A Novel Region-Based Image Registration Method for Multisource Remote Sensing Images Via CNN

Liang Zeng, Yanlei Du, *Member, IEEE*, Huiping Lin, Jing Wang, Junjun Yin, and Jian Yang, *Senior Member, IEEE*

***Abstract*—The comprehensive utilization of images from various satellite sensors can significantly increase the performance of remote sensing applications and has, therefore, attracted extensive research attention. One of the essential challenges that research encounters comes from multisource image registration. This article proposes a novel region-based image registration method for multisource images. The proposed method exploits the region features of input images, which provide more consistent and common information of the multisource data. The image region features are extracted based on image semantic segmentation using the deep convolutional neural network approach. The final registration result is a pixel-level output corresponding to the input images. The proposed registration scheme overcomes the limits of traditional feature extraction methods (e.g., point feature) adopted in previous registration schemes. Results indicate that the proposed method has good performance for the multisource remote sensing image registration and can serve as a building block for the fusion of multisource images.**

***Index Terms*—Image registration, radar imaging.**

I. Introduction

In RECENT years, significant improvements have been achieved in remote sensing (RS) sensors. As a consequence, massive and various multisource RS images (e.g., optical imaging, synthetic aperture radar (SAR), light detection and ranging, and multispectral and hyperspectral data) can be obtained [1]. The increasing amount of high-quality satellite RS imagery provides massive opportunities in civilian and military applications, e.g., land-cover and land-use analysis, agriculture, and forestry monitoring, change detection, disaster reduction, and data fusion [2]–[11].

Manuscript received September 14, 2020; revised December 6, 2020; accepted December 17, 2020. Date of publication December 28, 2020; date of current version January 21, 2021. This work was supported in part by NSFC under Grant 61771043 and Grant 61701454, in part by National Key Research and Development Program of China under Grant 2017YFB0502703, in part by Nature Science Foundation for Young Scientists of Jiangsu Province, China under Grant BK20160147, and in part by the China Postdoctoral Science Foundation under Grant 2020M680554. *(Corresponding author: Liang Zeng.)*

Liang Zeng, Yanlei Du, Huiping Lin, and Jian Yang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: zengliang14@mails.tsinghua.edu.cn; duyanlei@mail.tsinghua.edu.cn; linhp15@mails.tsinghua.edu.cn; yangjian_ee@mail.tsinghua.edu.cn).

Jing Wang is with the Science and Technology on Information System Engineering Laboratory, China Electronics Technology Group Corporation, Nanjing 320100, China (e-mail: wangjing@nuaa.edu.cn).

Junjun Yin is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: yinjj07@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2020.3047656

According to an uncompleted statistic, there are currently more than 700 RS satellites operating in space. Jointly using the multisource images obtained by different satellites and at different times would significantly improve the RS application [12]. For example, land-use monitoring typically needs to update maps frequently to continuously monitor the natural and man-made changes on the Earth's surface [13]. Therefore, better temporal resolution can be achieved by combining data from multiple sensors onboard different satellites.

One of the essential challenges of these applications lies in the correct alignment of multisource images, or the so-called *image registration* process. It is the process of transforming different sets of data into one coordinate system [14]. Image registration can be divided into two types in terms of simultaneity, namely, synchronous and asynchronous image registration. Synchronous image registration refers to matching and stitching images, which are obtained simultaneously, whereas asynchronous image registration refers to the processing of images obtained at different time intervals. A typically synchronous application is the 360° panoramic photography, and the asynchronous one is medical image registration. Another way to classify is based on the number of sensors, namely, the registration of single-sensor or multisource images. In general, the features of single-sensor images are relatively consistent, while those of multisource images show more variations, which lead to higher complexity of image registration.

Recently, a lot of research efforts have been devoted to image registration. Two kinds of commonly used registration methods are intensity-based methods and feature-based methods [15]–[20]. Intensity-based methods first normalize data into grayscale images and then calculate and compare their cross-correlation or mutual information to perform image matching. For instance, Suri and Reinartz proposed a mutual-information-based registration method [15]. Hasan *et al.* used an information-theoretic similarity measure known as cross-cumulative residual entropy to perform the registration of SAR data and optical images [16]. However, this kind of method is time consuming. Instead of normalizing the image grayscale, feature-based methods mainly exploit the invariant features in images such as corners, edges, or curves. Harris corner is one of the well-known feature extraction operators [17]. Based on these features, image registration is performed using certain similarity measures. Pan *et al.* presented a contour-based approach for multisource image registration, in which the contours are parameterized with nonuniform rational B-splines [18]. Scale-invariant feature transform (SIFT) [19] is

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

another popular approach for image registration. It can well address the issue of scale rotation of images. The SIFT algorithm is a very important method, but it may be no longer applicable when performing multisource imaging registration. This is because the image gradient information would be significantly different in such images. Based on the basic principle of SIFT, some improved methods were proposed for SAR image registration. An improved version of the SIFT is proposed by Fan *et al.* to obtain initial matching features from optical and SAR images [20].

However, challenges still exist for the proposed method especially in multisource images. There are many differences in image registration between the single-sensor image and the multisource image. For the registration of images from the same sensor, the challenges are mainly due to rotation, zoom, pan, etc., since the basic attributes of images are same. Compared to single-sensor registration, multisource image registration is more complicated. Besides the above differences, features of the same object from different sensors are usually different. The SAR image can be visualized in some ways such as Pauli decomposition and Freeman decomposition. The visualized SAR image differs greatly from the optical image.

To address the issues confronted with past methods, a novel region-based convolutional neural network (CNN) algorithm, including a network architecture and a registration procedure, is proposed for the multisource RS image registration. Unlike traditional methods, the proposed method exploits the *region features* of input images to perform registration. The image regional features represent the characteristics of the target in the image. They are extracted on the basis of image semantic segmentation via the CNN approach. Experimental results show that the regional feature registration method can achieve good results in multisource image registration. At the same time, the regional registration method is more robust and can still achieve good results within a certain error range.

Regional feature extracting can be regarded as semantic segmentation of images. Thus, semantic segmentation is the key step for registration. Over the past few decades, several kinds of image segmentation methods have been proposed, e.g., threshold-based segmentation [21]–[23], region growth segmentation [24]–[26], edge detection segmentation [27], [28], and specific-theory-based segmentation [29]–[31]. Recently, various CNN-based methods have been proposed. These methods present better performance and robustness. Some representative works are as follows. FCN is proposed in late 2014 [32]. FCN has a great improvement compared with the semantic segmentation algorithm before. However, this method ignores the correlation between the categories of pixels in the image and thus ignores the spatial smoothness of the image. U-Net is proposed in 2015 [33]. Although this work was primarily aimed at biomedical images, it has great inspiration for semantic segmentation of images in other areas. U-Net is fully symmetrical and uses the structure of an encoder and a decoder. Res-UNet [34] replaces each submodule of UNet with a form of residual connections. Recurrent Residual CNN-based U-Net (R2U-Net) [35] is a method that combines residual joining with cyclic convolution to replace the original submodule in U-Net. Attention U-Net [36] introduced an attention mechanism in U-Net. The pyramid scene parsing network (PSP-Net) [37] is another derivative of FCN. The net design includes a pyramid pooling module and also references auxiliary loss to speed up network convergence. These semantic segmentation studies make the image registration through regional features highly feasible.

