SAR–Optical Image Matching With Semantic Position Probability Distribution

Liangzhi Li, Ling Han, Ming Liu, Kyle Gao, *Graduate Student Member, IEEE*, Hongjie He, Lanying Wang, and Jonathan Li, *Fellow, IEEE*

***Abstract*— We propose a deep learning framework of semantic position probability distribution for synthetic aperture radar (SAR)-optical image matching, termed as SPPD. Unlike the pixel-by-pixel searching matching method, a correspondence is directly obtained by an outputted matching position probability distribution. First, multiscale pyramidal features are created for each pixel in the SAR and optical images by using two weight-sharing ResNet-50 + feature pyramid network (FPN) networks. The features containing high-level semantic information are then embedded into the proposed image position attention module to obtain the spatial position dependencies between two images. Then, we present a loss function for semantic position matching to optimize the network from both semantic information and pixel alignment perspectives, converting the probability distribution of semantic matching positions into a point-to-point matching problem. In this article, the SAR and optical images are set as the sensed and reference images. The effects of different image sizes, training label types, and loss function weights on matching accuracy are explored to obtain the optimal parameter settings for matching. The experimental results show that the proposed method is insensitive to image deformation and achieves cross-modal matching for SAR–optical images with high accuracy compared with the best matching method on different scene images, with several orders of magnitude faster inferences time.**

***Index Terms*— Image registration, position attention mechanism, pyramidal feature, semantic matching, synthetic aperture radar (SAR)-optical image.**

Manuscript received 9 June 2023; revised 18 September 2023 and 31 October 2023; accepted 4 November 2023. Date of publication 7 November 2023; date of current version 20 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42171348, in part by the Science and Technology Department of Shaanxi Province under Grant 211435220242, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2022JQ-247, in part by the China Center for Remote Sensing of Natural Resources Aerial Mapping under Grant 211735210034, and in part by the Shaanxi Key Laboratory of Land Consolidation through the Fund Project under Grant 2019-ZD04. *(Corresponding authors: Jonathan Li; Ling Han.)*

Liangzhi Li is with the College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710064, China, and also with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Ling Han and Ming Liu are with the School of Land Engineering, Chang'an University, Xi'an 710064, China (e-mail: hanling@chd.edu.cn).

Kyle Gao is with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Hongjie He, Lanying Wang, and Jonathan Li are with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Digital Object Identifier 10.1109/TGRS.2023.3330856

I. Introduction

SINCE the rapid development of sensor technology, various kinds of imagery can be accessed. While, all kinds of imagery have advantages and disadvantages, optical imagery can usually provide rich information in texture, color, and spectrum, but easily affected by cloud coverage, light, and other atmospheric conditions. As the active remote sensing technology, synthetic aperture radar (SAR) sensors are less susceptible to weather and light conditions [1], [2]. However, their imaging is usually in lower spatial resolution and affected by speckle noise, resulting in poorer interpretation quality. Therefore, fused these two modalities of imagery together will provide complementary information to each other. Image matching is the key technique to integrate these images to form a combined representation of observed scenes [3]. However, due to the heterogeneous representation of multimodal remote sensing images, matching SAR and optical images remains huge difficulties in term of lacking both heterogeneous pixel intensity and spatial feature information representation for solving the problem of feature invariance due to temporal variations [4].

Currently, multimodal SAR–optical image matching can be categorized into feature-based and area-based matching methods [5], [6]. Feature-based methods firstly extract salient features of the image, including points (Moravec and Harris detector [7]), lines (edges and contours) and surface features, and then measure their feature descriptions to obtain correspondence [8]. Many feature-based methods have been developed to detect and describe SAR images based on scale invariant feature transformation algorithms. Li et al. [9] combine the use of phase coherence to create the radiation change insensitive feature transform (RIFT), which is shown to be less sensitive to rotations and radiation differences across modes. Most feature-based approaches can identify correspondences between SAR and optical modes, and they are only applicable to images that conform to specific radiometric constraints with no geometric distortions.

Compared to feature-based methods, area-based methods have the following advantages: 1) area-based methods avoid feature detection and search for similarities with maximum features [10] and 2) area-based methods allow searching for initial geographic locations in remote sensing images within small regions to obtain geo-corrected images with position offsets of only a few pixels. Common similarity metrics include sum of squared differences (SSD) [11], the normalized

cross correlation (NCC) [8], [12], and mutual information (MI) [13]. SSD and NCC calculate the similarity and correlation of two images separately, which is vulnerable to nonlinear radiometric differences in SAR optical images. Therefore, neither method can effectively handle images with nonlinear radiometric differences. MI is robust to multimodal images and is widely used for remote sensing images and medical image registration [14], [15]. However, MI is sensitive to template distortion, while the pixel-by-pixel search for matching positions makes the computational overhead very large, which limits its application in remote sensing image matching. However, these area-based methods are weak in dealing with large geometric deformations, which is hard to enforce in a wide range of remote sensing image scenarios. Recently, Ye et al. [1] and [16] proposed a fast and robust matching framework based on structure similarity, which significantly outperform the intensity-based similarity metrics such as NCC and MI. Such image framework makes a great breakthrough to detect correspondences between multimodal remote sensing images with significant radiometric differences.

Recently, data-driven deep learning matching methods perform advanced image abstraction to obtain keypoints and feature descriptions, such as learning invariant feature transformation (LIFT) [17], SuperPoint [18], deep local feature (DELF) [19] methods, detection and description network (D2-Net) [20], and Superglue [21]. These feature-based methods are difficult to apply to SAR optical images with nonlinear radiometric differences where they struggle to achieve the distinguishability of keypoints and feature descriptions [22], [23]. To alleviate this limitation, many studies have developed SAR–optical matching models with repeatable keypoints. For example, MAP-Net [22] embeds SAR–optical image information containing high-level semantic features into cross-modal matching using self-attention, obtaining key features that are distinguishable. Xiang et al. [24] proposed a stable feature crossover-based keypoint detector as well as a cross-stage partial twin network to quickly extract feature descriptors containing deep and shallow features for SAR–optical image matching.

Furthermore, to address the problem of keypoint nonreproducibility, some studies have used deep neural networks to generate candidate matching regions based on the local features of patches, such as Goodness network [25], twin U-Net with a fast Fourier transform (FFT) [26], HardNet [27], pseudo-Siamese convolutional neural network (CNN) [28], and Siamese network followed by a similarity measure layer [29], [30]. These methods provide similarity matching regions and patches-based feature descriptions for SAR–optical image matching using deep neural network modeling. However, these methods are time-consuming when estimating the similarity between patches because of their sliding search strategy. Furthermore, these methods do not determine the spatial dependence of the pixels between the reference and sensed image. The network framework needs to be retrained to fit the matching scene when matching various regions, which limits the applicability of these methods.

To solve the problem of time-consuming similarity matches, Li et al. [31] proposed a semantic template matching framework that maps the template and reference image to the output prime position match as feature fusion to obtain the correspondence between images without considering the similarity between pixels. However, this mapping relationship is a fit on a dataset, which requires retraining of the network to accommodate the complex diversity of scenes in practical applications.

In summary, using deep neural matching framework conducts a SAR–optical imagery matching which mainly has two issues.

1) The network model needs to be retrained due to scene changes, which requires establishing the positional dependency between two image pixels.
2) Obtaining matching similarities via a pixel-by-pixel search strategy is time-consuming, necessitating localization.

Therefore, to response the above problems, first, we proposed a novel deep neural matching network that build multiscale pyramidal feature for each pixel in SAR and optical by using two weight-sharing ResNet-50 [32] + feature pyramid network (FPN) [33] which resist local distortions. We present a position attention module to obtain matching position dependencies between features. Second, to reduce computational complexity, we map the dependencies between image matching positions and semantic position probability distributions. Finally, a loss function based on a weighted average of the output position probabilities is proposed to solve the matching location problem by optimizing the network from both semantic and matching position perspectives. The main contributions are summarized as follows.

1) *Pyramid geometric invariant feature extraction network:* This structure obtains multiscale feature information by constructing a pyramidal feature for each local pixel to deal with the geometric differences between SAR and optical images.
2) *SAR–optical image semantic position attention module:* The network captures the dependency between two positions of the SAR–optical image. The position attention module aggregates the matching information from the reference and sensed images, which is crucial to obtain semantic position probability distributions of the network.
3) *A new loss function based on output position probability distribution:* This loss function performs a weighted average for the output probability distribution, which converts the position probability distribution boundary alignment into a point-to-point matching problem, lowering the computational complexity.

The rest of the article is organized as follows. Section II describes the pyramidal feature network structure and attention mechanisms. Section III presents the proposed method. Section IV describes the training dataset and the pipeline process for image registration. Section V details the effectiveness of the proposed network. Section VI concludes the article.

II. Related Works

This section describes the background of position attention mechanism.

