

Contrastive Learning of Multimodal Consistency Feature Representation for Remote Sensing Image Registration

Zhen Han , Ning Lv , Zhiyi Wang , Wei Han , Li Cong , Shaohua Wan , *Senior Member, IEEE*, and Chen Chen , *Senior Member, IEEE*

Abstract—Feature representation is a crucial issue in multimodal image registration. The handcrafted features extracted by traditional methods are highly sensitive to nonlinear radiation differences, while supervised learning methods are limited by deficient labeled samples in the remote sensing field. Therefore, this article proposes a consistency feature representation learning method for multimodal image registration, which involves mapping data into a common feature space to realize the accurate alignment of remote sensing images. First, a contrastive network with a spatial attention mechanism is driven to enhance the capability to highlight high-level features of images. Second, a positive sample augmentation strategy is implemented with contrastive loss, which helps the model learn the inherent features better, and imposes constraints on the sample similarity to optimize the feature projection. Finally, a multimodal image registration framework is proposed to enhance the stability of feature matching. The proposed framework achieves accurate feature extraction and consistency feature description for multimodal images, ensuring robustness against nonlinear radiometric differences. Experimental results demonstrate that the proposed method obtains more reliable registration results on the SEN1-2 dataset. Furthermore, the proposed algorithm achieves superior performance on data from other modalities, indicating strong generalization ability.

Index Terms—Contrastive learning, feature representation, multimodal image, remote sensing image registration.

I. INTRODUCTION

WITH the continuous advancement of global aerospace and aviation remote sensing technology, relying on a single sensor to acquire remote sensing image information no longer suffices to meet application requirements. In order to make up for the limited observation information acquired by a mono sensor and the inability to adapt to complex environmental challenges, multiple sensors are often employed in practical applications to fully leverage remote sensing information [1]. Image registration is the necessary prerequisite for cooperative processing of multimodal images, which aligns the spatial positions of remote sensing images captured by different sensors, enabling comprehensive analysis and utilization of the information. Multimodal remote sensing image registration is important for enhancing the spatial accuracy of images and facilitating spatiotemporal analysis [2], and it has found widespread application in various fields, such as meteorology, agriculture, and geology. However, differences in sensor imaging mechanisms [3] often result in salient geometric and grayscale discrepancies among remote sensing images, which exhibit different characteristics even on the same objective, making it challenging to acquire corresponding features.

To tackle this challenge, researchers have developed many methods in recent years. The handcrafted remote sensing image registration methods can be roughly divided into two categories: region-based methods and feature-based methods [4]. The region-based methods employ the global grayscale similarity information directly to alleviate local texture interference. CFOG [5] used the neighborhood information of channel features to carve the image structure features pixel by pixel, and conducted template matching in the frequency domain. SFOC [6] designed a coarse-to-fine registration system to cope with multimodal remote sensing images, extracting similar structural information. However, these methods are highly susceptible to geometric deformations and imaging noise in remote sensing imagery, and the computational cost is also high. Feature-based methods commonly extract significant features from two images, construct feature descriptors, and then estimate the geometric transformation matrix by comparing the similarity between the feature description. OS-SIFT algorithm [7] facilitated synthetic aperture radar (SAR) and optical image registration by detecting the consistency gradient with different operators according to the characteristics of images in various

Manuscript received 27 March 2024; revised 3 May 2024; accepted 21 May 2024. Date of publication 24 May 2024; date of current version 14 June 2024. This work was supported by the 2023 science and technology project “Research and application of multi communication system fusion networking technology for typical scenarios in power grids” under Grant 2023JBGS-11 of State Grid Jilin Electric Power Company. (Corresponding authors: Chen Chen; Ning Lv.)

Zhen Han is with the School of Telecommunication Engineering, Xidian University, Xi’an 710071, China (e-mail: hanzhen@stu.xidian.edu.cn).

Ning Lv is with the School of Electronic Engineering, Xidian University, Xi’an 710071, China, and also with the Xidian Hangzhou Institute of Technology, Hangzhou 311231, China (e-mail: nlv@mail.xidian.edu.cn).

Zhiyi Wang and Li Cong are with the State Grid Jilin Province Electric Power Company Limited Information Communication Company, Changchun 130000, China (e-mail: wwiizy@sina.com; congli8462@163.com).

Wei Han is with the Xi’an Beilin University-based Innovation Group Company Ltd., Xi’an 710000, China (e-mail: 847782433@qq.com).

Shaohua Wan is with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China (e-mail: shaohua.wan@ieee.org).

Chen Chen is with the School of Telecommunications Engineering, Xidian University, Xi’an 710071, China, also with the Key Laboratory of Industrial Internet of Things and Networked Control, Ministry of Education, Chongqing 400065, China, also with the Xidian Hangzhou Institute of Technology, Hangzhou 311231, China, and also with the Xidian Guangzhou Institute of Technology, Guangzhou 510555, China (e-mail: cc2000@mail.xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3405020

modalities. Nevertheless, due to the variations in intensity information among heterogeneous images, gradient-based operators often bring about erroneous matching. A representative approach is to design descriptors based on index maps that are insensitive to radiation variations. RIFT algorithm [8] combines the phase congruency feature with the maximum index map to construct the descriptor, which improves the robustness against the radiometric differences between multimodal images. Xiong et al. [9] proposed the adjacent self-similarity (ASS) feature, which leverages the minimum self-similarity map and the index map to calculate the distribution histogram, enabling fast heterologous image registration. But such methods tend to design descriptors for specific practical requirements, limiting their applicability and leading them prone to failure in complex scenarios.