In this article, two typical kinds of multisource scenarios are discussed: one is multiband SAR image registration (Bands C, L, and P), and the other one is the registration of optical image and SAR image. The rest of this article is structured as follows. Section II presents the details of the methodology. Experiment results as well as the discussions are presented in Section III. Finally, conclusions and outlooks are given in Section IV.

II. Methodology

In this section, a novel automatic registration method for multisource RS images based on the CNN is comprehensively introduced. Section II-A introduces the motivation of the proposed method. The whole algorithm architecture is introduced in Section II-B. The region features are extracted from input images after preprocessing. Then, region features are registered in the pixel level, and the corresponding transformation relationship is derived. Finally, the registered image is obtained according to the transformation relationship. Section II-C introduces the effects of region feature extraction on image registration accuracy. The key part of the algorithm is the accuracy of the region feature extraction, and our analysis shows that the relation is close to a step function, which is not the linear relationship (see Fig. 8).

A. Motivation

RS images from different sensors in different modes contain much more information than single-sensor ones since these multisource images reveal multidimensional characteristics of the object. Two typical kinds of multisource scenarios are SAR–optical image and multiband SAR image.

1) SAR–Optical Image: These two sensors are mutually complementary in the sense that SAR imaging collects information about the physical properties of the scene and follows a range-based imaging geometry, while optical imaging reflects the chemical characteristics of the scene and follows a perspective imaging geometry [38]. While SAR images provide high-resolution RS images in all-day and all-weather conditions, optical imaging is sensitive to sunlight or cloud coverage. Optical imaging is intuitive; SAR images, on the other hand, are harder to interpret. Even though many studies have transformed SAR images into pseudocolor images for the convenience of human interpretation, objects within these pseudocolor images, such as buildings, bridges, and roads, show significantly different appearances. Fig. 1 shows an example of the contrast between the SAR image (SPAN) and the optical image of the same location.The main difficulties in image registration of the SAR image and the optical image are as follows [39].

1) *Noise type:* The noise type of two kinds of image is significantly different.
2) *Radiation characteristics:* Significant differences in the imaging mechanism between these two imaging systems

(a) (b)

Fig. 1. Contrast between SAR and optical images of the same location SPAN of (a) SAR image and (b) optical image.

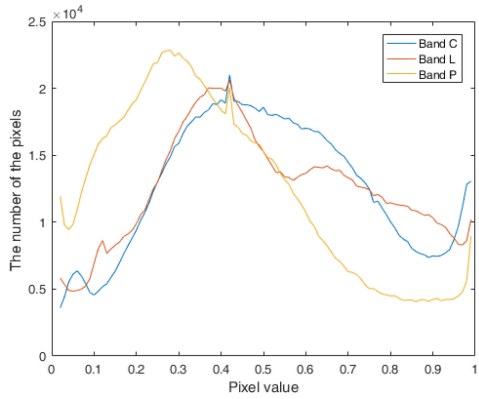


Fig. 3. Distribution of SPAN values of SAR images in different bands (Bands C, L, and P) in Flevoland.

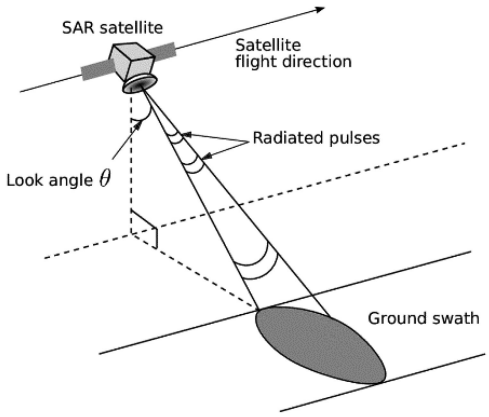


Fig. 2. Simplified geometry of a SAR system.

mean that even the same feature objects may exhibit totally different grayscale characteristics.

3) *Geometrical features:* The SAR system is a side view one to illuminate the ground scene, and a simplified geometry of SAR system is shown in Fig. 2. This inherent feature causes a lot of difficulties for registration, e.g., geometric deformation phenomena, perspective contraction, iteration, and slope shortening.

2) Multiband SAR Images: Multiband SAR images are very common data. Images of different bands can reflect the characteristics of different features. At the same time, this also makes it difficult to extract uniform features from the image. Fig. 3 shows an example that is the distribution of SPAN values of SAR images in different bands (Bands C, L, and P) in the same area (Flevoland). The main difficulties in multiband SAR image registration are as follows.

1) Ground objects have significantly different features in multisource images. Some of the key features, e.g., gray value, point, and edge, vary across different sensors and, therefore, cannot be used as registration features.
2) For asynchronous image registration, data are collected at different time periods, and changes (due to human or nature activities) may occur between these observations. As a result, key features may vary across different images. For instance, some feature points could be missing, and some edges could be interrupted.
3) Different sensors usually operate on different satellite platforms, which have various orbital heights and incident angles. These will lead to significant differences across various images.

It can be clearly seen from the above analysis that the characteristics of multisource images are significantly different at the pixel level, whether it is from the actual visual results or from the analysis of principle. However, common features still exist from a holistic perspective. The targets in image, e.g. airport runways, roads, and building areas, have one-to-one correspondence obviously. This actual ground objects are very suitable as a reference for registration images. Based on this observation, we try to divide an image into different regions and classify them into 10–20 objects as features for image registration, which are the regional features. Examples of image regional features are shown in Figs. 10 and 11. For example, the optical image I_O and the corresponding label feature image L_O is shown in Fig. 10(a) and (b). The SAR image I_S and the corresponding label feature image L_S are shown in Fig. 10(c) and (d). Obviously, the transform function T_L from L_S to L_O is equal to T_I that from I_S to I_O. And, T_L is easier to solve. Therefore, we try to simplify the problem in this way. In Section II-B, we propose a method based on this idea and the error analysis is in Section II-C.

B. CNN-Based Algorithm Architecture

The structure of the algorithm is shown in Fig. 4. The algorithm consists of three parts. Step 1 involves input and preprocessing. In this first step, the original picture has been filtered and denoised, where the SAR image and the optical image adopt different preprocessing methods. In step 2, region features are extracted via a conventional two-branch neural network. Based on the region features from step 2, the transformation relationship is calculated in step 3, and the final result is derived. Details of each step will be introduced in the following subsections.