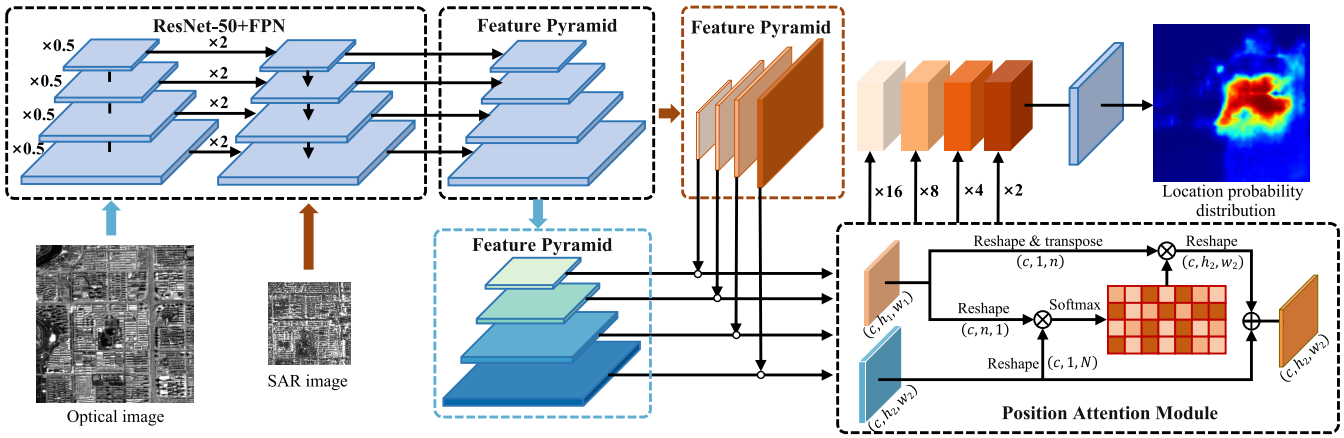


Fig. 1. Framework of the semantic position matching for SAR–optical images.

A. Position Attention Mechanism

The attention model (AM) [34] has become an important concept in neural networks. AM, a weight allocation mechanism, redistributes the otherwise equally distributed resources based on the importance of the attention object. The fundamental concept is to use the original data to identify the correlation between them and then highlight some of their most important characteristics. Channel attention [35], self-attention [36], and position attention [37] are some of the different types of AM.

The position attention mechanism is used to capture the dependencies between any two locations of the feature map. Any particular feature is weighted by its similarity to other features. Thus, any two existing positions with similar features can contribute to each other's lifting, regardless of the distance between them.

Since the position attention module is proposed for all the features at all positions of the same feature map, it will fuse all the features at the obtained positions. However, our goal is to establish the position dependence of the sensed image on the reference image. This operation involves two feature maps with different dimensions. Therefore, we have redesigned a position attention module. The detailed design of the position attention module in our framework will be explained in Section III-C.

III. Method

A. Overall Framework Description

Fig. 1 shows the SPPD architecture which maps the probability distribution of semantic matching positions between the SAR images on optical images. We first input the SAR and optical images into the weight-sharing ResNet-50 + FPN to generate the pyramid structure features. The weight sharing is used to obtain cross-modal information for SAR and optical images. ResNet-50 + FPN network structure is mainly used to enable the network to model the complex geometric distortion between SAR and optical images. The pyramidal features of the SAR and optical images are then input to the position attention module, and all the features obtained at the positions are fused to generate a feature map with position dependencies between the SAR and optical images. Finally, the upsampled semantic position feature maps are coupled and fed into CNN to output matching position probability distribution results.

The matching position probability distribution characterizes the probability that each pixel in the SAR image is located in the reference image. However, the outputted semantic location probability distribution is unordered, as shown in Fig. 1. It fails to determine the probability of each pixel for certain positions. Therefore, we propose a weighted average loss function for output matching position probability distribution, which involves computing the centroid of the output, changing pixel matching to point matching, and greatly increasing matching efficiency. The framework of ResNet-50 + FPN, the proposed position attention mechanism, and the loss function are described in detail below.

B. ResNet-50 + FPN Network

The first stage of our framework aims to perform cross-modal feature extraction for SAR and optical images by a weight sharing strategy to obtain multiscale high-level semantic information. This is because FPN utilizes both low-level features and high-level features, and the fused feature is output separately. The combination of features from the four layers of the FPN output forms a scale invariant feature transform (SIFT)-like [38] feature description. Therefore, ResNet-50 and FPN structures are used to build features at different levels (p_1, p_2, p_3, p_4) from each residual block.

The ResNet-50 + FPN network is illustrated in Fig. 2. The structure can be divided into three parts.

1) *A bottom-up residual network on the left:* The input first passes through a CNN with stride = 2, to reduce the feature map size and improve the computational efficiency in the position attention module. The feature map is half of the original size for each residual block, which constitutes a feature pyramid. Specifically, the features of the residual structure of each stage are used for the output. These bottom-up residual blocks are denoted as Res-1, Res-2, Res-3, and Res-4, where the output corresponds to the input of the left-hand structure.

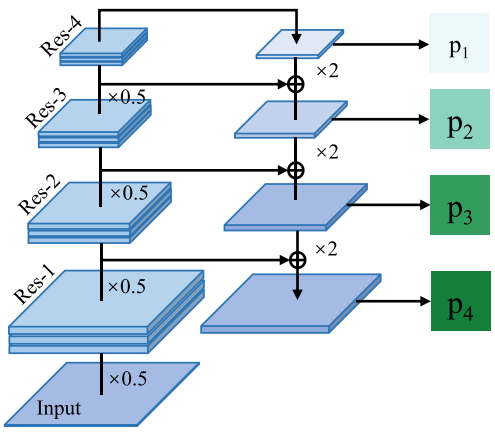


Fig. 2. Overall structure of ResNet-50 + FPN.

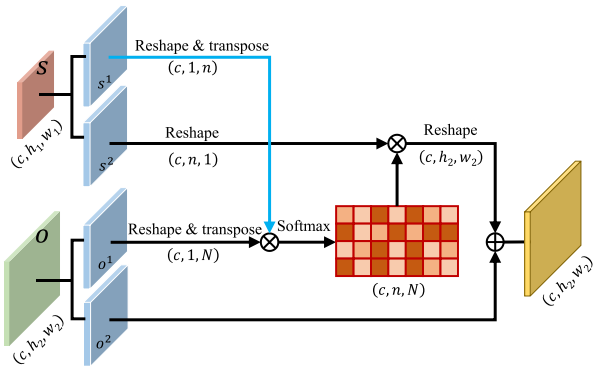


Fig. 3. Diagram of the of position attention module, where s and o denote the semantic feature maps of SAR–optical images with sizes (c, h_1, w_1), (c, h_2, w_2).

2) *A top-down upsampling on the right:* The top-down process is performed using the interpolation upsampling, which expands the feature map to twice the original size. This allows the upsampled feature map to extend to the dimensions same as the feature map of the next layer.
3) *A connection structure between the middle features:* The connection structure is to fuse the result of upsampling with the feature maps generated from the bottom-up. The feature maps output from the residual module are subjected to a 1×1 convolution and fused with the upsampled feature maps to obtain richer features of different layers. The 1×1 convolution aims to change the number of channels, which is required to be the same as the number of channels in the next layer. Then, a 3×3 convolution is used to eliminate the confounding effect of the upsampling and obtain multiple new feature maps (p_1, p_2, p_3, p_4).

C. Position Attention Module

The position attention module is mainly used to capture the dependencies between any two positions of the SAR–optical image. The position attention module is first proposed by the semantic task to establish the element information in the same channel. However, we need to establish the position dependency of the sensed image on the reference image, which is an operation on two feature maps. Therefore, we improve a position attention module for semantic matching.

Fig. 3 depicts the proposed position attention module for semantic matching. Let s and o be the feature maps of SAR and optical images with sizes (c, h_1, w_1), (c, h_2, w_2). From s and o, convolution layers are applied to obtain s_1, s_2, and o_1, o_2, respectively. For s_1, s_2, o_1, they are first reconstructed to obtain feature maps of sizes (c, 1, n), (c, 1, n), and (c, 1, N), where $n = h_1 \times w_1$ and $N = h_2 \times w_2$. Then, the transpose of s_1 is multiplied with o_1 and the SoftMax operation is applied to obtain the spatial attention map $A(N, n)$. Finally, A is multiplied by s_2 and reconstructed as (c, h_2, w_2), where the result is multiplied by a scale factor and then added to the feature map o_2 to obtain the final output feature map. The computational procedure can be described as follows:

$$A(N,n)_{i,j} = \text{SoftMax}\left(\frac{\exp\left(s_1^i \cdot o_1^j\right)}{\sum_{i=1}^{N} \exp\left(s_1^i \cdot o_1^j\right)}\right) \tag{1}$$

where $A(N, n)$ measures the influence of position i on position j. The more similar the feature representations of two locations are, the more they contribute to the correlation between them

$$E_j = \alpha \sum_{i=1}^{N} \left(A(N,n)_{i,j}\ s_2^i\right) + o_2^j \tag{2}$$

where α is a learned weight and initialized to 0. The result E for each position is a weighted sum of the features and optical image features of all positions of the SAR. Therefore, E has a global semantic position relationship. Similar semantic features realize the response for each other's positions, thus enhancing the matching of similar features and semantic consistency.

D. Loss Function

Since the output position probability distributions are unordered, they cannot be used to determine the coordinates and matching probabilities for the pixels. Furthermore, the final matching result cannot be based on the correspondence of a particular one-pixel probability value in the SAR image. Therefore, we compute the centroid of the output, which is the weighted average of the feature map.

Fig. 4 shows the geometric center and centroid of the affine transformed template. In Fig. 4(a), S is affine transformed to generate S', where the centroid of S and S' denote the same position pixel values. Intuitively, the centroid and center pixel of S, S' coincide after affine transformation. This means that the centroid coordinates of the output feature map are the matching positions of the template center pixels, which is used to replace all pixels matching with centroid matching.

