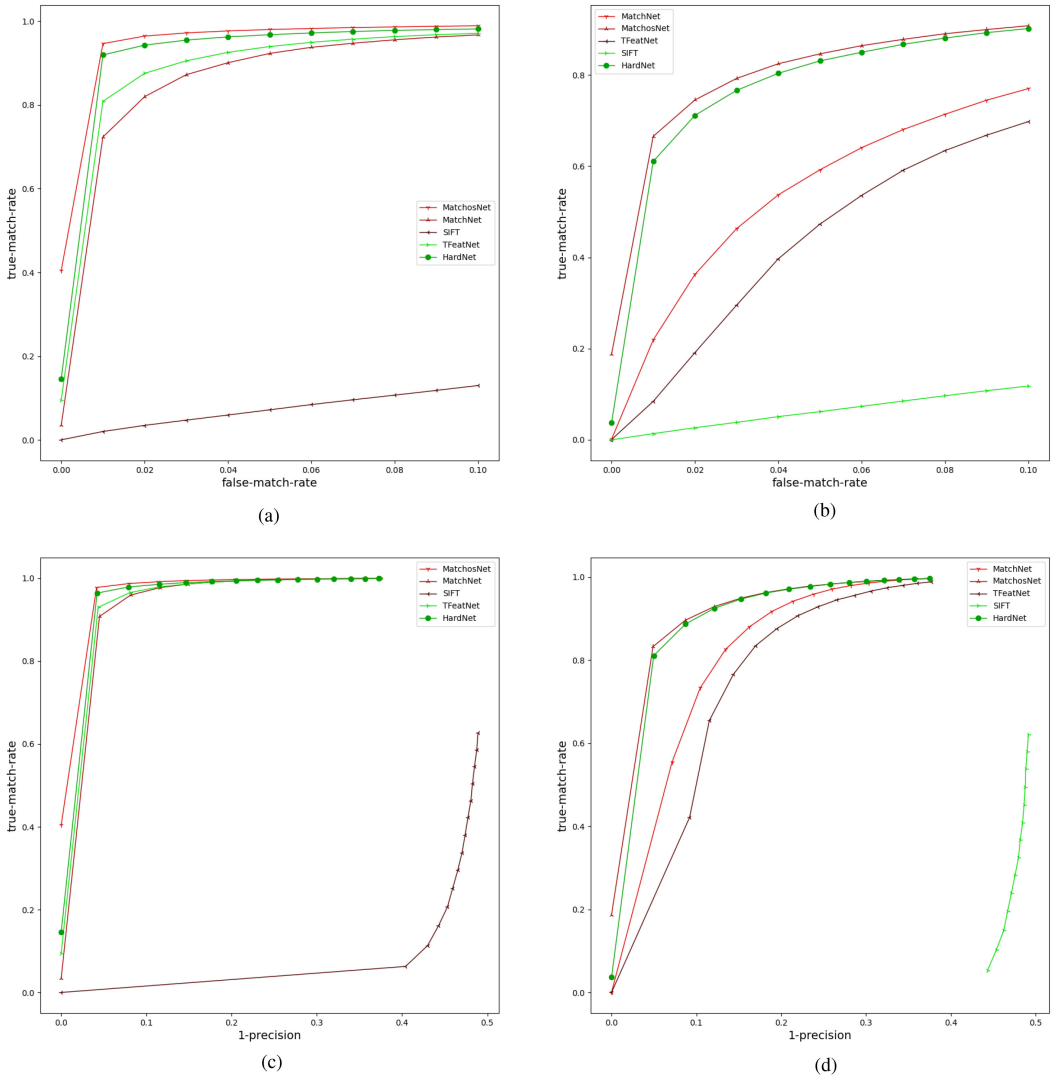


Fig. 16. ((False-match-rate)True-match-rate) curves and ((1-Precision) True-match-rate) curves with different methods. (a) No Noise FMR-TMR (b) With Noise FMR-TMR (c) No Noise (1-P)-TMR (d) With Noise (1-P)-TMR.