1) Preprocessing: The main function of image preprocessing is to process different input images to be registered into

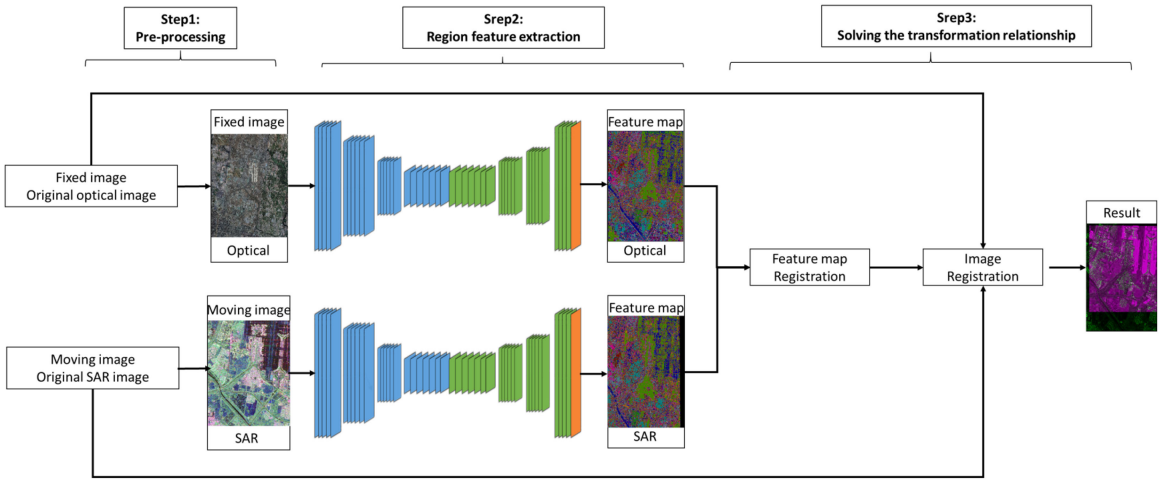


Fig. 4. Illustration of the algorithm architecture. Step 1 is input and preprocessing. Step 2 is region feature extraction. Step 3 is solving the transformation relationship and output result.

a data format that can be used by the network in step 2. Preprocessing methods of different source images are different. For optical imaging, the main processing is filtering to reduce image noise. A common method is Gaussian filtering. The input images and the output are 3-W. For the pol-SAR image, it includes filtering and decomposition. A common filter method is Lee-filtering. Pauli decomposition is a commonly used decomposition method. Its advantage is that it transmits the SAR image to 3-D data, which is similar to the RGB optical image. Assume that S is the scattering matrix of the pol-SAR image. SHH, SHV, SVH, and SVV are four channels of the pol-SAR image

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \quad (S_{HV} = S_{VH}). \tag{1}$$

The Pauli image is a pseudocolor image using intensities of three components, i.e.,

$$\mathrm{I}_{\mathrm{Pauli}} = \left[|S_{HV} - S_{HV}|^2, 4|S_{HV}|^2, |S_{HV} - S_{HV}|^2 \right]^T / 2. \tag{2}$$

2) Region Feature Extraction: Step 2 is a CNN network for region feature extraction, which is based on semantic segmentation. The backbone network consists of two branch networks corresponding to the fixed image and the moving image, respectively.

The structures of two branches are the same. They are composed of symmetric encoders and decoders and ended with a final pixel-level classification layer. The dimensions of the last classification layer's output must be the same because of the request of image registration. The encoder is a classification VGG network, which discards the final fully connected layer [40]. The encoder is composed of several basic units, each of which includes a convolution layer, a batch normalized layer, a rectified linear unit (ReLU), and a pooling layer. The basic unit of the decoder is quite similar to that of the encoder with some slight adaptations. It includes an upward sampling layer, a convolutional layer, a regularization layer, and a ReLU.

Considering the characteristics of RS images, e.g., high dimensions and large size, the design and selection of network must be subjected to this restrictions. Generally, larger input requires more memory and computing power, and deeper network produces better performance. Finally, in a balance of these two factors, the design of the network is shown in Fig. 5.

In order to alleviate the limitation of memory requirement and computation complexity caused by the size of the RS image, a method is specifically referred here [41]. The network adopts a special pooling method. The pool layer of the encoder stores the position of the maximum value in each pool window, which is used for feature mapping of each encoder. When a decoder performs upward sampling, it directly uses the index value stored in the encoder to place the data in the original position. Compared with U-Net, the feature map in the encoder phase does not need to be saved, which can reduce a lot of memory cost. In Fig. 5(a), the dotted line shows the pooling indices.

The parameter initialization of the model is random. For optical RS photos, we have tried to use the network parameters trained on the VOC2012 dataset. But experimental results show that these pretrained parameters present almost no benefits. The optimizer is an SGD, and the loss function uses cross entropy [42]. Suppose that the output result is x, and the true value is $class$. The loss function $\mathcal{L}$ is

$$\mathcal{L}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right). \tag{3}$$

3) Solving the Transformation Relationship: The process of image registration aims at solving the transformation relationship between the two images. The algorithm flowchart of solving the transformation relationship is shown in Fig. 6. Consider two images; the fixed image is F, the moving image is M, the fixed image feature map is F_f, and the moving image feature map is M_f. The transformation relationship is T_t, where t is the parameter in the transformation relationship. The similarity function is denoted as S, which is used to measure the similarity of the two images. The similarity function S is the objective

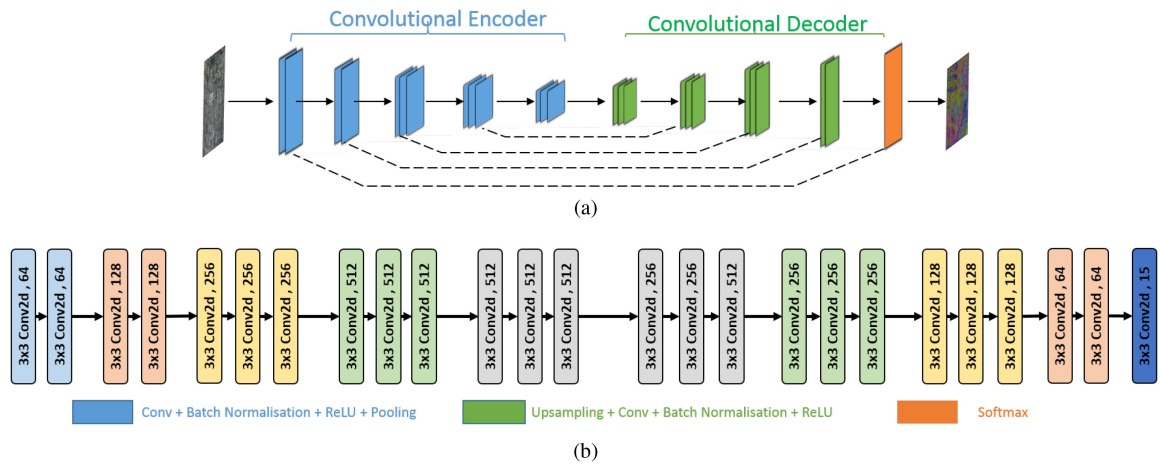


Fig. 5. (a) and (b) Architecture of the branch network.