Deep learning-based methods have gradually attracted attention due to their excellent performance in feature representation. Siamese networks [10] can measure similarity and meet the requirements of common feature learning. R²Net [11] designed a remote sensing framework [12] with global rectification and decoupled registration. DDFN [13] employed a Siamese network to learn dense pixel features, proposed a loss function based on the sum of squared differences, and enhanced computational efficiency. For the few-shot cases [14], PCNet [15] parsed the support mask into subregions to further form local descriptors. OSDescNet [16] proposed a local descriptor with an adaptive fusion convolution module to reduce SAR noise, and utilized DenseNet-CSP for characterization to match optical images with SAR images. Murugan et al. [17] introduced the deep reinforcement learning [18] to gain knowledge, registering multimodal images with a deep Q-network. These methods are capable to learn the intricate features of images and obtain a more robust representation. However, the supervised methods require a large amount of labeled sample data, which remains a rigorous challenge for Earth observation. With the development of self-supervised learning methods, novel feature representation models are proposed for image registration. Generative adversarial networks (GANs) can perform image-to-image translation, aligning multimodal images based on similar radiation feature information. KCG-GAN [19] input the K-means segmentation results of an image into the network to learn the spatial location, thereby enhancing the registration accuracy. To improve the recognition ability, RFM-GAN [20] applied neighborhood feature representation with two discriminators to train dissimilarity measurement networks.

However, methods based on GANs often excessively emphasize detailed information, rendering them subject to interference. By contrast, contrastive learning methods distinguish data in the feature space at the semantic level, exhibiting superior generalization capabilities. GLCNet [21] introduced a remote sensing semantic segmentation approach based on global style and local matching network, incorporating a matching contrastive loss to learn pixel-level information. ContraReg [22] achieved nonrigid multimodal image alignment by projecting the learned multiscale local patch features into the jointly learned interdomain embedding domain. Li et al. [23] proposed a template

matching method based on contrastive learning, which increases the matching number at finer details and performs intensive learning at the pixel level. Contrastive learning methods provide an effective solution for remote sensing tasks due to their excellent ability to learn the semantic embedding features without mass reference data.

Therefore, our work focuses on encoding consistency features with a contrastive learning network, and establishes a multimodal remote sensing image registration framework. The deep feature mining ability of the model is improved by integrating a spatial attention mechanism into the feature representation network, and the generalization ability is developed by the proposed sample design strategy and contrastive loss. The proposed image registration method fully exploits the features of multimodal images that exhibit significant radiation differences, achieving robust and reliable registration.

The main contributions are as follows.

- 1) A feature representation network with spatial attention modules is designed for embedding the consistency features between multimodal images. By performing efficient mapping to a common domain, the abstract semantic information is concerned and learned.
- 2) A positive sample augmentation strategy is presented to learn the intrinsic features of the data, and the robustness of model is developed under various conditions. The improved contrastive loss is proposed for the registration task to guide the distance constraints between samples.
- 3) A multimodal remote sensing image registration framework is proposed to effectively achieve feature extraction, description, and matching. Experiments demonstrate the superiority of the proposed work.

The rest of this article is organized as follows. The brief overview on contrastive learning methods are first introduced in Section II. Then the proposed model is explained in Section III in detail. Sections IV and V discuss the experimental results in various scenarios. Finally, Section VI concludes this article.

II. RELATED WORK

Annotating large datasets is laborious and costly, and this limitation hinders the applicability of deep learning-based methods in the remote sensing field. Self-supervised learning has become one of the effective ways to solve this problem. As a popular method of self-supervised learning, contrastive learning improves model performance by comparing the data with positive and negative samples in the feature space to learn the feature representation of the data. The core idea of contrastive learning is to narrow the similarity of data in the same class and to pull the encoding results of data in different classes as far as possible. SimCLR [24] set a paradigm for many subsequent contrastive learning research, and the related studies are mainly centered on three aspects: generation of sample construction, encoder mapping, and loss function constraints.

Proper selection of positive and negative samples can enhance model performance and avoid collapse. To reduce redundant

sampling, MoCo [25] proposed a momentum encoder to ensure real-time accuracy of the negative sample base, thus supporting larger sample sizes. SimCSE [26] introduced the idea of contrastive learning into sentence embedding, changing the dropout mask to generate positive samples, allowing the model to learn richer and fine-grained semantic information. In terms of encoder design, there are multifarious encoding networks can be applied in contrastive learning framework for feature extraction, such as Resnet-50 [27], U-Net [28], and Attention U-Net [29]. SCS-Co [30] applied a pretrained VGG-16 network as the style representation extractor. DINOv2 [31] used vision transformers as a backbone network and converted contrastive learning into a self-distillation learning task to learn robust visual features. CMAE [32] proposed a contrastive masked autoencoder consisting of two branches, the online branch and the target branch, aiming to improve the quality of representation and transfer learning performance. The core idea of the loss function in contrastive learning is to calculate the distance between representations, constraining the minimum and maximum distances between sample pairs. ProtoNCE [33] modified the similarity as the distance between samples and their clustering centers, learning both the clustering centers and the vector representations. SCL [34] introduced labels into contrastive learning, aiming to address the issue of neural networks being sensitive to noise labels.