A loss function is proposed to calculate the centroid position loss while considering the semantic loss between label and output, as shown in Fig. 4(b). The centroid position loss guides the output to correspond at the center point, and the semantic loss function optimizes the matching position of each pixel in

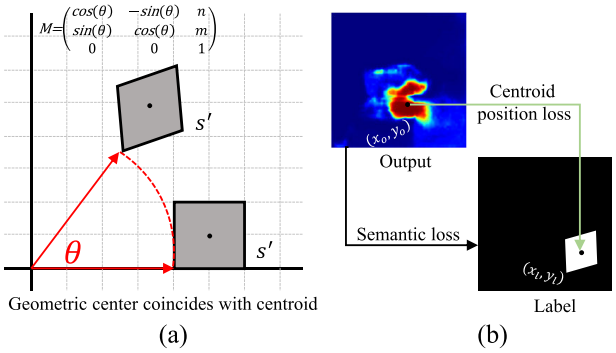


Fig. 4. Illustration of the loss function. (a) Geometric center coinciding with the centroid. (b) Loss function.

the sensed images. The formula of the feature map centroid and the loss function calculation process are as follows:

$$C_{IJ} = \sum_{i}^{P} \sum_{j}^{w} P'_{ij} \tag{3}$$

where P'_{ij} is the semantic position probability value of pixel (i, j)

$$C_I = \sum_{i}^{w} \sum_{j}^{w} i \cdot P'_{ij} \quad C_J = \sum_{i}^{w} \sum_{j}^{w} j \cdot P'_{ij} \tag{4}$$

where C_I and C_J represent the weights of all output positions in i and j coordinates, respectively. The centroid position coordinates of C'_{ij} are calculated as follows:

$$x = \frac{C_I}{C_{IJ}}, \quad y = \frac{C_J}{C_{IJ}}. \tag{5}$$

The position loss function L_p is calculated as follows:

$$L_p = (x - x_{\text{true}})^2 + (y - y_{\text{true}})^2 \tag{6}$$

where x_{true} and y_{true} denote the true coordinate position of the template centroid in the reference image.

During training, the truth label is the position of each pixel in the sensed image that corresponds to the reference image. The semantic loss function L_s is defined as the cross-entropy loss between the position dependence matching probability p and the true label, which is defined as

$$L_s = \frac{1}{N} \sum_{i} -\left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)\right] \tag{7}$$

where N is the number of matched images in a batch during training, y is the ground-truth label of the sensed image in sample i that matches the reference image, with match denoted as 1 and mismatch denoted as 0, and p denotes the matching position probability that the network architecture predicts sample i to be a match.

The final loss is described as

$$L = \alpha L_p + \beta L_s \tag{8}$$

where α and β are the weights of two loss functions, and $\alpha + \beta = 1$.

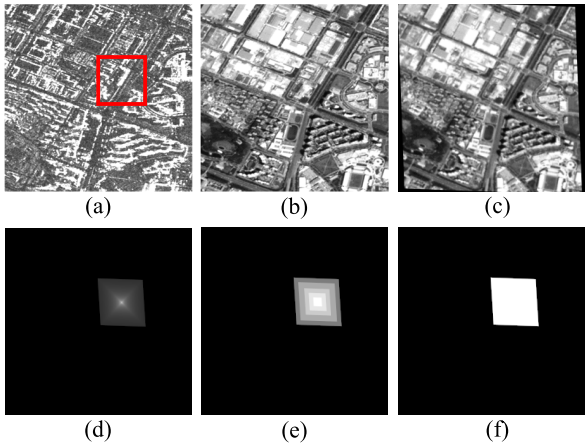


Fig. 5. Example to illustrate the dataset generation. (a) SAR image. (b) Optical image. (c) Warped image. (d)–(f) Three types of training labels.

IV. Datasets and Workflows

A. Datasets

To train the SPPD, SAR–optical image datasets are required. We use the SAR–intensity, PS-RGB from the SpaceNet [39] dataset for training, validating, and testing the network model. The SAR data come from Capella Space's X-band quadrupole sensor mounted on the aircraft. Optical imagery was acquired by Maxar WorldView-2 with a spatial resolution of 0.5 m. This imagery includes panchromatic bands, panchromatically sharpened RGB, and RGBNIR data, all with a resolution of 0.5 m.

Fig. 5 depicts the process of creating SAR–optical images and labels for training, and Figs. 5(a) and (b) are SAR and optical images, with pixel alignment. The optical image is first warped using a random affine transformation matrix, where the transformed image is the reference image in Fig. 5(c). The matching semantic labels with corresponding relationships are then generated based on the affine transformation relationship between the SAR–optical images, and randomly cropped image patches (red border) from Fig. 5(a) and (c) are used as training data.

In addition, we generate three types of labels separately, as shown in Fig. 5(d)–(f).

1) *Equivariant label:* From the outside to the inside, the padding values of each row and column present an equal-variance arrangement. For example, we set the template width to w, and the padding sequence will be $(1/0.5w)$, $(2/0.5w)$, $(3/0.5w)$, $\ldots$, $(i/0.5w)$, where i is $0.5w$ and only the centroid matching probability value is 1.0.
2) *Stepwise label:* From outside to inside, multiple rows are filled with the same value, and each step filling presents an equal arrangement. For example, let the template width be w, the sequence of padding is $(1/n)$, $(2/n)$, $\ldots$, 1, where n denotes the number of steps.
3) *0-1 label:* They are filled with 1.0. These three types of labels are used to assign different confidence values to the pixels surrounding the matching points, which are then compared for validity.

Fig. 6. Example of large image matching to obtain matching correspondence points. (a) SAR image. (b) Optical images. The dataset utilizes a unique combination of SAR imagery (0.5 m) from Capella Space and electro-optical imagery (0.5 m) from the Maxar WorldView-2 satellite.

B. Training and Matching

We crop the data in SpaceNet to generate 250,000 pairs of training data. In this article, the experiments are run on an AMAX workstation with Ubuntu 18.04 LTS and RTX3090Ti GPU and 128 GB RAM, where the initial learning rate is 0.0001 and the model is trained for ten epochs. We adopt the Adam optimizer, which is an adaptive learning rate optimization algorithm. The weight combination (α, β) of the loss function is set to $(0.3, 0.7)$. In the registration step, we crop the SAR image to obtain the sensed image. The trained network uses the reference and sensed image as image pairs to predict the position of a correspondence between the two images. We use sliding-based clipping on the SAR and optical image to obtain the set of matching points, based on the need for overall image registration.

Fig. 6 depicts the process of SAR and optical image cropping to generate the sensed and reference image pairs, where Figs. 6(a) and (b) are the SAR and optical images. Depending on the size of the network, the SAR and optical images are cropped according to their respective sizes to obtain $T_{11}, T_{12}, \ldots, T_{ij}$, as the sensed image, and $R_1, \ldots, R_n$. Each T_{ij} from Fig. 6(a) and each R_n from Fig. 6(b) are used as corresponding input pairs to obtain the matching points. For the obtained set of matching points, we use the random sample consensus algorithm (RANSAC) to globally constrain the false matches. RANSAC eliminates the erroneous matches from a set of obtained point set data by a random sampling and voting scheme.

V. Experiment

In this section, we first evaluate the effect of SAR image size on image matching accuracy. Compared with the existing methods, the matching performance of the network and its effectiveness on SAR–optical image matching are evaluated. Then, the overall performance of the entire pipeline in a larger test scenario is evaluated. Finally, a network module and loss function ablation studies are performed to motivate the selection of our network structure and loss function weights.

A. Selection of Template Size

There is a wide range of patch correlation between the reference and sensed images, necessitating further research into what types of semantic information may be used to identify the matching relationship between images. Specifically, the effect of the sensed image size on the matching accuracy was investigated. We expect to choose a smaller size of SAR images when possible to improve the computational efficiency while providing more freedom for the reference image. Let the size of the reference image be $w \times w$, while the minimum size of the SAR image is 1×1 pixel and the maximum is $w \times w$ pixels. However, using a pixel as a matching template cannot provide more semantic information, leading to wrong correspondence. A suitable SAR image size needs to be selected while ensuring matching accuracy. We experimented with reference image with sizes of 256×256 pixels and SAR image size 48, 64, 96, 128 pixels to evaluate the effect of the SAR image size on the matching accuracy. Moreover, the model's training hyperparameters are kept consistent, including batch size, epoch number, and learning rate. The $L2$ error of the matching position is used to evaluate the effect of SAR image size on matching accuracy

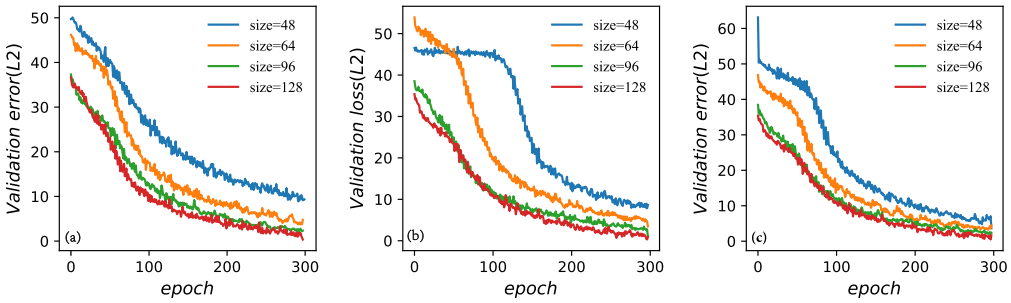
$$L2 = \sqrt{\left(x_i^p - x_i^t\right)^2 + \left(y_i^p - y_i^t\right)^2} \tag{9}$$

where (x_i^p, y_i^p) is the centroid of prediction and (x_i^t, y_i^t) is the true matching position.