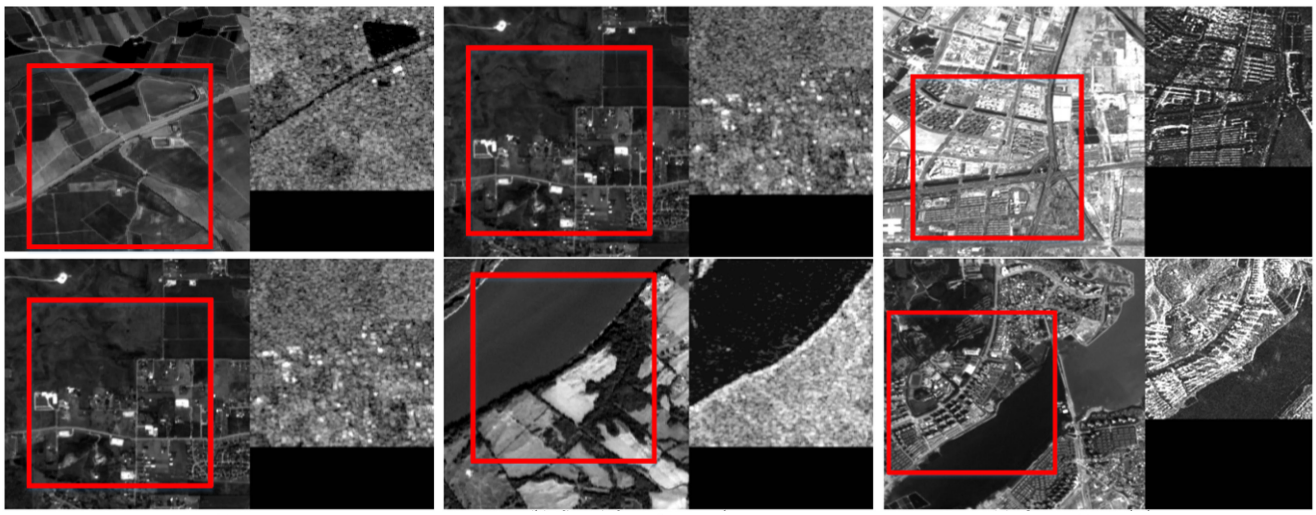


Fig. 17. Position matching graph of MatchosNet in different datasets. (a) SEN1-2 SUMMER dataset (b) SEN1-2 WINTER dataset (c) Our proposed dataset.

TABLE II
AVERAGE ERROR XRMSE,YRMSE, AND XYRMSE OF MATCHED POINTS FOR FOUR DIFFERENT METHODS

Different Methods	XYrmse	Xrmse	Yrmse
TFeat	1.3383	0.7616	0.7537
MatchNet	1.3403	0.7452	0.7714
HardNet	1.3044	0.6793	0.7870
MatchosNet	1.3033	0.7432	0.7415

TABLE III
POSITION FEATURE MATCHING RESULT IN DIFFERENT DATASETS

Different Methods	Match count	XYrmse	Xrmse	Yrmse
TFeat	816	1.3084	0.9608	0.6029
MatchNet	808	1.4531	1.0248	0.7413
HardNet	830	1.1768	0.8614	0.5325
MatchosNet	833	1.1647	0.8558	0.5212

TABLE IV
PERFORMANCE OF THE DIFFERENT NETWORKS

Methods	SEN1-2 SUMMER dataset		SARptical dataset	
	AUC	FPR80	AUC	FPR80
TFeat	0.9859	0.10000	0.9004	0.1647
MatchNet	0.9625	0.05813	0.9010	0.1594
HARDNet	0.9881	0.00075	0.9575	0.0445
MatchosNet	0.9899	0.0001	0.9810	0.0168

TABLE V
COMPUTATIONAL COMPLEXITY AND TIME PERFORMANCE FOR FOUR DIFFERENT METHODS IN THE TESTING PROCESS

Different Methods	Params (M)	Computational Complexity (GMac)	Inference Time Consume (ms)
TFeat	0.34	0.06	2.652
HardNet	1.34	0.05	1.897
MatchNet	1.34	0.05	1.486
MatchosNet	0.39	0.02	1.271

other information. It provided a great reference value for the following methods.

C. Contrast Feature Matching Tests of Different Methods

1) The Feature Matching Test: In the feature matching test, the experiment evaluates the performance of MatchosNet, HardNet, TFeat, and MatchNet trained with different training datasets (SEN1-2 SUMMER dataset, SEN1-2 WINTER dataset and our proposed dataset) to judge whether the two patches correspond to each other. These four methods use the same dataset and are trained in the batch of same size on the same server to ensure the objectivity and fairness of the experiments. In the test of the experiment, if the distance error of the corresponding matching points in the optical image and SAR image is less than 2 pixels, they will be regarded as a pair of correct matching points.