These existing contrastive learning methods have reached the level of supervised learning on common visual tasks [35], such as classification and detection, and the powerful feature representation capability indicates that contrastive learning has a broader application prospect. Except for the lack of samples, the problem of multimodal remote sensing image registration is the representation difficulty between cross-modal features. Therefore, as contrastive learning shows increasingly superior characterization ability, it raises considerable attention in remote sensing. Considering the spatial properties of different modalities, Cha et al. [36] proposed a multimodal representation learning method for SAR semantic segmentation based on contrastive multiview coding. Pielawski et al. [37] proposed contrastive multimodal image representation for registration model, generating MI-related dense representations by maximizing the mutual information noise-contrastive estimation. DINO-MM [38] achieved SAR-optical image representation based on DINO network, which combined heterologous image pairs by connecting all channels to a uniform input, and enhances the data by randomly masking out channels of one modality. AdaSSIR [39] mined the potential features of the keypoints from two images. It treats each keypoint as an independent category, converting keypoints from one image to the other to construct training and test samples, ultimately achieving image alignment.

In summary, contrastive learning has shown excellent semantic embedding ability in many remote sensing tasks. By mapping remote sensing data into the same embedded space, significant radiation differences caused by various sensor imaging can be decreased. So this article proposes a contrastive learning representation method for consistency features in multimodal images, aiming to achieve remote sensing image registration based on

the feature representation. And the complexity of remote sensing image puts forward more challenging requirements for sample selection, so we design an effective sample augmentation strategy to improve the robustness of features.

III. METHOD

In this section, we introduce the proposed contrastive learning-based feature representation algorithm for multimodal image registration in detail, including a consistency feature representation network, a contrastive loss with sample augmentation, and a novel multimodal image registration framework.

A. Consistency Feature Representation Based on Spatial Attention

The proposed feature representation model is shown in Fig. 1, which consists of three stages: sample augmentation, consistency feature representation learning, and contrastive loss optimization. To achieve the alignment of cross-modal images, it is necessary to learn the mapping of consistency features, thus a suitable network needs to be designed for feature encoding. In this article, we choose U-Net [28] with a skip connection structure as the backbone net, so that the model can better retain and utilize feature information from different levels better, and effectively capture both local and global information in the image. Due to the complexity of remote sensing images, the network should have the ability to enhance the features of terrain. Therefore, we introduce the spatial attention mechanism to highlight the representation of the salient structural regions. The purpose of the spatial attention mechanism is to learn the weights and attention distributions of different locations, helping the model focus on the spatial structure in the input data, thereby improving the quality of the feature representations. It is worth mentioning that the dual path networks are trained alternately for two modalities, which is beneficial for extending this representation model to the training of multimodal images.

The feature embedding network is shown in Fig. 2. For the downsampling encoder, the features of the input sample are extracted through the convolution layer by layer. The input to each layer of the upsampling decoder is the aggregation of output from the previous layer and the skip connection structure. The spatial attention module is introduced related to the skip connection section, and the encoder output of each layer is weighted with attention and concatenated with the corresponding layer of the decoder. Finally, the feature map is recovered by 1×1 convolution layer of the same size as input sample. The inputs of the spatial attention module are convolved by $1 \times 1 \times 1$ respectively and concatenated, then the features are activated with a ReLU function. Convolved at $1 \times 1 \times 1$ and activated with a sigmoid function, the weight information is multiplied by the original input. Therefore, the spatial attention mechanism can be calculated as