Fig. 7 depicts the $L2$ errors of the validation data on the equivariant, stepwise, and 0-1 labels with different sizes of SAR images during the training process. The first three columns show the $L2$ errors of equivariant, stepwise, and 0-1 labels on the validation dataset at 0–300 epoch with image sizes of 48, 64, 96, 128 pixels. The last column shows the average $L2$ errors of the three labels with different SAR image sizes, where Equ, Ste, and Zer are abbreviations for equivariant, stepwise, and 0-1 labels. The $L2$ error trend of the validation data is consistent. They gradually decrease as the size of the SAR image grows larger, eventually stabilizing in a range of values. Their $L2$ errors on the three types of labels are almost the same as those of the 96×96 pixels size image as a whole, when the SAR image size increases from 96×96 to 128×128 pixels. Therefore, to improve the matching accuracy and computational efficiency simultaneously, we chose 96×96 pixels as SAR image size in the subsequent experiments. In addition, we compared the overall accuracy of the three types of labels on the validation dataset. The last one in Fig. 7 shows the overall matching performance of the three labels, where the $L2$ error of the 0-1 label is lower than that of the equivariant and stepwise labels. For different scenes, the models trained using the three different labels may have varying matching performances. Therefore, in Section V-F, we detail the matching responses of the models trained on the three types of labels.

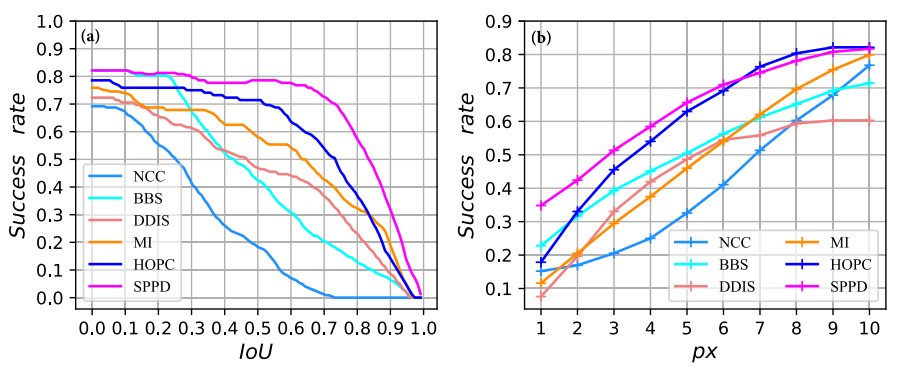
B. Matching Performance

Since the matching accuracy plays a crucial role in the overall image registration, we evaluated the performance of our method relative to the state-of-the-art methods. We compared SPPD with NCC, best buddies similarity (BBS) [40], deformable diversity similarity (DDIS) [41], MI and HOPC, where our method uses 0-1 labels with $(\alpha, \beta) = (0.3, 0.7)$.



Sizes	Equ	Ste	Zer
$s = 48$	14.25	13.97.1	14.15
$s = 64$	6.25	6.14	6.08
$s = 96$	2.16	2.23	2.10
$s = 128$	2.06	2.17	1.97

Fig. 7. $L2$ errors of the validation dataset with different SAR image sizes on three types of labels. (a) Equivariant label. (b) Stepwise label. (c) 0-1 label.



Methods	L2(px)	Average(sec.) CPU	GPU
NCC	9.126	50.60	–
BBS	6.497	15.46	–
DDIS	5.613	7.16	–
MI	5.016	70.12	–
HOPC	4.670	32.16	–
SPPD	3.175	1.71	0.93

Fig. 8. Matching performance comparison by matching success rate. (a) IoU. (b) $L2$.

We evaluated the proposed method on a test dataset and used the intersection over union (IoU) of template matching, matching accuracy with different pixel thresholds, average $L2$ error and average processing speed (sec/sample) as evaluation metrics. The performance of the proposed and all compared methods is shown in Fig. 8. The first two graphs show the matching success rate with different IoU and different pixel thresholds, respectively. The last graph shows the average $L2$ error of matching on the dataset and the average consumption time of data processing on CPU, GPU, where "-" means data processing cannot be implemented on GPU. It can be seen that the performance of SPPD is better than the compared methods. The success rate curves at different IoU thresholds are overall higher than those of the compared methods. The average $L2$ error of SPPD in the test dataset is 3.2, which is significantly lower than NCC (9.1), BBS (6.5), DDIS (5.6), MI (5.0), and HOPC (4.7).

SPPD took an average time of 1.71 s to process an image match on the CPU. This performance surpassed the average time taken by other methods such as NCC, BBS, DDIS, MI, and HOPC. Distinct from these methods, SPPD employs a weighted semantic position probability feature map to derive the primary match as the result, which significantly enhances the matching efficiency. The inherent design mechanism of SPPD allows for acceleration using GPU, thereby further boosting its matching speed. Additionally, we evaluated the complexity of SPPD using two metrics: the size of the model parameters and the number of floating point operations per second (FLOPs). Our findings indicate that the SPPD has a parameter size of 170 MB and a computational requirement of 8 GFLOPs. This data further underscore the efficiency and practicality of SPPD in image matching tasks.

Fig. 9 provides the qualitative matching results of SPPD in some typical scenes, where the deep red color represents the region as the matching position. The results showed that SPPD produced more accurate matches. The output of the comparison method ideally has a response at only one pixel or one region on the feature map. However, for the other methods, the response is obtained at different positions in the generated feature map and far from the correct matching point.

C. Overall Matching Performance

In Section IV-B, the effectiveness of SPPD is assessed. For a comprehensive evaluation of SPPD's registration performance on entire images, correspondence was established on large-scale remote sensing images. Four pairs of substantial SAR and optical images from the SpaceNet dataset were chosen. These images predominantly capture urban scenes, as depicted in Figs. 10–13, with dimensions of 1796 × 1146 pixels. Keypoints were manually identified to ascertain the correspondence between the images. The four pairs of images are denoted by I_1, I_2, I_3, and I_4.

The proposed framework identifies correspondence points from a given matching template. While theoretically possible to crop a template for each pixel in the reference image to determine the correspondence, such an approach would lead to computational redundancy. For a consistent comparison, the image has been segmented based on the grid size of the matching template (96 × 96 pixels). The RANSAC method is employed to eliminate incorrect matching points. The performance of SPPD is evaluated against other methods including SIFT, Affine-SIFT, PSO-SIFT, SAR-SIFT, RIFT, and

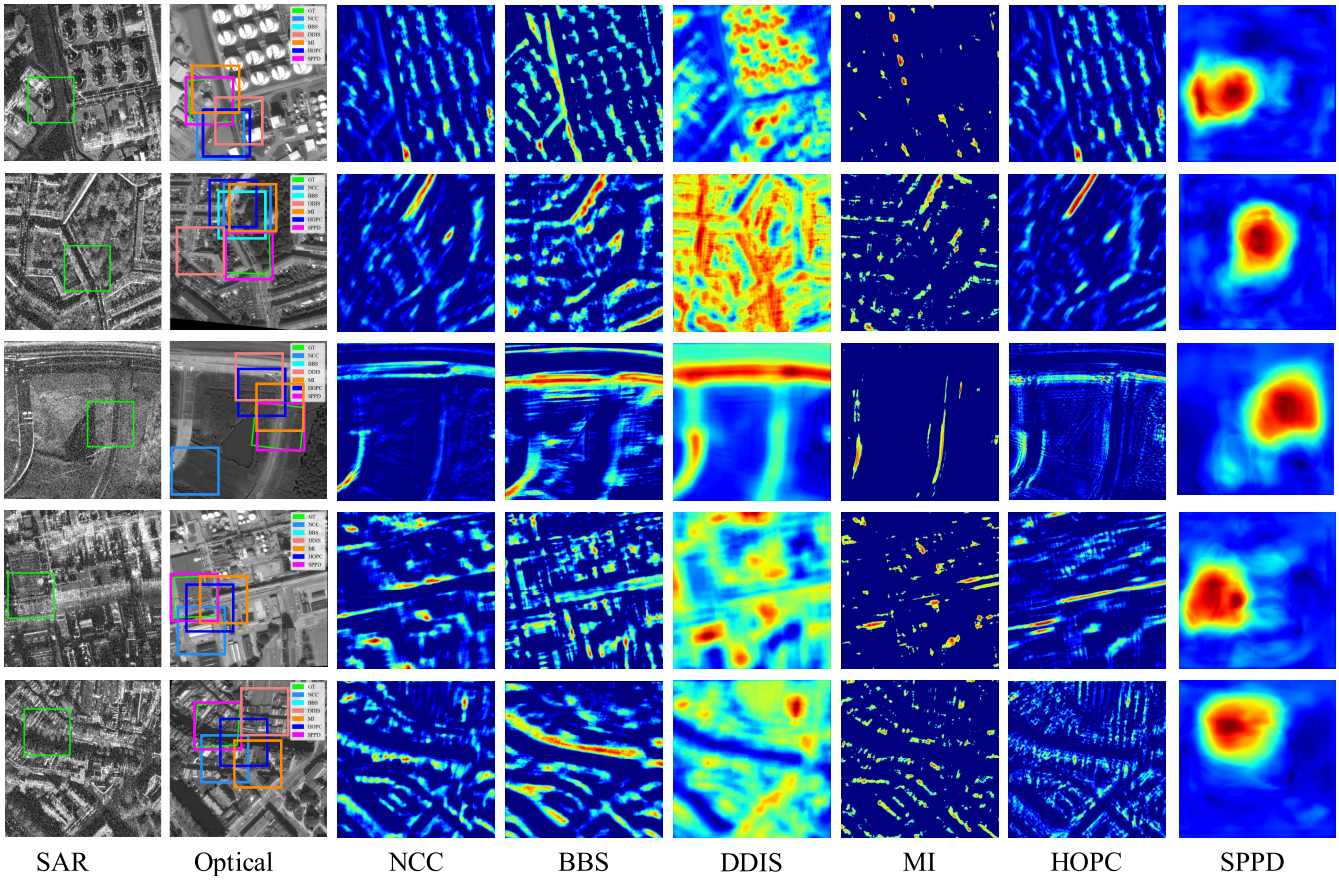


Fig. 9. Qualitative matching performance of a typical scene image. (Left to right) Cropped sensed images, and matching results on the reference image (with colors representing different methods).