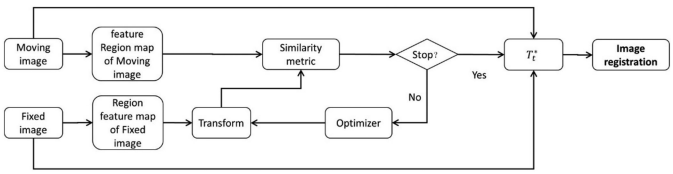


Fig. 6. Algorithm flowchart of solving the transformation relationship.

function. The goal of image registration is to find the transformation relationship T_t^*, which can make S achieve its maximum value. The final registered image is M_R

$$M_R = T_t^*(M) \tag{4}$$

$$T_t^* = \arg\max_{T_t} S\left(F_f, T_t\left(M_f\right)\right). \tag{5}$$

A common choice of objective function S for nonrigid registration is mutual information. The information-theoretic measure of mutual information measures the amount of information one random variable contains about another random variable. The mutual information of the two images is maximal when they are geometrically aligned. Let A and B be two images with conditional entropy $H(A)$ and H(B), and joint entropy $H(A, B)$; then, the expression of mutual information $I(A, B)$ can be written as

$$I\left(A, B\right) = H(A) + H(B) - H\left(A, B\right). \tag{6}$$

C. Effects of Region Feature Extraction on Image Registration Accuracy

In the proposed method, image registration is based on the regional features generated by image semantic segmentation. Image registration error comes from two aspects: one is the error in region feature map extraction and the other is the error in transform relation. In this section, we analyze the effects of regional feature extraction on image registration accuracy.

There are several ways to evaluate the similarity of the two feature images after registration: pixel accuracy (PA), mean pixel accuracy (MPA), mean intersection over union (MIoU), etc. Suppose i and j represent the pixel values of the two images, respectively. $K+1$ is the number of categories. $i, j \in [0, \mathrm{K}](i, j \in \mathbf{N}^+)$. p_{ii} represents the number of equal pixel values in the two images. p_{ij} and p_{ji} represent the number of pixels with unequal values, respectively:

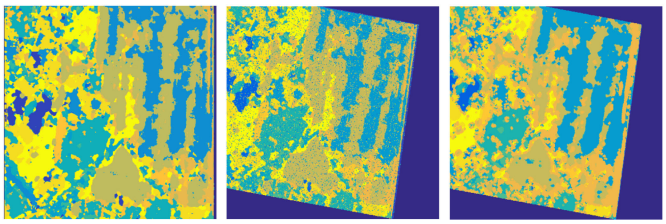
$$\mathrm{PA} = \left(\sum_{i=0}^{k} p_{ii}\right) \Big/ \left(\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}\right) \tag{7}$$

$$\mathrm{MPA} = \frac{1}{k+1}\left(\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}\right) \tag{8}$$

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{9}$$

In order to analyze the relationship between the error of the regional feature map and the final registration error. The algorithm flow of accuracy evaluation is shown in Algorithm 1. The regional feature image is I_f. Generate the registered image I_m by randomly transforming the image and I_N by adding random noise to I_m. Suppose that $K+1$ is the number of region categories. The pixel value of image I_f, I_m, I_N is integer in $[0, \mathrm{K}]$. Register I_f and I_N to derive the transformation relation T. Get registered image I_r ($I_r = T(I_m)$). Evaluate the registration accuracy of I_r and I_f by (7)–(9). In order to avoid the impact of randomness, we conduct this evaluation process for multiple times and calculate the average score. Two methods are used to simulate noise. One is to randomly superimpose the noise on the whole image, and the amount of increased noise is expressed by the number of pixels changed by the noise accounting for the number of pixels in the entire image. Other is to superimpose the noise on the edge of the region feature map. This is due to the consideration that noise at the edge has a greater impact on registration. The result of adding two types of noise separately is shown in Fig. 7.

The relationship between the accuracy of image registration and the noise intensity is shown in Fig. 8. Different colored



(a) (b) (c)

Fig. 7. (a) Original image. (b) Image superimposed the noise on the whole image. (c) Image superimposed the noise on the edge.

Algorithm 1: The Algorithm Flow of Accuracy Evaluation.

Input: A regional feature image I_f,
Initialization: Image Noise intensity N_i, the number of loops L
Begin
1 While $l < L$
2 Generate the registered image I_m by randomly transforming the image and I_N by adding random noise to I_m.
3 Register I_f and I_N to derive the transformation relation T
4 Get registered image I_r : $I_r = T(I_m)$
5 Evaluate the registration accuracy of I_r and I_f by 7, 8, 9
6 N_i++
7 $l++$
End
Output: The relationship between the accuracy of image registration and noise intensity

curves represent different evaluation methods. Although different indicators give specific accuracy rate values, the overall trend of change is consistent. The change trend can be roughly divided into three stages according to the weak to strong noise. In the first stage, the accuracy rate is relatively high and relatively stable. In the second stage, there is a cliff-like decline. In the third stage, the accuracy rate is low and relatively stable. Note that the total error is, however, not the sum of these two errors. Instead, the process of image registration is very robust to the error of region feature map. Under a certain threshold, the error of the region feature map does not necessarily affect the result of image registration. The following can be concluded through this simulation experiment.

1) The effect of region feature extraction accuracy on image registration accuracy approximates a step change function at a certain threshold. The increase in the feature extraction error within a certain range will not lead to the increase in the image registration error, and the method is robust overall. As shown in Fig. 8, if the type of noise is randomly distributed throughout the whole image, the proportion of noisy pixels in the whole image is preferably within 10%. If the noise type is concentrated on the edge of the region feature maps, the number of noisy pixels is preferably within 6% at the edge.
2) This method can tolerate a certain degree of feature extraction errors without reducing the accuracy of image registration. This indicates that the feature extraction process can achieve a better balance between accuracy and computation complexity.