$$a^l = \Omega_2 (\psi (\Omega_1 (y^l C_y + x C_x + b_x))) + b_\psi \quad (1)$$

$$y^{l_a} = y^l a^l \quad (2)$$

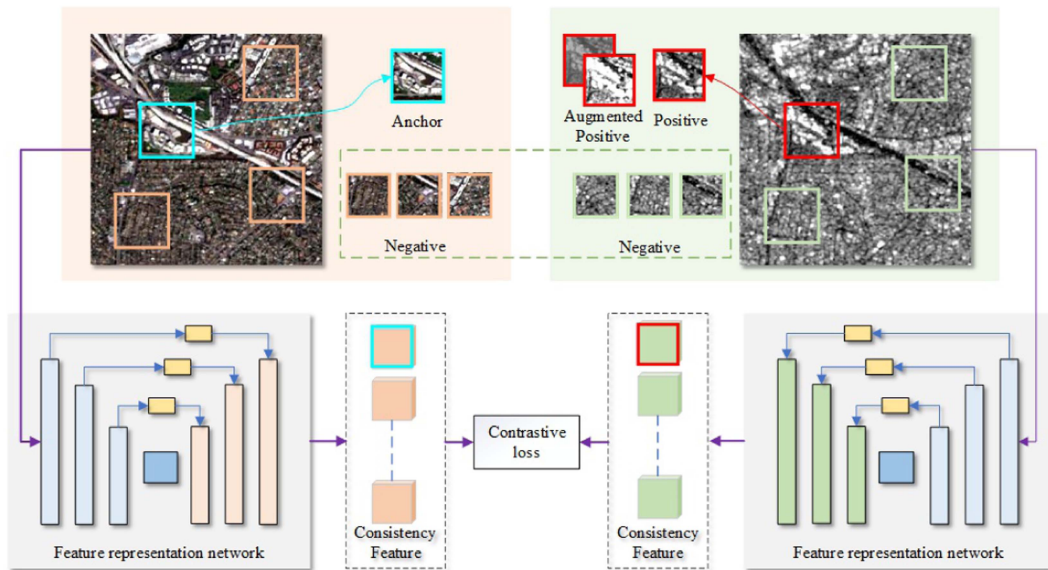


Fig. 1. Proposed consistency feature representation method.

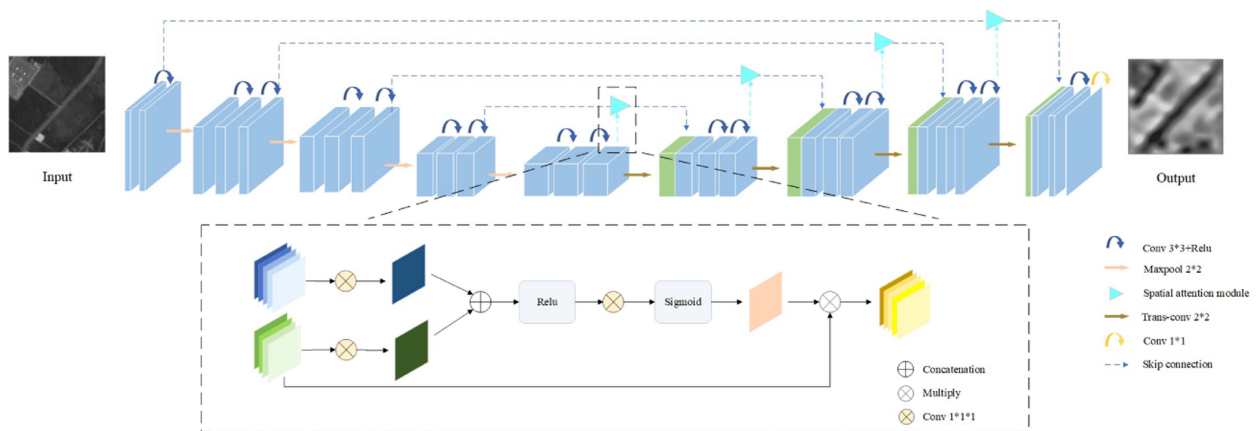


Fig. 2. Structure of feature encoding network.

where a^l represents the attention weight, ψ denotes the $1 \times 1 \times 1$ convolution, x and y^l represent feature map from the decoder and the previous layer of the encoder, which are operated with matrix C_x and C_y , respectively. Ω_1 and Ω_2 represent ReLU and sigmoid activation functions, respectively. b_x and b_ψ are bias parameters. Fig. 3 presents the heatmap of derived attention weights. It can be observed that the model can focus on structural areas, such as roads and shores. The combination of skip connection structure and spatial attention modules helps the contrastive network concentrate more on the important spatial position of the image. The consistency features are captured by the proposed contrastive learning network, providing reliable spatial information for subsequent matching.

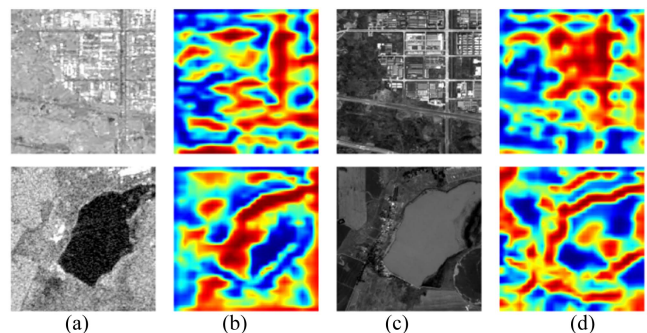


Fig. 3. Examples of spatial attention heatmap. (a) Image A. (b) Att. Map A. (c) Image B. (d) Att. Map B.

B. Contrastive Loss With Sample Augmentation

Contrastive learning methods aim to promote the feature representation of positive samples closer and that of negative samples more dispersed in the absence of labeled data.

Therefore, the sample construction is crucial for the performance of the model. The design of the sample is determined by the downstream task, and for the image registration, the learning goal is to map the consistency features of the same object in

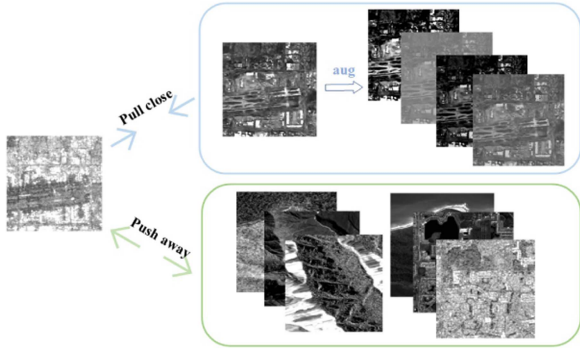


Fig. 4. Positive and negative samples.

different modalities to a common space. Therefore, assuming that a SAR image is taken as the anchor sample, then the optical image at the corresponding position is the positive sample, and other images in the batch are regarded as negative samples. In other words, for each batch with $2N$ samples, once a sample is anchored, there is one original positive sample and $2N - 2$ negative samples corresponding to it.