Fig. 14 compares the proposed MatchosNet with the state-of-the-art methods. The top two sets of images are the results of the SEN1-2 SUMMER dataset, the middle two sets of images are the results of the SEN1-2 WINTER dataset, and the bottom two sets of images are the results of our proposed dataset. As can be seen in Fig. 14, the MatchosNet method is able to obtain more feature matching points on the same dataset compared to the other three excellent methods, indicating that MatchosNet has a very strong capability in solving the feature matching problem.

TABLE VI
IMPACT OF NETWORK AND LOSS FUNCTION ON POSITION MATCHING

Method	XYrmse	Xrmse	Yrmse	Average Correct number in a batch (64)
w/o CSPNet	1.1257	0.7888	0.6034	62.66
w/o DenseNet	1.2685	0.8271	0.7103	62.13
w/o ARCpatch loss	1.4242	0.9120	0.8472	60.25
w/o ARCpatch and hard l2 loss	2.6775	1.7213	1.6721	20.62
MatchosNet	1.0586	0.6774	0.5806	63.19

Fig. 15(a)–(c), respectively, show the number of correctly matched feature points of different methods in the three different datasets. It can be seen that, under the same conditions, MatchosNet can obtain more correctly matched feature points than the other three excellent methods, which further proves that MatchosNet has a strong ability of learning and matching features.

Then, we calculate the final registration results of SAR-optics using a variety of datasets including Gaussian noise and salt-and-pepper noise. By calculating the matching results of Optical-Patch and SAR-Patch, the experiment uses a criterion [52], [53] calculated based on the number of true and false matches obtained per image pair. Assume two detected key points, A, and B, with their descriptors, DA, and DB, are selected from reference and target images, respectively. If the distance between descriptors DA and DB is below a threshold T and simultaneously, A and B are correct matches verified by ground-truth (Correspondence regions data), A and B will be the true match. If A and B are not the correct matches confirmed by ground truth, but the distance between the descriptors DA and DB is less than T, then A and B are false matches, and vice versa.

True-match-rate, False-match-rate, and 1-precision factors are declared as follows:

$$\text{True-match-rate} = \frac{\text{Number of trueMatch}}{\text{Number of Correspondence}} \tag{13}$$

$$\text{False-match-rate} = \frac{\text{Number of falseMatch}}{\text{Number of non correspondence}} \tag{14}$$

$$1 - \text{precision} = \frac{\text{Number of falseMatch}}{\text{Number of trueMatch} + \text{Number of falseMatch}}. \tag{15}$$

The threshold T is varied to obtain the curves. A perfect descriptor would give a recall equal to 1 for any precision [52], [53]. In another word, both the curve [(1-Precision) True-match-rate] and the curve [(False-match-rate)True-match-rate] are above and left, the efficiency of its algorithm is higher. As

shown in Fig. 16, (a) is the [(False-match-rate) True-match-rate] curve generated by the dataset without added noise; (b) is the [(False-match-rate)True-match-rate] curve generated by the dataset with Gaussian noise and salt-and-pepper noise; (c) is the [(1-Precision) True-match-rate] curve generated by the dataset without added noise; (d) is the [(1-Precision) True-match-rate] curve generated by the dataset with Gaussian noise and salt-and-pepper noise. The MatchosNet method performed best on both measures regardless of the dataset and regardless of whether noise was added, which proved the superiority of the MatchosNet method.

2) Distance Error Test of Feature Matching: To prove the accuracy of the matching points, the distance error of the matching feature points are detected. The experiment uses Xrmse for error in horizontal distance, Yrmse for error in vertical distance, and XYrmse for error in distance on the image. The units of all the error measurements are pixels. The functions of Xrmse,Yrmse, and XYrmse are calculated as follows:

$$x_{\mathrm{rmse}} = \sqrt{\frac{1}{N}\sum_{i}\left(x_i^1 - x_i^2\right)^2} \tag{16}$$

$$y_{\mathrm{rmse}} = \sqrt{\frac{1}{N}\sum_{i}\left(y_i^1 - y_i^2\right)^2} \tag{17}$$

$$\mathrm{xy}_{\mathrm{rmse}} = \sqrt{\frac{1}{N}\sum_{i}\left(\left(x_i^1 - x_i^2\right)^2 + \left(y_i^1 - y_i^2\right)^2\right)} \tag{18}$$

where (x_i^1, y_i^1) denotes the coordinates of the matching points in the SAR image, (x_i^2, y_i^2) indicates the coordinates of the matching point in the optical image. *N* denotes the total number of matched points.