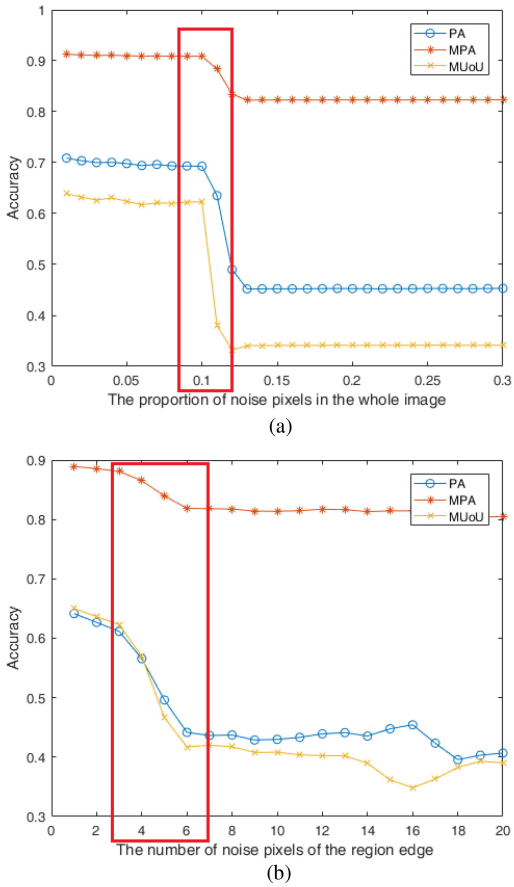
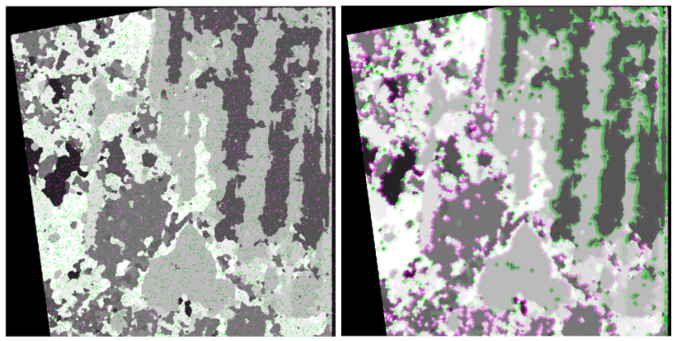


Fig. 8. Relationship between the accuracy of image registration and the noise. Two different methods are used to simulate noise. (a) Type of noise is randomly distributed throughout the picture. (b) Noise type is concentrated on the edge of the region feature maps.



(a) (b)

Fig. 9. Pseudocolor image superimposed noise in two different ways. (a) Noise is randomly distributed throughout the picture. (b) Noise is randomly imposed on the edge of the region feature maps.

In addition to quantitative evaluation, a pseudo-color image superimposed on the two pictures is used to visually compare the error of image registration, as shown in Fig. 9. In the figure, red and green pixels indicate inconsistent parts between the two

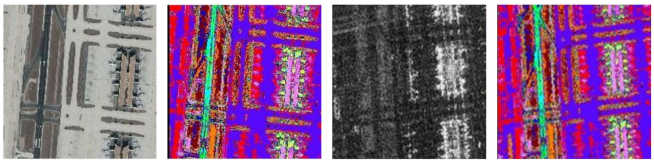


Fig. 10. (a) Optical imaging. (b) Region feature labels of optical imaging. (c) SAR image. (d) Region feature labels of the SAR image.

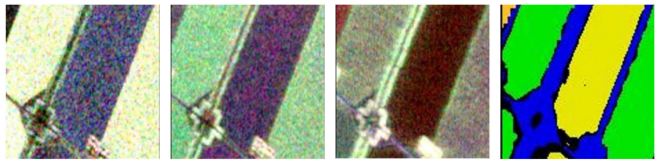


Fig. 11. Pauli image of Flevoland in three bands. (a) C-band. (b) L-band. (c) P-band. (d) Region feature labels of the image.

pictures. The left and right figures correspond to the two noise generation methods respectively.

III. Experiments and Discussions

This section is mainly about experimental analysis and discussion. Experiments were done on two datasets. In Section III-C, we compared the pros and cons of different feature extraction networks. In Section III-D, comparative experiments were done with four other methods: one SIFT method, and three CNN-based methods.

A. Datasets

The experiment was performed on two datasets. The first dataset is for SAR–optical image registration and the second one is for multiband SAR image registration. A brief introduction is as follows.

1) Dataset 1: It consists of GF-3 quad-polarized SAR images and corresponding optical image located in Beijing, China. The Gaofen-3 satellite is a Chinese satellite carrying a C-band SAR (5.4 GHz), launched in August 2016. The optical imaging is from Google Maps. The resolution is both about 8 m/pixel but not totally the same. The whole dataset is centered at E116.4 N40.0 in Beijing near the airport and contains 7469 groups of image chips. Each group consists of four image chips, which are optical imaging and its region feature labels and SAR image and its region labels. An example is shown in Fig. 10.

The image and corresponding labels are automatically generated according to the following methods.

1) Each GF3 dataset contains image files and parameter files. According to the parameters in the parameter file, the latitude and longitude of each pixel can be calculated.
2) The optical image is downloaded from Google Maps, and the latitude and longitude of each pixel can be obtained as well.
3) The labeling information can be extracted from purchased commercial navigation maps. These data contain markers for various objects, including roads, parks, rivers, lakes, etc., and the corresponding latitude and longitude.

According to the above three steps, three kinds of data and corresponding coordinate values can be obtained. Using the latitude and longitude coordinates of all pixel values as the link, the labels of image can be obtained. For example, the road feature type is recorded as a positive integer N. Get the longitude and latitude coordinates of points in the road area from the navigation data $[long, lat]$. According to $[long, lat]$, the corresponding coordinate point can be found in the optical image and marked as N. Thus, the corresponding label image of the optical image can be obtained, as shown in Fig. 10(a) and (b). In the same way, the label image corresponding to the SAR image can be obtained, as shown in Fig. 10(c) and (d). For an ideal transform function, the registered label images can have a pixel-level one-to-one correspondence.

The advantage of the method is that the dataset can be automatically generated. Similarly, more multisource RS datasets can be produced, so that the method proposed has good generalization.

2) Dataset 2: It consists of ARISAR SAR data in three bands (P, L, and C) in Flevoland (The Netherlands). The dataset is located in Flevoland, The Netherlands, centered at 52.4° North latitude, 5.4° East longitude. The area shown is approximately 25 by 28 km. Flevoland, which fills the lower two-thirds of the image, is a very flat area made up of reclaimed land that is used for agriculture and forestry. The resolution is 6.6 m in the slant range direction and 12.1 m in the azimuth direction. The source of the data can be referred to [43]. An example is shown in Fig. 11.