In order to increase the information of the positive sample with fewer training batches, the positive sample is augmented in this article for stronger feature representation ability. Random transformations of different contrast and brightness are conducted to increase the data diversity, and improve the capacity of model to process images in different illumination scenes. An example of sample composition is shown in Fig. 4. The blue box represents positive samples with the augmentation strategy that require attraction, and the green box represents negative samples from two modalities that need to be repelled. The augmentation strategy introduces changes to samples, which is conducive for the model to learn the features of objects under different imaging conditions.

The contrastive loss is also extended as the constraint on augmented positive samples. The information multiformity of the positive sample pairs can be improved by estimating the expectation of each augmentation of the positive sample instead of the original one, promoting the adaptive capacity of the model. The proposed contrastive loss is formulated as

$$L_{i,j} = -\log \frac{\exp[\hat{s}(z_i, z_j)/\tau]}{\sum_{k=1}^N \exp[s(z_i, z_k)/\tau]} \quad (3)$$

$$\hat{s}(z_i, z_j) = \frac{1}{n} \sum_{j=1}^n s(z_i, \bar{z}_j) \quad (4)$$

$$s(z_i, z_j) = (z_i \cdot z_j) / (\|z_i\| \|z_j\|). \quad (5)$$

In addition, the loss function adopts the similarity of negative samples instead of total samples to stretch the distance, enhancing sensitivity to discrepancies in negative samples. In this way, the coupling between positive and negative samples is avoided, and the optimization efficiency of the network is improved. Therefore, the loss function for multimodal remote sensing image registration is defined as (6). By constraining the consistency mapping between samples, the proposed loss

function can encourage the model to learn deep features

$$L_{\text{Reg}} = -\frac{1}{N} \sum_{i=1}^N \left(\log \frac{\exp[\hat{s}(z_i, z_j)/\tau]}{\sum_{k=1}^N \exp[s(z_i, z_k)/\tau]} \right). \quad (6)$$

Suppose that z_i , z_j , and z_k correspond to feature representation of the anchor sample, positive sample, and negative sample, respectively. τ is a temperature coefficient, and N means the number of negative samples. n denotes the total number of positive samples \bar{z}_j with augmentation. \hat{s} is calculated as (4), and s represents a similarity function defined by (5), to represent the distance of the feature encoding vector in the measure space.

C. Multimodal Image Registration Framework

On the basis of the above-mentioned representation network construction, this article proposes a framework for multimodal remote sensing image registration based on contrastive learning. The procedure is depicted in Fig. 5. The reference and sensed images are first preprocessed with specific modal characteristics. The processed images are then trained by contrastive network for consistency feature representation. It should be noted that the feature encoding will inevitably cause pixel deviation in the representation map, to ensure the precision of the registration results, the proposed framework simultaneously feeds the pre-processed images $I^{(1)}$ and $I^{(2)}$ into a candidate feature detection module, to provide a more accurate and complete candidate set. The design of the candidate detection module aims to provide an accurate spatial reference, and simple lightweight detection algorithms are able to meet the demand, such as oriented FAST and rotated BRIEF [40] selected in this article.

After generating the contrastive feature representation map $I_{\text{rep}}^{(\cdot)}$, we traverse the feature points obtained on the candidate image $I_{\text{can}}^{(\cdot)}$, and extract the corresponding neighborhood of key-points on the representation map. Then, the local descriptors are constructed based on the neighborhood of feature. Histogram of oriented gradients [41] is adopted in this article, which counts the gradient orientation histogram of the image area to form the feature description. Subsequently, the correspondences are matched with the consistency feature descriptors, and the spatial transformation parameters can be obtained after the outlier removal.

The proposed framework simplifies the issue of multimodal image registration by aligning images to the same modality, making full use of the strong nonlinear representation capabilities of deep learning networks as well as preserving matching accuracy. Moreover, this framework is adaptable to various modalities or scene inputs, making it suitable for a wide range of applications.

IV. EXPERIMENTS AND ANALYSES

In this section, the dataset and implementation details are first introduced, followed by an explanation of the evaluation metrics used in the experiments.