Table II shows the average error Xrmse, Yrmse, and XYrmse of the different methods. MatchosNet has the lowest average error XYrmse in the three datasets and its Xrmse and Yrmse are also very excellent. It indicates that MatchosNet has very high precision of feature matching points and strong feature matching ability.

D. Position Matching Tests of MatchosNet

Fig. 17 shows the implementation of position matching between SAR images and optical images made by MatchosNet. It can be seen that MatchosNet can achieve an accurate position matching of SAR image in optical image under different datasets. The experiment uses the same key point extraction method to extract features from MatchosNet, MatchNet, TFeat, and HardNet, and then uses the position matching algorithm presented in this article to conduct position matching test. Table III shows the position feature matching result in different datasets. We extracted a total of 850 optical and SAR images from SEN1-2 SUMMER Dataset, SEN1-2 WINTER Dataset and our proposed dataset and calculated their matching coordinates in the upper left corner. In the position matching between SAR image and optical image, when the distance error of image position matching is less than 5.0 pixels, it is regarded as a pair of correct matching images.

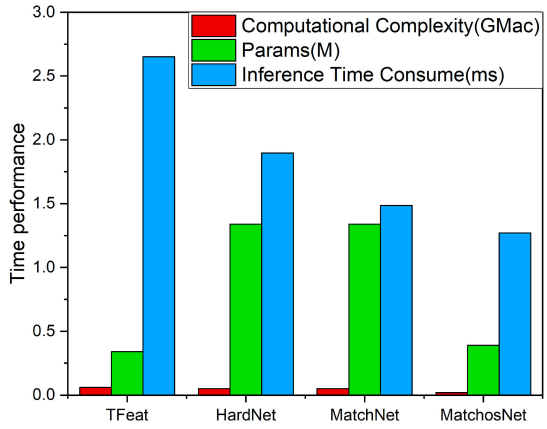


Fig. 18. Computational complexity and time performance of the testing process.

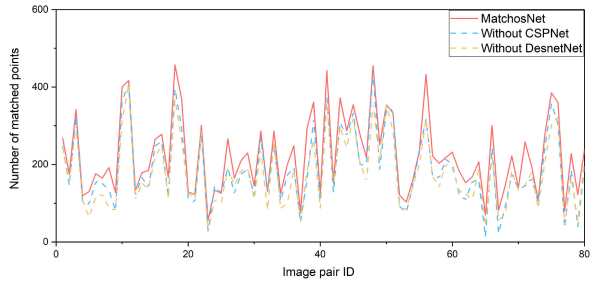


Fig. 19. Impact of network on feature matching.

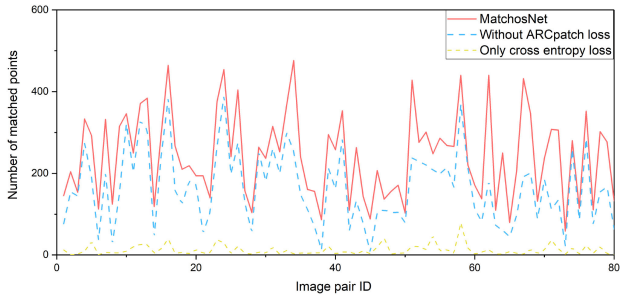


Fig. 20. Impact of loss function on feature matching.

As can be seen from Table III, compared with other methods, MatchosNet matches the most images in the correct position and also has the least XYrmse, Xrmse, and Yrmse. These experimental results are sufficient to prove that MatchosNet's position matching capability is very competitive and has great practical value.

E. Classification Benchmark Test

In the test dataset, the optical and SAR images are completely scrambled. Through the classification experiment of this step, it can be proved that MatchosNet has the ability of one-to-one correspondence between optical images and SAR images.