In the experiment, five-cross-validation was used, and the data were randomly divided into five parts, of which four parts were used for training and one part was used for verification.

B. Experimental Environment

The program is executed on a workstation with Intel 64 Core i7 CPUs with 128-GB RAM and a TITAN-V GPU with 12 GB of memory capacity. The operating system is Ubuntu 16.04. The code is programmed in MATLAB 2017b and Python 3.7.

C. Comparisons of Network Architecture

The CNN is undoubtedly a very good choice for extracting regional features. The feature extraction process can also be almost equivalent to an image semantic segmentation process. In the field of semantic segmentation, there are many effective methods. These methods can be used as reference methods for feature extraction. But, obviously, it is necessary to select and design a suitable network structure according to the requirements of RS image registration. According to the previous analysis, after the accuracy rate exceeds a certain threshold, the improvement of the accuracy of feature extraction has no effect on the improvement of the registration accuracy. Therefore, the accuracy of the results is not the highest priority factor in the network design.

However, there are some factors that have to be considered for RS images. The two images to be registered are generally

TABLE I
COMPARISONS OF NETWORK ARCHITECTURE

Method	Params	(512,512) Memory	(512,512) Flops	(1024,1024) Memory	(1024,1024) Flops
Ours	20,743,170	1.66G	113.34G	6.42GB	453.34G
FCN8	134,367,618	12.87G	8.45T	18.71G	34.53T
FCN16	134,361,964	4.09G	2.04T	15.21G	8.52T
PSP-Net	56,147,926	177.56G	11.19T	-	-
Res-Unet	14,482,564	5.99G	283.34G	23.79G	1.13T
R2U-Net	39,091,393	4.4G	262.76G	17.18G	1.05T
Attention U-net	34,878,573	4.68G	266.45G	18.32G	1.07T
R2AttU-Net	39,442,925	5.17G	267.19G	20.24G	1.07T

different in size and resolution. Therefore, the network should be adapted to input images with different sizes and different dimensions. The input RS image is relatively large. Larger input data and more complex networks will lead to more consumption of computing resources. The consumption of resources is an important limitation that cannot be ignored in RS image processing. This requires network design to keep memory requirements and computational complexity within an acceptable range. Therefore, we compared our method with some classic network structures and compared them in three aspects: total number of network parameters, floating point arithmetic (FLOPs), and memory usage. The toolkit is torchstat [44]. The size of the input image used is 512×512 and 1024×1024. As shown in Table I, it can be seen from the comparisons that our network structure requires less computational resource consumption and can be deployed on GPUs with higher cost performance, such as the NVIDIA TITAN-V GPU used in this article.

D. Results and Evaluations

Three methods are used to evaluate the result of the image registration and are represented in three subsections.

1) *Mosaic display:* To show the results of registration more intuitively, the method of mosaic display is adopted. More registration details can be observed by changing the grid size. The focus of this method is to see the continuity of the edges at the junctions of the image grids. If the registration result is good, the edges are continuous; otherwise, the edges will appear dislocated.
2) *Region-based method:* It includes PA, MPA, and MIoU. These criteria have been introduced in (7)–(9).
3) *Point-based method:* The average probability of correct key point (PCK) refers to the proportion of key points that are correctly matched. A key point is considered to be matched correctly if its predicted location is within a distance of its actual location [20], [45]–[49].

1) Result of Mosaic Display: The result of mosaic display is shown in Figs. 12–14. This display method judges the accuracy of registration by observing the continuity of the edges. For example, the part marked by the red rectangle in Fig. 12 is an edge of runway. The part marked by the red rectangle in Fig. 13 is a part of river. The part marked by the red rectangle in Fig. 14 is a part of road. Compared with other features, roads are most easily affected by registration. If there is a deviation in registration, the roads will not be aligned. In order to observe the registration results more closely, the registration results are presented in two different grid sizes. Different details can be seen in different grid sizes. As can be seen from the following results, the registration result is good.

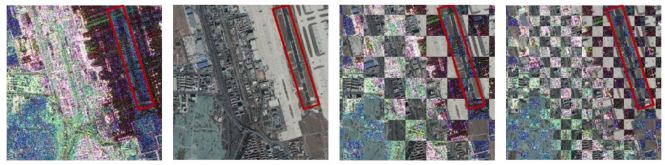


Fig. 12. Results of the registration for SAR image and optic photograph. The part in red rectangle is the edge of runway. (a) Pauli pseudocolor image of GF3 SAR image. (b) Optical imaging. (c) Big grid mosaic display. (d) Small grid mosaic display.

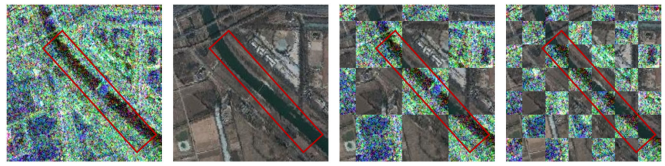


Fig. 13. Results of the registration for SAR image and optic photograph. The part in red rectangle is the river. (a) Pauli pseudocolor image of GF3 SAR image. (b) Optical imaging. (c) Big grid mosaic display. (d) Small grid mosaic display.

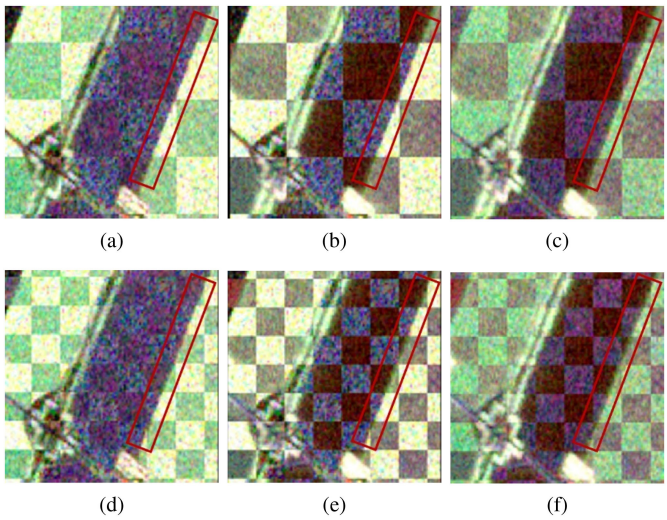
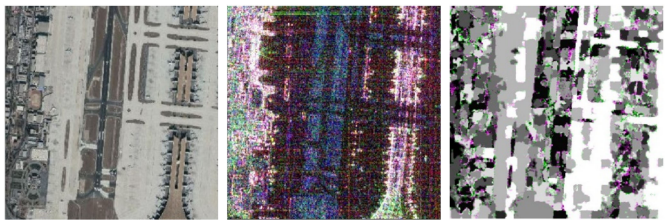


Fig. 14. Registration results are shown in mosaic. The part in red rectangle is the road. (a) C-band and L-band in big mosaic display. (b) C-band and P-band in big mosaic display. (c) L-band and P-band in big mosaic display. (d) C-band and L-band in small mosaic display. (e) C-band and P-band in small mosaic display. (f) L-band and P-band in small mosaic display.