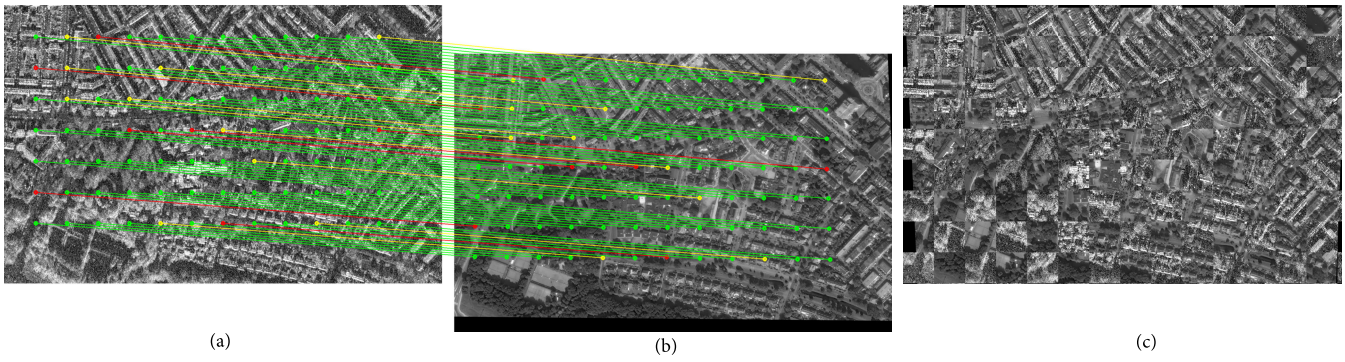


Fig. 10. Qualitative matching results for the I_1. (a) SAR image. (b) Optical image. (c) Checkerboard mosaicked image.

HOPC. Evaluation metrics include the mean average precision with a 2-pixel threshold (mAP) and the average $L2$ error across all matching points. For methods such as SIFT, SAR-SIFT, and PSO-SIFT, matching is conducted based on the Euclidean distance ratio between the nearest and the second nearest neighbors of the respective features. Ratios of 0.6, 0.7, 0.8, and 0.9 are tested, and the lowest $L2$ error is chosen for further comparison. Results from the online algorithm application are adopted for Affine-SIFT. For RIFT, parameters and algorithms provided in [9] are utilized. Lastly, the Harris detection algorithm is employed to identify corner points for HOPC matching.

Table I shows the matching results for the four pairs of large size images. A larger mAP and smaller $L2$ indicate higher matching performance. As can be seen from Table I, SPPD achieves the lowest $L2$ of 3.162, 3.361, 3.034, 3.217. The results show that the overall matching performance of SPPD outperforms other comparative methods in all image pairs. SIFT and Affine-SIFT, with an mAP of 0, have the worst accuracy, which may be due to the low repeatability of keypoints due to the radiometric differences between SAR and optical images. The PSO-SIFT method applies multiple constraints and has a lower $L2$ than SIFT, but the overall matching results are not satisfactory. The detector used in SAR-SIFT is

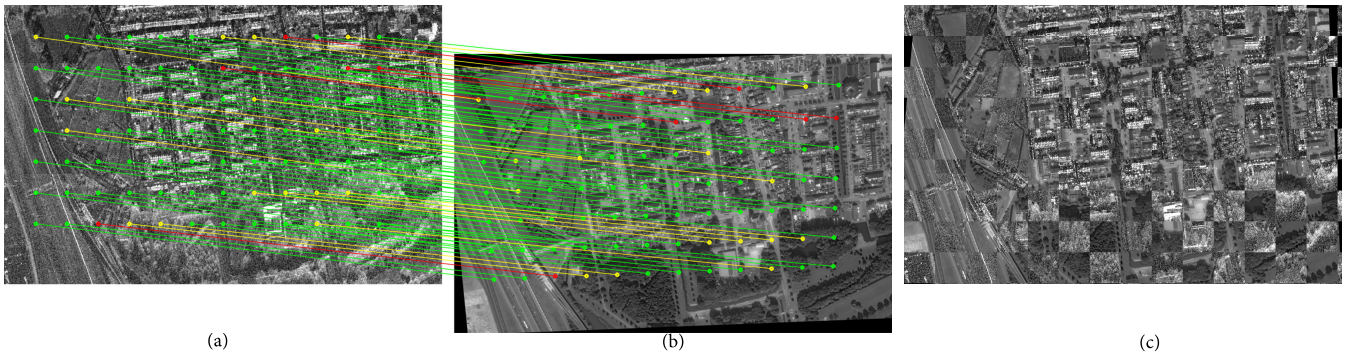


Fig. 11. Qualitative matching results for the I_2. (a) SAR image. (b) Optical image. (c) Checkerboard mosaicked image.

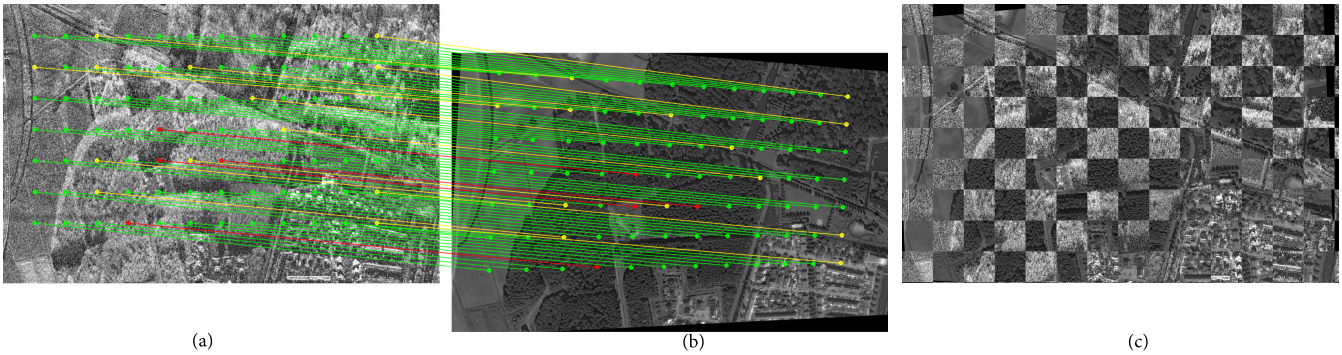


Fig. 12. Qualitative matching results for the I_3. (a) SAR image. (b) Optical image. (c) Checkerboard mosaicked image.

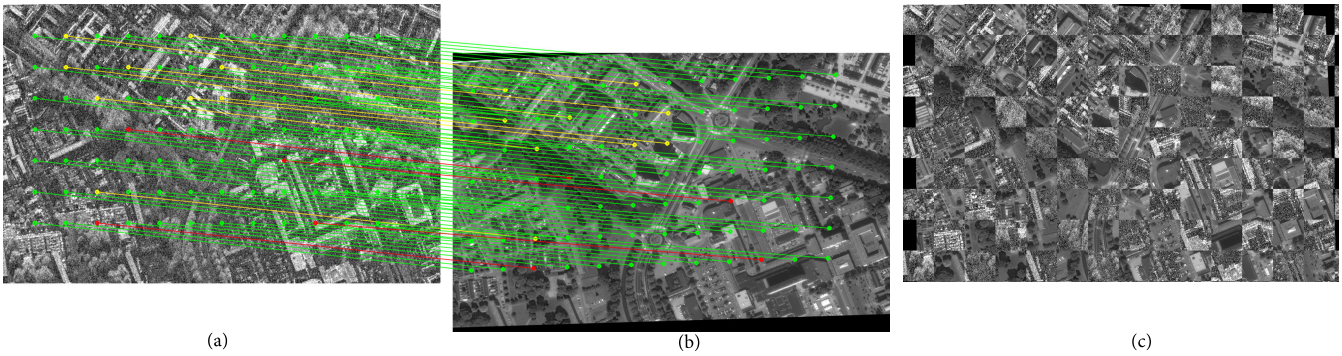


Fig. 13. Qualitative matching results for the I_4. (a) SAR image. (b) Optical image. (c) Checkerboard mosaicked image.

TABLE I

MATCHING PERFORMANCE ON I_1, I_2, I_3, I_4. THE SYMBOL "-" REPRESENTS THE SEVER MISREGISTRATION

	SIFT		**ASIFT**		**PSO-SIFT**		**SAR-SIFT**		**RIFT**		**HOPC**		**SPPD**	
	mAP (%)	$L2$ (**pixel**)	**mAP** (%)	$L2$ (**pixel**)	**mAP** (%)	$L2$ (**pixel**)	**mAP** (%)	$L2$ (**pixel**)	**mAP** (%)	$L2$ (**pixel**)	**mAP** (%)	$L2$ (**pixel**)	**mAP** (%)	$L2$ (**pixel**)
I_1	11.31	9.9	–	12.8	10.4	10.6	24.10	5.8	34.70	4.4	23.42	3.8	41.00	3.2
I_2	–	11.2	–	14.7	14.3	9.7	19.70	6.0	32.30	4.0	15.70	4.1	45.20	3.4
I_3	7.22	10.5	–	15.1	9.6	11.5	28.60	5.8	31.20	3.4	23.92	4.1	43.64	3.0
I_4	10.60	10.0	–	15.2	11.7	11.1	19.60	5.9	32.90	3.6	24.40	4.3	41.93	3.2

too sensitive to nonlinear radiation differences, resulting in low matching performance accuracy. HOPC obtains high matching accuracy on I_1, and a decrease in accuracy can be seen on I_2, I_3, I_4, which may be due to the large deformation. The matching accuracy and precision of RIFT on I_1, I_2, I_3, and I_4 are better than SIFT, Affine-SIFT, PSO-SIFT, SAR-SIFT, and HOPC. Overall, the accuracy and precision of SPPD are higher than that of RIFT.