The metrics of area under curve (AUC) and fpr80 (false positive rate at point of 0.80 true positive recall) are reported in the Table IV. The AUC ideal value is 1. The larger the AUC is, the better the network performe. For the FPR80, the smaller

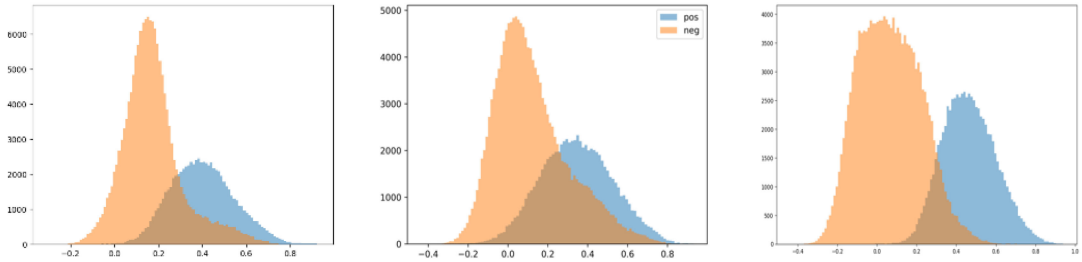


Fig. 21. Angle distributions of all positive pairs and random negative pairs (0.5 M) from SEN1-2 dataset. Orange area indicates positive pairs while blue indicates negative pairs. All angles are represented in degree. (a) L_{ARCpatch} (b) L_{hardl2} (c) $L_{\text{ARCpatch}} + L_{\text{hardl2}}$.

it is, the better the network performe. The FPR80's ideal value is 0. The Table IV shows the performance of the different networks trained from SEN1-2 SUMMER dataset and SARptical dataset. MatchosNet has the highest AUC value and the lowest FPR80 value in both datasets. It can prove that MatchosNet is very competitive in terms of classification capability.

F. Evaluation of Computational Complexity and Time Performance

As shown in in Algorithm 1 in Section III-C, the pseudocode of MatchosNet in the training process is showed. It can be concluded that the computational complexity of MatchosNet in the training process is $o(n^2)$. In the testing process, as shown in Table V and Fig. 18, MatchosNet is compared in detail with several other comparison algorithms.

Obviously, the computational complexity and inference time consume of MatchosNet are minimal, and the number of parameters is very small. Although TFeat has fewer parameters than MatchosNet, it has the highest computational complexity and inference time consume. Through the experiment, it can be obviously found that the computational complexity and time performance of MatchosNet algorithm is far better than the comparison algorithm.

G. Ablation Studies and Analysis

Ablation studies are conducted to evaluate the effect of the network and loss function. The results for the feature matching with different networks are shown in Fig. 19, MatchosNet is compared with the model without CSPNet and DenseNet. The results for the feature matching with different loss functions are shown in Fig. 20, MatchosNet is compared with the model without ARCpatch loss and the model with only cross entropy loss. It is clear that MatchosNet is able to obtain more feature matches than the other methods when they are tested with the same dataset under the same conditions. As illustrated in Fig. 21, the distributions of all positive pairs and random negative pairs are tested by ablating partial loss function. Obviously, the compound loss function has a more distinct margin than ARCpatch-Loss and hard l2-Loss.

Table VI shows the results of position matching for MatchosNet compared to the other methods. It is obvious that MatchosNet has the lowest Xrmse, Yrmse, and XYrmse and has more correct images.

V. Conclusion

In this article, a new deep learning method—MatchosNet is designed to implement the feature matching between optical images and SAR images with size differences, and further implemented the position matching of SAR images on the optical images. A new dataset is proposed by collecting a large number of corresponding SAR images and optical images. Then, we created samples of training patches for optical and SAR images and designed a special network with dense blocks and cross stage partial networks to generate deep feature descriptors. In addition, a compound loss function composed of the hard L2 loss function and ARCpatch loss function is designed. Finally, a 2-D Gaussian function voting algorithm is designed to match the position of the SAR images and optical images.

In the experiment, we first collected a large number of SAR images and optical images of China, and put forward a dataset of corresponding SAR images and optical images. The collated dataset was compared with the SEN1-2 dataset and the SARpitcal dataset. It had been shown that the proposed dataset was reasonable and useful. The MatchosNet was compared with several other excellent methods on different datasets, and the experimental results showed that the feature matching effect of MatchosNet was obviously better than other methods. In addition, MatchosNet was also compared with other methods in the test of computational complexity and time performance, which proved that MatchosNet had excellent computational complexity and time performance.