The Flevoland dataset contains three bands, i.e., L, P, and C. For visualization purpose, the image is first preprocessed with Pauli decomposition. The images of three bands are registered in a pairwise fashion, and three groups of results are obtained after image registration, as shown in Fig. 14: the first is C and L, the second is C and P, and the last one is L and P.

TABLE II
RESULT OF REGION-BASED EVALUATION

		PA	MPA	MIoU
Dataset 1		91.43%	92.51%	93.05%
Dataset 2	Band C and Band L	91.43%	92.51%	93.05%
	Band C and Band P	80.43%	96.48%	94.78%
	Band L and Band P	82.46%	96.43%	93.73%



(a) (b) (c)

Fig. 15. Results of the region feature map. (a) Optical imaging. (b) Pauli pseudocolor image of the GF3 SAR image. (c) Pseudocolor image showing the difference of the two region feature maps. Gray regions in the composite image are the areas where the two images have the same intensities. Magenta and green regions are the areas where the intensities are different.

TABLE III
RESULT OF POINT-BASED EVALUATION

	Dataset 1	Dataset 2		
		Band C-L	Band C-P	Band L-P
Our method	92.49%	98.89%	98.40%	98.26%
SIFT method	-	98.49%	98.20%	98.86%
Dense-SNCNet	82.70%	97.81%	98.54%	98.51%
CNNReg	0.19%	97.90%	97.57%	98.76%
VoxelMorph	36.22%	98.01%	97.40%	98.51%

2) Region-Based Evaluation: The region-based method mainly evaluates the registration performance by judging the accuracy of the region map after registration. The calculation method is based on (7)–(9). The results are shown in Table II. According to previous analysis, the accuracy of the region and the accuracy of the final registration are not linearly related. Therefore, after the accuracy reaches a certain threshold, the registration can be successful. According to the results shown in Fig. 8, the threshold of PA is 70%, that of the MPA is 90%, and that of the MIoU is 65%.

The experimental results are further analyzed by comparing the extraction results of the region feature map. The results of the region feature map are shown in Fig. 15, and the differences between the two region feature maps are shown in Fig. 15(c). Gray regions in the composite image show the areas where the two images have the same intensities. Magenta and green regions show the areas where the intensities are different.

3) Point-Based Evaluation: Currently, in the research of image registration, the most commonly used registration method is PCK, which refers to the proportion of key points that are correctly matched. The PCK results of dataset 1 and dataset 2 are shown in Table III. There are four methods for comparison.

1) *SIFT:* This a very popular method for image registration.
2) *Dense-SNCNet [50]:* This method proposes a framework for end-to-end polarimetric SAR image registration that is based on weakly supervised learning and uses no image patch processing or iterative parameter estimation.
3) *CNNReg [51]:* This article presents a CNN feature-based multitemporal RS image registration method with two key contributions: 1) it uses a CNN to generate robust multiscale feature descriptors and 2) it designs a gradually increasing selection of inliers to improve the robustness of feature point registration.
4) *VoxelMorph [52]:* This is a fast learning-based framework for deformable pairwise image registration. This method formulates registration as a function that maps an input image pair to a deformation field that aligns these images. It parameterizes the function via a CNN and optimizes the parameters of the neural network on a set of images.

It can be seen from the results that our method has obtained good results on both datasets. Particularly, in dataset 1, compared with other methods, the advantages are more obvious. For the SIFT method, the accuracy of the SIFT method completely depends on the accuracy of the selection of feature points. In a multisource image, it is more difficult to extract many feature points. If the proper feature points cannot be obtained, the registration will almost completely fail, so the PCK value is almost zero. The Dense-SNCNet is to be more susceptible to noise interference during feature extraction. In the band C-P and band L-P experiments of dataset 2, the quality of the data itself is good, so that the registration result is very good. However, in other high-noise data, experimental results are poor. In comparison, the proposed method has stronger anti-interference ability and better generalization. CNNReg is not balanced in feature extraction, and the effect is better in areas with strong features, but in areas with weak features, it is easy to be interfered by strong features. VoxelMorph's idea is very novel and can make finer registration, but in RS images, this will cause distortion of the image registration, and the effect is poor.

IV. CONCLUSION

In this article, we proposed a novel automatic registration method for high-resolution multisource and multiband RS images via a CNN. The main idea of the proposed method is the usage of region features in images, which is different from previous methods. The architecture of the network is composed of two branch networks. Each branch independently generates the region feature map for one of the input images, which represents the common semantic characters of the input multisource data. The result is a pixel-level one-to-one correspondence output. The proposed method outperforms previous intensity-based or edge-based methods by overcoming the limits of the instability of features of multisource data. The method is tested on datasets from two different scenarios. The results illustrate that the proposed method has good performance in RS data registration of multisource images.

ACKNOWLEDGMENT

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

REFERENCES

[1] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[2] D. Tuia, D. Marcos, and G. Camps-Valls, "Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization," *ISPRS J. Photogrammetry Remote Sens.*, vol. 120, pp. 1–12, 2016.

[3] B. Dousset and F. Gourmelon, "Satellite multi-sensor data analysis of urban surface temperatures and landcover," *ISPRS J. Photogrammetry Remote Sens.*, vol. 58, nos. 1/2, pp. 43–54, 2003.

[4] P. Hyde, R. Dubayah, W. Walker, J. B. Blair, M. Hofton, and C. Hunsaker, "Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM, Quickbird) synergy," *Remote Sens. Environ.*, vol. 102, nos. 1/2, pp. 63–73, 2006.

[5] J. Rhee, J. Im, and G. J. Carbone, "Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data," *Remote Sens. Environ.*, vol. 114, no. 12, pp. 2875–2887, 2010.

[6] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[7] D. M. Tralli, R. G. Blom, V. Zlotnicki, A. Donnellan, and D. L. Evans, "Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards," *ISPRS J. Photogrammetry Remote Sens.*, vol. 59, no. 4, pp. 185–198, 2005.

[8] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[9] K. Rokni, A. Ahmad, K. Solaimani, and S. Hazini, "A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 226–234, 2015.

[10] L. Wan, Y. Xiang, and H. You, "A post-classification comparison method for SAR and optical images change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1026–1030, Dec. 2019.

[11] Y. Zhang, "Understanding image fusion," *Photogrammetric Eng. Remote Sens.*, vol. 70, no. 6, pp. 657–661, 2004.