The qualitative evaluation results of SPPD are illustrated in the tessellated mosaic images as in Figs. 10–13, where green and yellow indicate 3, 5-pixel errors and red indicates mismatching. For the tessellated mosaic images, the edges of their

TABLE II
NETWORK CONFIGURATION USED IN THE ABLATION STUDY

Network	ResNet-50	FPN	Pos-loss	Sem-loss
$PosAttenS$	N	N	N	Y
$PosAttenSR$	Y	N	N	Y
$PosAttenSRF$	Y	Y	N	Y
$PosAttenRFC$	Y	Y	Y	N
$PosAttenSRFC$	Y	Y	Y	Y

TABLE III
MATCHING PERFORMANCE OF THE CORRESPONDENCE NETWORK

Network	Matching accuracy	Matching precision	
	mAP (%)	Average $L2$ (pixel)	STD (pixel)
$PosAttenS$	10.26	6.5	2.3
$PosAttenSR$	15.76	6.1	2.1
$PosAttenSRF$	25.61	5.6	1.3
$PosAttenRFC$	39.49	3.8	1.9
$PosAttenSRFC$	42.19	3.2	1.1

features were continuous and the overall area overlapped very well in our results. The results above provide both quantitative and qualitative evaluations of the effectiveness of SPPD in large-scene remote sensing images.

D. Ablation Study

1) Network Configuration: An ablation study was performed to compare the performance of adding various network architectures. Since the position attention module is central to building the semantic position probability distribution used for image matching, we combined in different ways the position attention module, multilayer CNN as the base network ($PosAttenS$), ResNet-50, FPN, centroid location, and semantic loss function. In Table II, the use of a specific network, centroid position (Pos-loss) or semantic loss (Sem-loss) function, is indicated by a yes (Y) or no (N).

The networks were trained according to Section IV. All networks were trained on the same training data. We evaluated the performance based on the mean and standard deviation of the $L2$ between matching positions, and the matching accuracy with 2-pixel thresholds (mAP). The ablation models are described in Table II.

Table III shows that $PosAttenSRF$ has a greater average matching accuracy than $PosAttenSR$, which is due to FPN's ability to boost information fusion between distinct residual blocks. The addition of the loss function based on centroid position achieved a significant improvement in the matching accuracy.

Furthermore, simultaneously using the semantic and centroid position loss functions achieved the best result. Therefore, the $PosAttenSRFC$ was selected as our matching framework, and all further experiments were conducted used this setup.

2) Position Attention Module: We experimented with the positional attention module using the following state-of-the-art semantic segmentation networks (PSPNet [42], DeepLabV3 [43], PSANet [42], DeepLabV3+ [44], UPerNet [45], NonLocal Net [46], EncNet [47], DANet [37], DANet [37], FastFCN [48], Fast-SCNN [49], CGNet [50], BiSeNetV2 [51], STDC [52], STDC [52], SETR [53], DPT [54], DPT [54], Segmenter [55], and SegFormer [56]) to train the datasets, verifying the effectiveness of the position attention module. The basic idea behind the SPPD structure was to use two weight-sharing ResNet-50 + FPN architectures to construct pyramidal features for both the reference and sensed images, and then used the position attention mechanism to establish position dependencies between two images. To use the current segmentation network, we complemented the input sensed images by 0 to form the same size as the reference image. Then, they were joined together to obtain the size of $(W, H, 2)$ and input to the semantic segmentation network, as shown in Fig. 14. All other parameters were kept consistent, where the training labels were 0-1 labels and the loss functions were semantic and centroid position loss functions. The detailed experimental results are shown in Table IV.

TABLE IV
MATCHING RESULTS ON THE STATE-OF-THE-ART SEMANTIC SEGMENTATION NETWORKS

Network	Matching accuracy	Matching precision	
	mAP (%)	Average $L2$ (pixel)	STD (pixel)
PSPNet	1.03	8.2	1.3
DeepLabV3	1.83	7.2	1.9
PSANet	1.44	7.0	1.5
DeepLabV3+	1.92	7.2	1.8
UPerNet	1.06	8.6	2.0
NonLocal Net	1.11	7.5	1.8
EncNet	1.58	7.1	2.7
DANet	1.77	8.4	2.4
FastFCN	1.80	7.2	1.5
Fast-SCNN	1.87	7.3	2.8
CGNet	1.71	7.1	2.8
BiSeNetV2	1.96	7.9	1.6
STDC	1.88	8.0	2.3
SETR	1.05	8.6	2.9
DPT	1.88	7.1	2.7
Segmenter	1.57	8.9	2.9
SegFormer	1.77	8.8	2.9

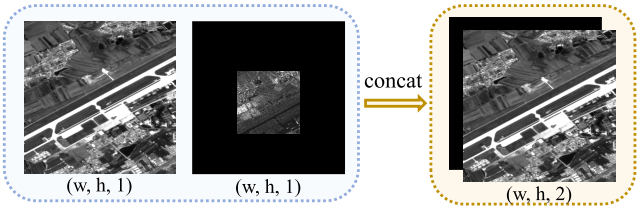


Fig. 14. Illustration of production for sensed and reference images that can be input to the semantic segmentation network.

As can be seen from Table IV, the average $L2$ scores between matched positions of the semantic segmentation frameworks are overall lower than that of our proposed method. The average matching accuracy at 2-pixel thresholds

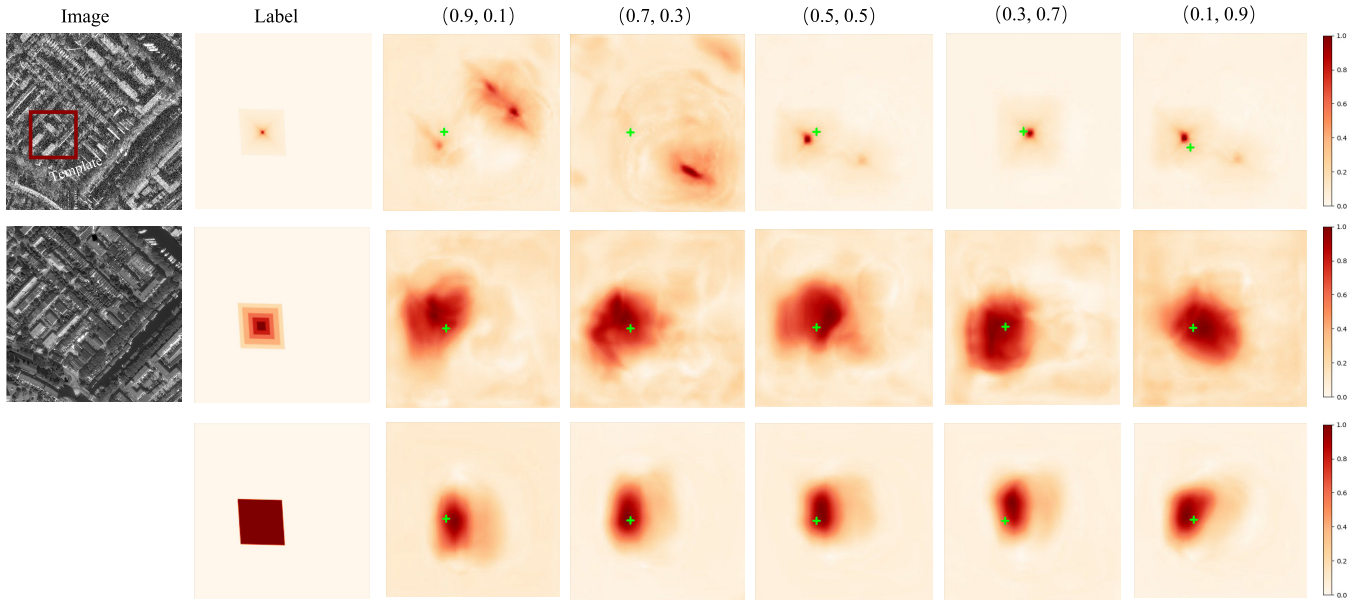


Fig. 15. Qualitative comparison of heat maps generated by different weight combinations of the loss function on the three training labels.

is less than 2.0%. These semantic segmentation algorithms rely solely on a data-driven fit from input to output, focusing on the training dataset instead of image correspondence. In contrast, SPPD creates a matching model based on the pixel alignment relationship by mapping matching position probabilities between two images. Thus, the position attention module achieves high matching accuracy, illustrating its effectiveness.

TABLE V
*L*2 ERROR WITH (α, β) COMBINATIONS BETWEEN LOSS FUNCTIONS

	Combination(α, β)				
Label	**(0.9, 0.1)**	**(0.7, 0.3)**	**(0.5, 0.5)**	**(0.3, 0.7)**	**(0.1, 0.9)**
Equ	3.6	3.6	3.5	3.5	**3.4**
Ste	3.7	3.5	3.1	**2.9**	3.3
Zer	3.9	3.2	2.8	**2.8**	3.0

E. Parameter Analysis

To further describe the design of the loss function in Section III, we conducted a hyperparameter search on the loss function to compare the performance of the semantic and centroid position loss functions for determining the matching results. The semantic and centroid position loss functions were tested separately in Section V-D. The total matching accuracy when utilizing semantic or centroid location loss functions alone is lower than when combining both loss functions, as evidenced by the experimental findings. This is because the semantic loss function leads the network output to suit the labels, whereas the centroid position loss function maximizes matching from a single point without having to fit the training labels completely. Combining the two loss functions can further improve the matching accuracy because the semantic loss function can guide the network output to fit the labels, while the centroid position loss function can offset some of the pixel-independent bias due to radiation differences between the SAR and optical images.