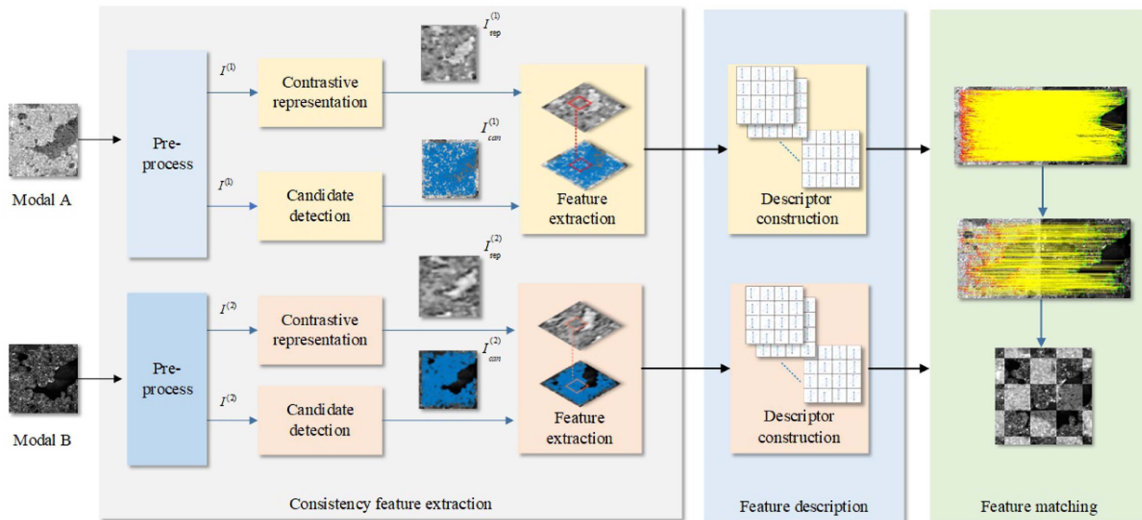


Fig. 5. Multimodal image registration framework.

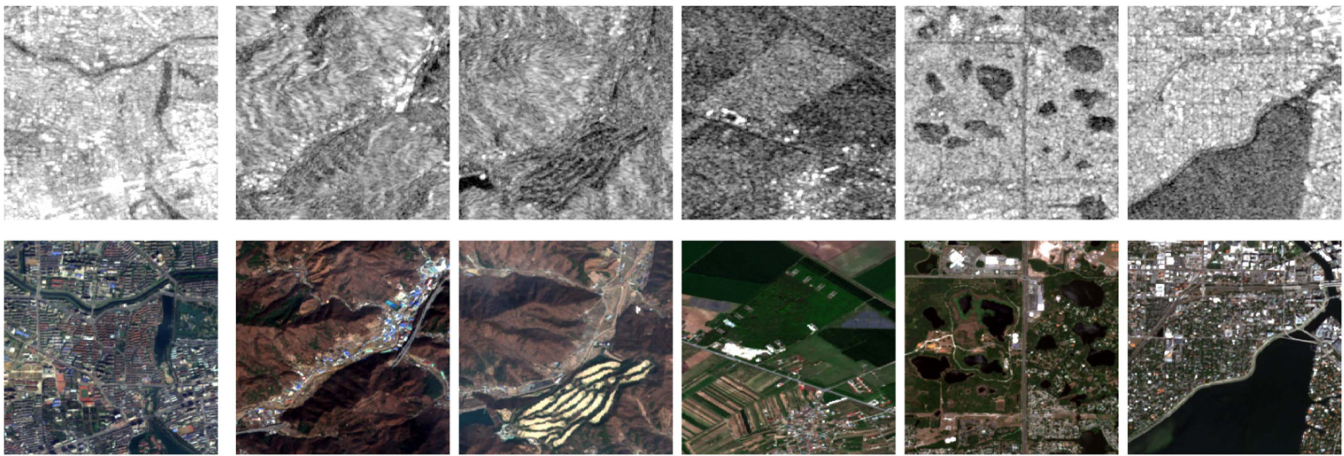


Fig. 6. Examples of SEN1-2 dataset.

A. Experimental Dataset and Settings

We employ the public SEN1-2 dataset¹ [42] for training and validation, which comprises paired SAR and optical images. This dataset is commonly employed in the multimodal remote sensing images registration, change detection, and fusion tasks. Examples of the SEN1-2 dataset are displayed in Fig. 6. The SAR images are collected by the Sentinel-1 satellite in IW mode, and the optical images are composed of the RGB bands from the Sentinel-2 satellite. Both SAR and optical images have a uniform sampling resolution of 10 m and are cropped to a size of 256×256 pixels. The data eliminates the interference of cloud cover to image information and covers diverse regions and terrains.

The trained model is tested on images from the SEN1-2 dataset and a multimodal image dataset to assess the registration performance and generalization ability. In the experiment, the SAR images are denoised with the nonlocal means

TABLE I
PARAMETERS SETTING

Parameter	Value
Epoch	48
Batch size	16
Learning rate	1e-3
Optimizer	Adam
Tau	0.1
Weight decay	1e-4

algorithm [43], and the optical images are grayed. The multimodal dataset consists of SAR, optical, infrared, depth, and map images captured by diverse sensors, covering different regions and time periods. With significant radiation differences, the test images pose a great challenge for registration methods.