[12] Y. Xiang, F. Wang, L. Wan, N. Jiao, and H. You, "OS-Flow: A robust algorithm for dense optical and SAR image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6335–6354, Sep. 2019.

[13] D. Marcos, R. Hamid, and D. Tuia, "Geospatial correspondences for multimodal registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5091–5100.

[14] Accessed: Jan. 13, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Image_registration

[15] 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.

[16] M. Hasan, M. R. Pickering, and X. Jia, "Multi-modal registration of SAR and optical satellite images," in *Proc. Digit. Image Comput., Techn. Appl*, 2009, pp. 447–453.

[17] C. G. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 10–5244.

[18] C. Pan, Z. Zhang, H. Yan, G. Wu, and S. Ma, "Multisource data registration based on NURBS description of contours," *Int. J. Remote Sens.*, vol. 29, no. 2, pp. 569–591, 2008.

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

[20] B. Fan, C. Huo, C. Pan, and Q. Kong, "Registration of optical and SAR satellite images by exploring the spatial relationship of the improved sift," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 657–661, Jul. 2013.

[21] K. J. Batenburg and J. Sijbers, "Optimal threshold selection for tomogram segmentation by projection distance minimization," *IEEE Trans. Med. Imag.*, vol. 28, no. 5, pp. 676–686, May 2009.

[22] K. J. Batenburg and J. Sijbers, "Adaptive thresholding of tomograms by projection distance minimization," *Pattern Recognit.*, vol. 42, no. 10, pp. 2297–2305, 2009.

[23] A. Kashanipour, N. S. Milani, A. R. Kashanipour, and H. H. Eghrary, "Robust color classification using fuzzy rule-based particle swarm optimization," in *Proc. Congr. Image Signal Process.*, 2008, pp. 110–114.

[24] L. Chen, H. Cheng, and J. Zhang, "Fuzzy subfiber and its application to seismic lithology classification," *Inf. Sci.—Appl.*, vol. 1, no. 2, pp. 77–95, 1994.

[25] L. Chen, "The lambda-connected segmentation and the optimal algorithm for split-and-merge segmentation," *Chin. J. Comput.*, vol. 14, no. 5, pp. 321–331, 1991.

[26] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1452–1458, Nov. 2004.

[27] T. Lindeberg and M.-X. Li, "Segmentation and classification of edges using minimum description length approximation and complementary junction cues," *Comput. Vis. Image Understanding*, vol. 67, no. 1, pp. 88–98, 1997.

[28] L. Barghout, "Visual taxometric approach to image segmentation using fuzzy-spatial taxon cut yields contextually relevant regions," in *Proc. Int. Conf. Inf. Process. Manage. Uncertainty Knowl. Based Syst.*, 2014, pp. 163–173.

[29] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.

[30] S. Osher and N. Paragios, *Geometric Level Set Methods in Imaging, Vision, and Graphics*. New York, NY, USA: Springer, 2003.

[31] A. Yezzi, S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum, "A geometric snake model for segmentation of medical imagery," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 199–209, Apr. 1997.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, 2015, pp. 234–241.

[34] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ.*, 2018, pp. 327–331.

[35] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," *J. Med. Imag. (Bellingham, Wash.)*, vol. 6, no. 1, 2019, Art. no. 014006.

[36] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," *Med. Imag. Deep Learn.*, pp. 1–10, 2018.

[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[38] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[39] L. Kai and Z. Xueqing, "Review of research on registration of SAR and optical remote sensing image based on feature," in *Proc. IEEE 3rd Int. Conf. Signal Image Process.*, 2018, pp. 111–115.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Biol. Learn. Soc.*, pp. 1–14, 2015.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[42] Accessed: Jan. 13, 2021. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

[43] W. Gao, J. Yang, and W. Ma, "Land cover classification for polarimetric SAR images based on mixture models," *Remote Sens.*, vol. 6, no. 5, pp. 3770–3790, 2014.

[44] Accessed: Jan. 13, 2021. [Online]. Available: https://github.com/Swall0w/torchstat

[45] 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.

[46] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A sift-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2014.

[47] Y. Xiang, R. Tao, L. Wan, F. Wang, and H. You, "OS-PC: Combining feature representation and 3-D phase correlation for subpixel optical and SAR image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6451–6466, Sep. 2020.

[48] M.-S. Kang and K.-T. Kim, "Automatic SAR image registration via Tsallis entropy and iterative search process," *IEEE Sens. J.*, vol. 20, no. 14, pp. 7711–7720, Jul. 2020.

[49] S. Paul and U. C. Pati, "Automatic optical-to-SAR image registration using a structural descriptor," *IET Image Process.*, vol. 14, no. 1, pp. 62–73, 2019.
[50] Q. Zhu, J. Yin, L. Zeng, and J. Yang, "Polarimetric SAR image affine registration based on neighborhood consensus," *J. Radars*, vol. 9, pp. 1–12, 2020.
[51] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38 544–38555, 2018.
[52] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

Liang Zeng is currently working toward the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering, Tsinghua University, Beijing, China.

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

Yanlei Du (Member, IEEE) received Ph.D. degree in cartography and geographic information systems from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2019.

He is currently a Postdoctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, Beijing. His research interests include computational electromagnetics in applications of ocean remote sensing and satellite oceanography.

Huiping Lin received the B.S. degree in information and communication engineering, in 2015 from Tsinghua University, Beijing, China, where he is currently working toward the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering.

His research interests include polarimetric synthetic aperture radar image segmentation, object detection, and target classification.

Jing Wang received the Ph.D. degree in radar signal processing from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2010.

She is currently a Senior Engineer with the Science and Technology on Information System Engineering Laboratory, Electronics Technology Group Corporation, Nanjing. Her research interest includes artificial intelligence in applications of radar signal processing.

Junjun Yin received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2013.

From 2013 to 2015, she was a Postdoctoral Research Fellow with the Department of Electronic Engineering, Tsinghua University, and with the Department of Geological Sciences, University of Manitoba, Winnipeg, MB, Canada. She is currently a Faculty Member with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing. Her current research interests include polarimetric synthetic aperture radar (SAR) image interpretation, compact polarimetry, ship detection and oil spill observation, change detection, and polarimetric SAR image classification and segmentation.

Jian Yang (Senior Member, IEEE) received the Ph.D. degree in polarimetric radar remote sensing from Niigata University, Niigata, Japan, in 1999.

In 1985, he joined the Department of Applied Mathematics, Northwestern Polytechnical University. From 1999 to 2000, he was an Assistant Professor with Niigata University. In April 2000, he joined the Department of Electronic Engineering, Tsinghua University, Beijing, China, and became a Full Professor in 2002. He has authored or coauthored more than 300 papers and received many awards. His research interests include radar polarimetry feature extraction, target detection, and target classification.