We assigned different weights (α, β) to the two loss functions, and experimented on the three labels. The matching accuracy of the network with different combinations is detailed in Table V. The highest matching accuracy for the stepwise and 0-1 labels is obtained on the combination $(0.3, 0.7)$, whereas the best matching accuracy for the equivariant label is found on the combination $(0.1, 0.9)$.

The response of each label represents the correspondence of the semantic position probabilities of the sensed image in a reference image, which may contain some small groups of pixels in a neighborhood. These pixels would theoretically not exist until the network is fully trained. The heat maps of the three labels with different weight combinations are shown in Fig. 15, where the deeper color denotes the value of the region closer to 1.0 and the cross mark represents the centroid position of the matched result. The plots reveal that none of the generated feature maps are identical to the labels, and instead have activated responses around the centroid position centered at the labels. The extent and placement of these output results' responses are irregular, with a shape that differs significantly from that of a regular geometric quadrilateral. There is no way to identify that location as the centroid. This means that by containing the centroid position loss function in the deep learning network, the features are semantically abstracted at a high level.

More specifically, the weights $(\alpha, \beta) = (0.1, 0.9)$ mainly emphasize semantic information. The resulting feature maps are close to the labels. However, the accuracy of the matching results for the centroid position is lower than the combination $(0.3, 0.7)$. Similarly, for the combination $(\alpha, \beta) = (0.9, 0.1)$, the overall matching accuracy is also less than that of the combination $(0.3, 0.7)$. The original intention for designing the network is to find a dependency or match for each pixel

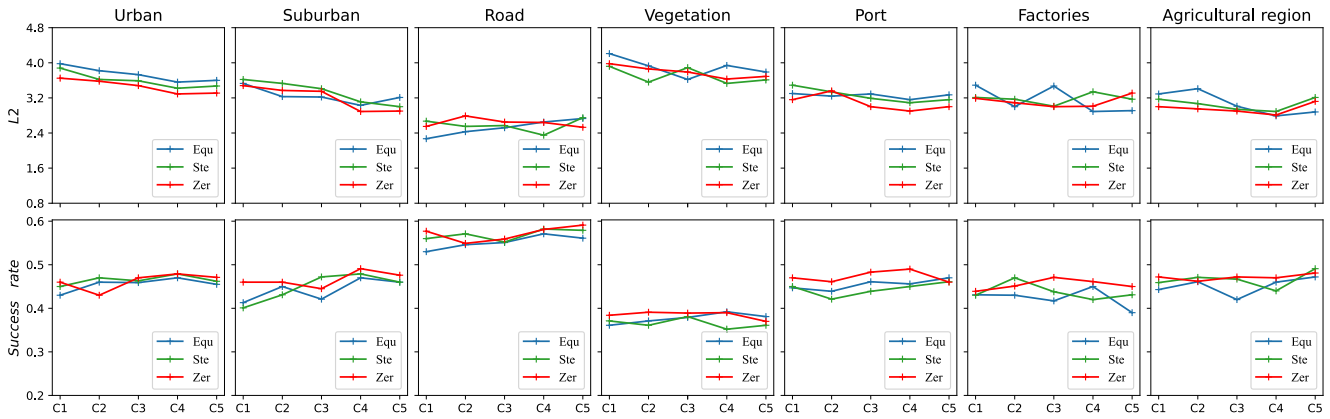


Fig. 16. Quantitative matching results on the three training labels with different weight combinations of the loss function. (First row) Matching average $L2$ error. (Second row) Matching success rate at 3-pixel error.

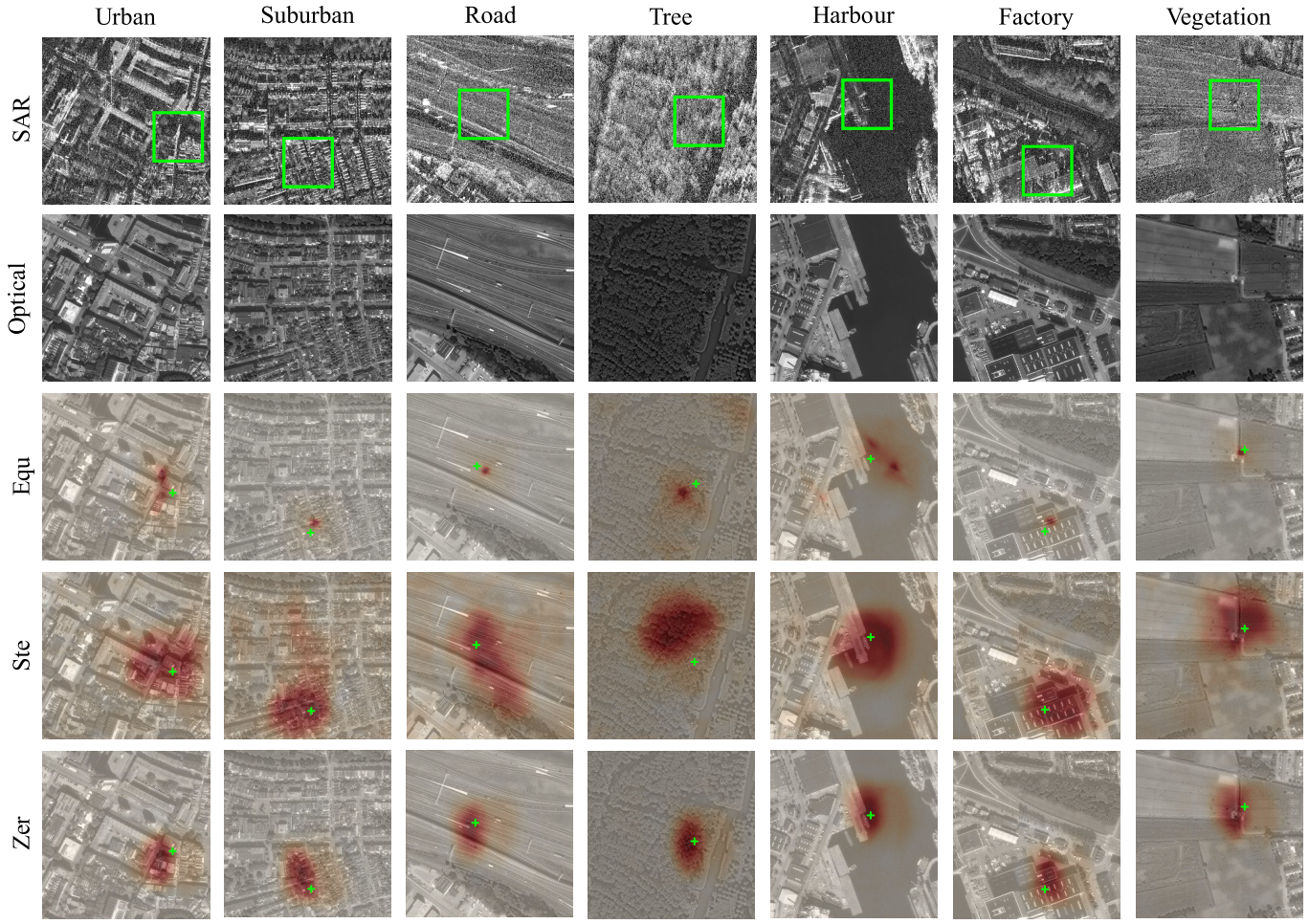


Fig. 17. Matching results for typical scenes on the three training labels.

of the sensed image based on the reference image. However, due to the nonlinear radiometric variations between the SAR and optical images, an exact match for each pixel on the reference image is impossible to achieve. Therefore, adjusting the weights (α, β) of the semantic position dependence and the centroid position loss function during training enable the network to obtain an accurate matching.

F. Training Label Analysis

The optimal loss function weight combinations for the equivariant labels were found to be $(0.1, 0.9)$ and $(0.3, 0.7)$ for the stepwise and 0-1 labels in Section V-E. The SpaceNet dataset contains a variety of feature scenes, including urban, suburban, roads, vegetation, ports, factories, and agricultural

areas. They exhibit different intensity values and texture structures on optical and SAR images. Thus, we evaluate the matching accuracy of the three types of labels on different feature scenes. We count the overall division of the features in the SpaceNet dataset, according to the percentage of the stitched features compared to the total number of pixels. The percentages of urban buildings, suburban buildings, roads, vegetation, ports, factories, and agriculture are 20.60%, 22.90%, 7.60%, 23.40%, 5.10%, 6.40%, and 14.00%, respectively. For a fair comparison, we select the same number of feature scenes on the whole dataset for training and comparison. $L2$ error is used as an evaluation metric. Fig. 16 shows the matching results for each scene with different combinations of loss function weights for the three types of labels.

The results show that the 0-1 labels with the combination $(0.3, 0.7)$ achieve the highest matching accuracy and precision for suburban and urban buildings, which is similar to the results obtained in Section V-E. The matching precision and accuracy are higher in road scenes than in other scenes, with the stepwise label with the combination $(0.5, 0.5)$ achieving the highest matching precision and the equivariant label with the combination $(0.9, 0.1)$ achieving the highest accuracy. This could be because the roads have a more pronounced texture structure and do not suffer from geometric distortion induced by the SAR–optical imaging shooting angle, allowing our network to achieve better matching accuracy and precision. It also demonstrates that the equivariant labels with the combination $(0.9, 0.1)$ are primarily intended to boost centroid position dependent probability and improve centroid position matching accuracy.

Moreover, the overall accuracy and precision of their matching results are the lowest in the matching of vegetation scenes. The intensity values of the SAR images show that there are no distinguishable features in the vegetation, which degrades the overall performance. Considering the context in the vegetation, a larger scale, increasing the size of the SAR image, is also required to obtain a higher matching accuracy. Furthermore, the port, factory, and agricultural scenes, all of which have considerable texture differences. They achieve better matching accuracy and precision with 0-1 labels. The qualitative comparison results are shown in Fig. 17, where the deeper color represents the matching position of the response and the green cross indicates the matching centroid position.