In the future, we will further explore better networks and loss functions by mathematical research in the field of feature matching, so as to design better feature matching methods and make more further applications.

Disclosures

The authors declare no conflict of interest.

References

[1] G. Tochon, M. D. Mura, M. A. Veganzones, T. Géraud, and J. Chanussot, "Braids of partitions for the hierarchical representation and segmentation of multimodal images," *Pattern Recognit.*, vol. 95, pp. 162–172, 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.05.029

[2] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021. [Online]. Available: https://doi.org/10.1109/TGRS.2020.3016820

[3] A. Pournemat, P. Adibi, and J. Chanussot, "Semisupervised charting for spectral multimodal manifold learning and alignment," *Pattern Recognit.*, vol. 111, 2021, Art. no. 107645. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107645

[4] T. Ma, J. Ma, K. Yu, J. Zhang, and W. Fu, "Multispectral remote sensing image matching via image transfer by regularized conditional generative adversarial networks and local feature," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 2, pp. 351–355, Feb. 2021. [Online]. Available: https://doi.org/10.1109/LGRS.2020.2972361

[5] 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 Geosci. Remote Sens. Mag.*, to be published, doi: 10.1109/MGRS.2021.3081763.

[6] A. Wong and D. A. Clausi, "ARRSI: Automatic registration of remote-sensing images," *IEEE Trans. Geosci. Remote. Sens.*, vol. 45, no. 5/2, pp. 1483–1493, May 2007. [Online]. Available: https://doi.org/10.1109/TGRS.2007.892601

[7] Y. Ban and A. W. Jacob, "Object-based fusion of multitemporal multiangle ENVISAT ASAR and HJ-1B multispectral data for urban land-cover mapping," *IEEE Trans. Geosci. Remote. Sens.*, vol. 51, no. 4/1, pp. 1998–2006, 2013. [Online]. Available: https://doi.org/10.1109/TGRS.2012.2236560

[8] D. K. Seo, Y. Kim, Y. D. Eo, M. H. Lee, and W. Y. Park, "Fusion of SAR and multispectral images using random forest regression for change detection," *ISPRS Int. J. Geo. Inf.*, vol. 7, no. 10, 2018, Art. no. 401. [Online]. Available: https://doi.org/10.3390/ijgi7100401

[9] F. Argenti, A. Lapini, T. Bianchi, and L. Alparone, "A tutorial on speckle reduction in synthetic aperture radar images," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 3, pp. 6–35, Sep. 2013.

[10] M. F. Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "Sar-to-optical image translation based on conditional generative adversarial networks - Optimization, opportunities and limits," *Remote. Sens.*, vol. 11, no. 17, 2019, Art. no. 2067. [Online]. Available: https://doi.org/10.3390/rs11172067

[11] Z. Li, D. Mahapatra, J. A. W. Tielbeek, J. Stoker, L. J. van Vliet, and F. M. Vos, "Image registration based on autocorrelation of local structure," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 63–75, Jan. 2016. [Online]. Available: https://doi.org/10.1109/TMI.2015.2455416

[12] X. Xiong, Q. Xu, G. Jin, H. Zhang, and X. Gao, "Rank-based local self-similarity descriptor for Optical-to-SAR image matching," *IEEE Geosci. Remote. Sens. Lett.*, vol. 17, no. 10, pp. 1742–1746, Oct. 2020. [Online]. Available: https://doi.org/10.1109/LGRS.2019.2955153

[13] S. Cao, H. Shen, S. Chen, and C. Li, "Boosting structure consistency for multispectral and multimodal image registration," *IEEE Trans. Image Process.*, vol. 29, pp. 5147–5162, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.2980972

[14] E. Ferrante and N. Paragios, "Slice-to-volume medical image registration: A survey," *Med. Image Anal.*, vol. 39, pp. 101–123, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.04.010

[15] Y. Zhang, Z. Wan, X. Jiang, and X. Mei, "Automatic stitching for hyperspectral images using robust feature matching and elastic warp," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 13, pp. 3145–3154, 2020. [Online]. Available: https://doi.org/10.1109/JSTARS.2020.3001022

[16] Q. Guo, M. He, and A. Li, "High-resolution remote-sensing image registration based on angle matching of edge point features," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 11, no. 8, pp. 2881–2895, Aug. 2018. [Online]. Available: https://doi.org/10.1109/JSTARS.2018.2844295