The experiments are conducted on NVIDIA RTX 1080 Ti, and the training parameters of the model are shown in Table I. For feature point detection, the threshold number is set to 5000,

¹SEN1-2 dataset: <https://mediatum.ub.tum.de/1436631>

TABLE II
NCM RESULTS OF DIFFERENT METHODS

Method	1	2	3	4	5	6	7	8	9	10
RIFT	33	33	26	51	15	51	24	104	--	58
ASS	19	22	--	53	25	--	--	16	22	65
sRIFD	34	--	35	48	19	76	33	119	--	203
Pix2pix	--	53	--	64	--	--	118	--	55	59
Pix2pixHD	--	79	--	--	109	--	147	--	--	163
KCG-GAN	77	66	--	20	110	175	60	--	80	302
ReDFeat	--	9	14	--	--	21	13	40	7	11
Proposed	334	175	150	157	217	111	157	136	152	162

and the number of decomposition levels is 8. The patch size in descriptor construction is 128×128 pixels, and matching threshold is 100. To ensure the comparability of the results, each algorithm was tested with the optimal parameters given in the literature, and the parameters for outlier removal in the algorithm were kept consistent across all algorithms.

B. Evaluation Metric

The registration performance is evaluated by the number of correct matching (NCM) and root-mean-square error (RMSE). During the feature point matching process, candidate pairs with large errors are iteratively eliminated, and the remaining pairs that meet accuracy requirements are deemed as correct matches. A higher NCM indicates greater registration reliability, with the algorithm demonstrating increased stability.

The RMSE is used to evaluate the precision of the registration results, which is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (x_{2i} - x_{1i})^2 + (y_{2i} - y_{1i})^2}{n}} \quad (7)$$

where i represents the index of keypoint, (x_{1i}, y_{1i}) is the keypoint coordinates of the reference image, and (x_{2i}, y_{2i}) is the keypoint coordinates obtained after affine transformation of the moving image, and n denotes the number of matching pairs. The lower the RMSE, the higher the matching accuracy.

C. Experimental Results

The performance is compared with other remote sensing image registration methods and analyzed to prove the superiority of our method. Then, a multimodal dataset is applied to verify its generalization ability.

1) *Performance Comparison to Other Methods*: The performance of the proposed methods is quantitatively and qualitatively compared with some popular multimodal image registration methods, including RIFT [8], ASS [9], sRIFD [44], Pix2pix [45], pix2pixHD [46], KCG-GAN [19], and ReDFeat [47]. Ten pairs are picked at random, and Table II lists the NCM results for eight algorithms. In cases where a method failed to register due to insufficient matching pairs, we denoted it with "--". Bold numbers indicate the best result in the column.

The RIFT algorithm achieves stable performance on most data, but the NCM obtained is limited. Therefore, the image registration of the ninth pair is offset because the amount of correct matching pairs is not enough. sRIFD obtains more matching pairs than RIFT, but performs weakly on two pairs. The ASS,

pix2pix, pix2pixHD, and ReDFeat algorithms fail on three or more images, indicating the vulnerability of their performance. The KCG-GAN method scores the highest on two pairs but fails on the third pair and eighth pair, displaying relatively unstable performance across different scenes. For the proposed method, it achieves the best results on most images with NCM values consistently over a hundred. The RMSE results, as illustrated in Fig. 7, further confirm the superior performance of the proposed method. We set the value of invalid registration as ∞ in this figure for clarity. It is clear that the proposed method steadily ranks at the forefront of precision with a high success ratio.

Besides, we choose some representative terrains to analyze the performance of the model, including roads, water bodies, buildings, farmland, mountains, and airports. The visualization of matching results are displayed in Fig. 8. Most algorithms only excel in terrains with sharp edges, such as roads and water. The edges in the multimodal image provide consistent information for the registration task. However, for areas such as buildings and airports, whose geomorphic boundaries are ambiguous or features are heavily disturbed by strong scattering effects, algorithms probably fail. By comparison, the proposed method outperforms the other methods with outstanding robustness. Visually, matching pairs of the proposed method are numerous and uniformly distributed, achieving reliable registration results. With sufficient matching pairs on all terrains, features are registered with high accuracy.

The corresponding checkerboard mosaic maps of the proposed method are displayed in Fig. 9. The junction areas of red boxes in the figure are enlarged and shown on the right. It can be observed that the edges of the regions are perfectly overlapped. The results demonstrate that our method can register multimodal remote sensing images effectively.

We also enlarge the intersection of checkerboard maps with different methods for comparison. In Fig. 10, the orange, green, and cyan boxes outline the three regions in different positions with road connections. There are certain deviations on the results from other methods, manifesting as the discontinuity of the roads. The results of RIFT and KCG-GAN methods show significant displacements in the orange area. For the green area, the road is mismatched by ASS, sRIFD, pix2pix, and ReDFeat algorithms. And pix2pixHD algorithm makes an error in the cyan region. While for the results of the proposed method, the lines are aligned and connected perfectly in each area. The global alignments illustrate that the proposed method achieves stable and reliable matching, with superior performance to other methods.

Inference time is also an important aspect to evaluate algorithms. We compare the time consumption of four deep learning-based methods in the experiment. The three generative models cost about 40 ms for each image, while the ReDFeat and proposed models take 11.82 and 12.01 ms, respectively. Generally, the proposed method requires less inference time than generative methods, and the cost is comparable to the ReDFeat model.