VI. Conclusion

In this article, we proposed a deep neural network model for solving current SAR and optical image registration. First, a weight-sharing ResNet-50 + FPN was utilized to create a pyramidal feature structure for SAR and optical images, thereby providing multiscale features. Second, those features containing high-level semantic information were inserted into our proposed position attention module to obtain spatial position dependencies between the two images. Different from the pixel-by-pixel search matching method, each pixel position of SAR imagery was mapped onto the optical image to obtain the matching position probability distribution in this network. The matching probability position weighted average loss function transformed the semantic position probability distribution alignment into a point-to-point matching problem, consequently improving the matching accuracy. Moreover, we designed three types of training labels to compare their impact on matching accuracy. The proposed method outperformed state-of-the-art methods of SAR–optical images matching with various different scale scenes, which demonstrated its effectiveness in multimodal remote sensing image registration. This method could offer a new solution for the multimodal imagery registration challenges faced by the remote sensing community.

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] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, and G. Zhang, "A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 331–350, Jun. 2022.

[3] M. Schmitt, F. Tupin, and X. X. Zhu, "Fusion of SAR and optical remote sensing data—Challenges and recent trends," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5458–5461.

[4] 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.

[5] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, Sep. 2021.

[6] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, and Y. Ye, "R_2FD_2: Fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606115.

[7] N. Dey, P. Nandi, N. Barman, D. Das, and S. Chakraborty, "A comparative study between Moravec and Harris corner detection of noisy images using adaptive wavelet thresholding technique," 2012, *arXiv:1209.1558*.

[8] 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.

[9] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[10] K. Yang, A. Pan, Y. Yang, S. Zhang, S. Ong, and H. Tang, "Remote sensing image registration using multiple image features," *Remote Sens.*, vol. 9, no. 6, p. 581, Jun. 2017.

[11] M. B. Hisham, S. N. Yaakob, R. A. A. Raof, A. B. A. Nazren, and N. M. Wafi, "Template matching using sum of squared difference and normalized cross correlation," in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, Dec. 2015, pp. 100–104.

[12] J. P. Kern and M. S. Pattichis, "Robust multispectral image registration using mutual-information models," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1494–1505, May 2007.

[13] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[14] A. A. Cole-Rhodes, K. L. Johnson, J. LeMoigne, and I. Zavorin, "Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient," *IEEE Trans. Image Process.*, vol. 12, no. 12, pp. 1495–1511, Dec. 2003.

[15] S. Klein, M. Staring, and J. P. W. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2879–2890, Dec. 2007.

[16] Y. Ye, J. Shan, S. Hao, L. Bruzzone, and Y. Qin, "A local phase based invariant feature for remote sensing image matching," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 205–221, Aug. 2018.

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

[18] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[19] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3456–3465.

[20] M. Dusmanu et al., "D2-Net: A trainable CNN for joint detection and description of local features," 2019, *arXiv:1905.03561*.

[21] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4938–4947.

[22] S. Cui, A. Ma, L. Zhang, M. Xu, and Y. Zhong, "MAP-Net: SAR and optical image matching via image-based convolutional network with attention mechanism and spatial pyramid aggregated pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 1000513, doi: 10.1109/TGRS.2021.3066432.

[23] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, "A multiscale framework with unsupervised learning for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622215.

[24] D. Xiang, Y. Xu, J. Cheng, Y. Xie, and D. Guan, "Progressive keypoint detection with dense Siamese network for SAR image registration," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 5, pp. 5847–5858, Oct. 2023.

[25] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 166–179, Nov. 2020.

[26] Y. Fang, J. Hu, C. Du, Z. Liu, and L. Zhang, "SAR-optical image matching by integrating Siamese U-Net with FFT correlation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3100531.

[27] T. Bürgmann, W. Koppe, and M. Schmitt, "Matching of TerraSAR-X derived ground control points to optical image patches using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 241–248, Dec. 2019.

[28] 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.

[29] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, "Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images," *Remote Sens.*, vol. 9, no. 6, p. 586, Jun. 2017, doi: 10.3390/rs9060586.

[30] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and SAR image registration based on feature decoupling network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5235913.

[31] L. Li, L. Han, M. Ding, H. Cao, and H. Hu, "A deep learning semantic template matching framework for remote sensing image registration," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 205–217, Nov. 2021.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[34] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–32, Oct. 2021, doi: 10.1145/3465055.

[35] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural. Inf. Process Syst. (NeurIPS)*, vol. 30, 2017, pp. 1–11.

[37] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

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

[39] J. Shermeyer et al., "SpaceNet 6: Multi-sensor all weather mapping dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 196–197.

[40] T. Dekel, S. Oron, M. Rubinstein, S. Avidan, and W. T. Freeman, "Best-buddies similarity for robust template matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2021–2029.

[41] I. Talmi, R. Mechrez, and L. Zelnik-Manor, "Template matching with deformable diversity similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 175–183.

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

[43] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[44] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV).*, 2018, pp. 801–818.

[45] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV).*, 2018, pp. 418–434.

[46] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[47] H. Zhang et al., "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[48] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation," 2019, *arXiv:1903.11816*.

[49] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*.

[50] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[51] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.

[52] M. Fan et al., "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9716–9725.

[53] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[54] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[55] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[56] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural. Inf. Process Syst. (NeurIPS)*, vol. 34, 2021, pp. 12077–12090.

Liangzhi Li received the B.Sc. degree in surveying and mapping engineering from Shandong Jiaotong University, Jinan, China, in 2017, and the M.Sc. degree from the School of Geology Engineering and Geomatics, Chang'an University, Xi'an, China, in 2019, where he is currently pursuing the Ph.D. degree.

His research interests include remote sensing image matching, classification, and segmentation. For more details about his work, please visit: https://www.researchgate.net/profile/Liangzhi-Li-4.

Ling Han received the B.Sc. and M.Sc. degrees in remote sensing from Wuhan University, Wuhan, China, in 1991 and 1994, respectively, and the Ph.D. degree in remote sensing from Northwest University, Xi'an, China, in 2005.

She has significantly contributed to the field, having published more than 100 SCI articles. Her primary research interests encompass deep learning applications in remote sensing image classification, registration, and artificial intelligence. Furthermore, she has delved into cloud detection in snow-covered images and landslide interpretation in the loess plateau areas.

Dr. Han, in recognition of her expertise, has served as a Reviewer for esteemed journals, such as IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL), and the *International Society for Photogrammetry and Remote Sensing (ISPRS) Journal of Photogrammetry and Remote Sensing*.

Ming Liu received the B.Sc. degree in geographic information science and the M.Sc. degree in surveying and mapping from the China University of Geosciences, Beijing, China, in 2014 and 2017, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2020.

Her main research interests include atmospheric remote sensing, environmental justice, and quantitative Earth observations.

Kyle Gao (Graduate Student Member, IEEE) received the bachelor's degree in mathematics from the University of Waterloo, Waterloo, ON, Canada, in 2016, and the master's degree in physics from the University of Victoria, Victoria, BC, Canada, in 2020. He is currently pursuing the Ph.D. degree in systems design engineering with the Geospatial Intelligence and Mapping Group, University of Waterloo.

He has published articles in the *International Journal of Applied Earth Observation and Geoinformation*, IEEE ACCESS, and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. His research interests include computer vision and deep learning.

Hongjie He received the B.Sc. degree in geomatics from the China University of Petroleum, Huadong, China, in 2016, the M.Sc. degree in cartography and geographic information systems from Lanzhou University, Lanzhou, China, in 2019, and the Ph.D. degree in geography specializing in applied Earth observations from the Geospatial Intelligence and Mapping Laboratory, University of Waterloo, Waterloo, ON, Canada, in 2023.

He has published articles in the *International Journal of Applied Earth Observation and Geoinformation*, *Canadian Journal of Remote Sensing*, and *Geomatica*, and flagship conferences, including International Geoscience and Remote Sensing Symposium (IGARSS) and International Society for Photogrammetry and Remote Sensing (ISPRS). His research interests include AI-based algorithms and software tools for information extraction from Earth observation images.

Lanying Wang received the B.Sc. degree in geomatics from Tianjin Normal University, Tianjin, China, in 2011, and the M.Sc. degree in geomatics from the University of Waterloo, Waterloo, ON, Canada, in 2016, where she is currently pursuing the Ph.D. degree with the Geospatial Intelligence and Mapping Group, Department of Geography and Environment Management.

Her research interests are developing applications of LiDAR point cloud and Earth observation imagery by deep learning techniques.

Jonathan Li (Fellow, IEEE) received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000.

He is currently a Professor of geomatics and systems design engineering with the University of Waterloo, Waterloo, ON, Canada. He has coauthored almost 600 publications, more than 150 of which were published in top remote sensing journals, including *Remote Sensing of Environment*, *International Society for Photogrammetry and Remote Sensing (ISPRS) Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *International Journal of Applied Earth Observation and Geoinformation* (JAG). He has also published papers in flagship conferences in computer vision and AI, including Computer Vision and Pattern Recognition Conference (CVPR), Association for the Advancement of Artificial Intelligence (AAAI), and International Joint Conferences on Artificial Intelligence (IJCAI). He has supervised nearly 200 master's/Ph.D. students as well as post-doctoral fellows/visiting scholars to completion. His main research interests include AI-based information extraction from Earth observation images and LiDAR point clouds, photogrammetry and remote sensing, GeoAI and 3-D vision for digital twin cities, and autonomous driving.

Dr. Li is a fellow of the Canadian Academy of Engineering, the Royal Society of Canada (Academy of Science), and the Engineering Institute of Canada. He is also the Editor-in-Chief of JAG and an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.