[17] J. Dong and S. Soatto, "Domain-size pooling in local descriptors: DSP-SIFT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 5097–5106. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7299145

[18] D. Bhattacharjee and H. Roy, "Pattern of local gravitational force (PLGF): A novel local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 595–607, Feb. 2021. [Online]. Available: https://doi.org/10.1109/TPAMI.2019.2930192

[19] T. Ma, J. Ma, and K. Yu, "A local feature descriptor based on oriented structure maps with guided filtering for multispectral remote sensing image matching," *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 951. [Online]. Available: https://doi.org/10.3390/rs11080951

[20] M. A. Ghannadi and M. Saadatseresht, "A modified local binary pattern descriptor for SAR image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 568–572, Apr. 2019. [Online]. Available: https://doi.org/10.1109/LGRS.2018.2876661

[21] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu, "Robust Optical-to-SAR image matching based on shape properties," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 564–568, Apr. 2017. [Online]. Available: https://doi.org/10.1109/LGRS.2017.2660067

[22] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017. [Online]. Available: https://doi.org/10.1109/TGRS.2017.2656380

[23] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3279–3286. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298948

[24] Y. Tian, B. Fan, and F. Wu, "L2-Net: Deep learning of discriminative patch descriptor in euclidean space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6128–6136. [Online]. Available: https://doi.org/10.1109/CVPR.2017.649

[25] Z. Zhou, Q. M. J. Wu, S. Wan, W. Sun, and X. Sun, "Integrating SIFT and CNN feature matching for partial-duplicate image detection," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 5, pp. 593–604, Oct. 2020. [Online]. Available: https://doi.org/10.1109/TETCI.2019.2909936

[26] Y. Dong *et al.*, "Local deep descriptor for remote sensing image feature matching," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 430. [Online]. Available: https://doi.org/10.3390/rs11040430

[27] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, 2018, pp. 224–236. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018_workshops/w9/html/DeTone_SuperPoint_Self-Supervised_Interest_CVPR_2018_paper.html

[28] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 7260–7268. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Revisiting_Local_Descriptor_Based_Image-To-Class_Measure_for_Few-Shot_Learning_CVPR_2019_paper.html

[29] 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, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621001842

[30] L. H. Hughes, N. Merkle, T. Bürgmann, S. Auer, and M. Schmitt, "Deep learning for SAR-optical image matching," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, 2019, pp. 4877–4880. [Online]. Available: https://doi.org/10.1109/IGARSS.2019.8898635

[31] P. Kovesi, "Image features from phase congruency," *Videre J. Comput. Vis. Res.*, vol. 1, Jan. 1999.

[32] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, Minnesota, USA, 2007, pp. 1–8. [Online]. Available: https://doi.org/10.1109/CVPR.2007.383198

[33] J. Fan, Y. Wu, M. Li, W. Liang, and Y. Cao, "SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5368–5379, Sep. 2018. [Online]. Available: https://doi.org/10.1109/TGRS.2018.2815523

[34] K. Gao, H. Aliakbarpour, G. Seetharaman, and K. Palaniappan, "DCT-based local descriptor for robust matching and feature tracking in wide area motion imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1441–1445, Aug. 2021. [Online]. Available: https://doi.org/10.1109/LGRS.2020.3000762

[35] C. A. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, vol. 12, no. 9, pp. 12661–12672, 2012. [Online]. Available: https://doi.org/10.3390/s120912661

[36] C. F. G. Nunes and F. L. C. Pádua, "A local feature descriptor based on log-gabor filters for keypoint matching in multispectral images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 14, no. 10, pp. 1850–1854, Oct. 2017. [Online]. Available: https://doi.org/10.1109/LGRS.2017.2738632

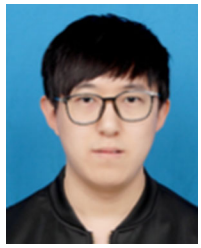
[37] Z. Fu, Q. Qin, B. Luo, C. Wu, and H. Sun, "A local feature descriptor based on combination of structure and texture information for multispectral image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 100–104, Jan. 2019. [Online]. Available: https://doi.org/10.1109/LGRS.2018.2867635

[38] H. Qian, J. Yue, M. Chen, M. Wang, and H. Xin, "Subpixel-level edge feature matching for SAR and optical images based on zernike moments," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Waikoloa, HI, USA, 2020, pp. 2723–2726. [Online]. Available: https://doi.org/10.1109/IGARSS39084.2020.9324579

[39] X. Jiang *et al.*, "Robust feature matching for remote sensing image registration via linear adaptive filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1577–1591, Feb. 2021. [Online]. Available: https://doi.org/10.1109/TGRS.2020.3001089
[40] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. Brit. Mach. Vis. Conf.*, York, U.K., 2016. [Online]. Available: http://www.bmva.org/bmvc/2016/papers/paper119/index.html
[41] A. Mishchuk *et al.*, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 4826–4837. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/831caa1b600f852b7844499430ecac17-Abstract.html
[42] W. Du, Y. Zhou, J. Zhao, X. Tian, Z. Yang, and F. Bian, "Exploring the potential of unsupervised image synthesis for SAR-optical image matching," *IEEE Access*, vol. 9, pp. 71022–71033, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3079327
[43] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *CoRR*, vol. abs/1807.01569, 2018. [Online]. Available: http://arxiv.org/abs/1807.01569
[44] Y. Wang and X. X. Zhu, "The sarptical dataset for joint analysis of SAR and optical image in dense urban area," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 6840–6843. [Online]. Available: https://doi.org/10.1109/IGARSS.2018.8518298
[45] A. Foi *et al.*, "Difference of Gaussians revolved along elliptical paths for ultrasound fetal head segmentation," *Comput. Med. Imag. Graph.*, vol. 38, no. 8, pp. 774–784, 2014. [Online]. Available: https://doi.org/10.1016/j.compmedimag.2014.09.006
[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, United States, 2012, pp. 1106–1114. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90
[48] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2261–2269. [Online]. Available: https://doi.org/10.1109/CVPR.2017.243
[49] C. Wang, H. M. Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 1571–1580. [Online]. Available: https://doi.org/10.1109/CVPRW50498.2020.00203
[50] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 4690–4699. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94
[52] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005. [Online]. Available: https://doi.org/10.1109/TPAMI.2005.188
[53] B. Sadeghi, K. Jamshidi, A. Vafaei, and S. A. Monadjemi, "A local image descriptor based on radial and angular gradient intensity histogram for blurred image matching," *Vis. Comput.*, vol. 35, no. 10, pp. 1373–1391, 2019. [Online]. Available: https://doi.org/10.1007/s00371-018-01616-z
[54] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed., Harlow, U.K.: Addison-Wesley, 1999.

Yun Liao He received the BE degree in communication engineering and the M.S. degree in software engineering from the Yunnan University, Kunming, China, in 2002 and 2009, respectively, and the Ph.D. degree in software engineering from Yunnan University, in 2012.

He is currently a Lecturer with the School of Software, Yunnan University, China. His current research interests include machine learning, image processing, and deep learning.



Yide Di received the B.E. and M.E. degrees in software engineering from Yunnan University, Kunming, China, in 2017 and 2021, respectively. He is currently working toward the Ph.D. degree with the School of Information Science and Technology, Dalian Maritime University, Dalian, China.

His current research interests include machine learning, image processing, and deep learning.

Hao Zhou graduated from Yunnan University, Kunming, China, in June 2018.

He join the Yunnan Lanyi Network Technology Co. His current research interests include computer vision and deep learning.

Anran Li received the BE degree in electrical engineering from Guangxi University, Nanning, China, in 2018.

He is currently working as an Algorithmic Engineer with Lanyi Network Technology Co, China. His research interests include deep learning and image processing.

Junhui Liu received the Ph.D. degree in systems analysis and integration from the Yunnan University, Kunming, China, in 2009.

He is currently a Lecturer with the School of Software, Yunnan University, China. His current research interests include deep learning and domain-specific modelling.

Mingyu Lu born in 1963. He received the doctor's degree from Tsinghua University, Beijing, China, in 2002.

He is currently a Profession and Doctoral Supervisor with Dalian Maritime University. His research interests include data mining, pattern recognition, machine learning and natural language processing.

Qing Duan born in 1975. She received the Ph.D. degree in software engineering from De Montfort University, Leicester, U.K., in 2010.

She is currently an Associate Researcher with the School of Software, Yunnan University, Key Laboratory in Software Engineering of Yunnan Province, China. Her current research interests include data mining and knowledge engineering, intelligent information processing, and software engineering.