2) *Generalization Ability*: In the above-mentioned experiments, the test data adopted is from the same source as the training data. A question worth verifying is whether the

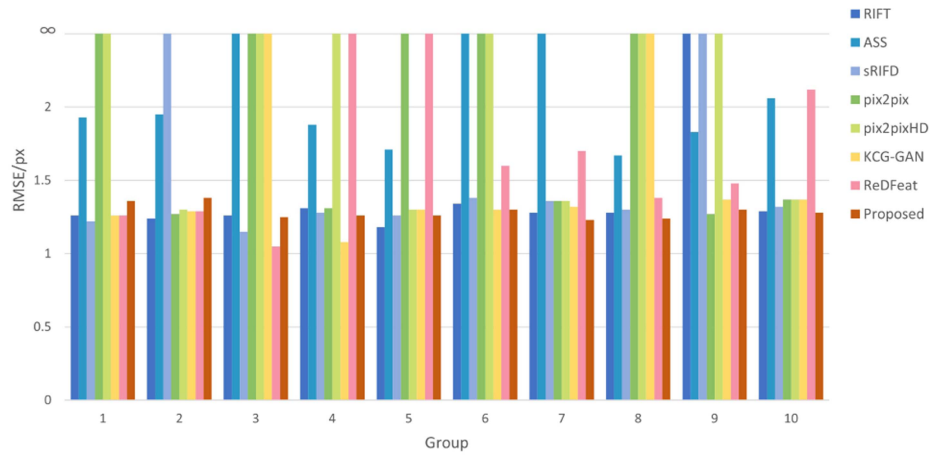


Fig. 7. RMSE results of different methods.

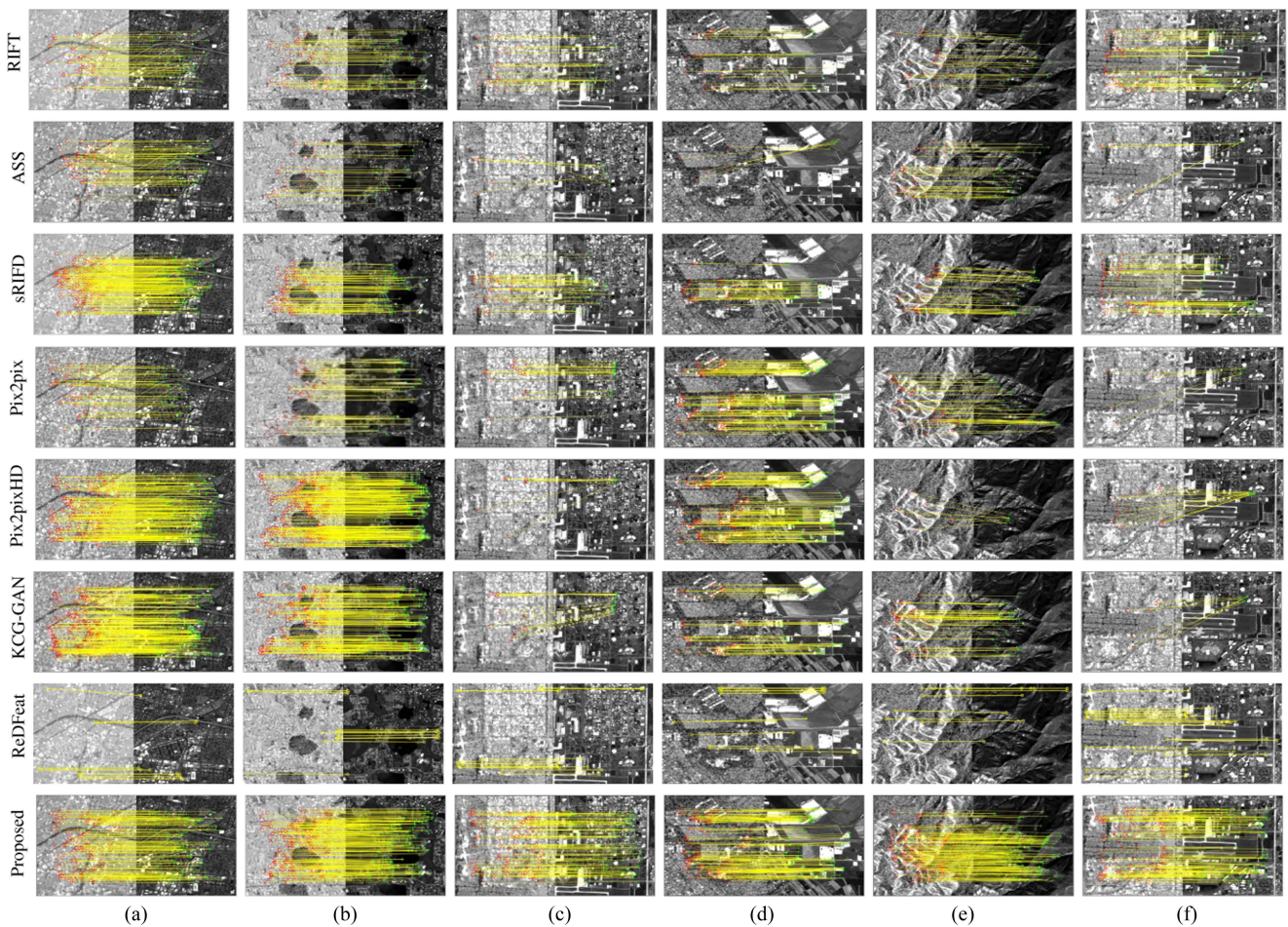


Fig. 8. Matching results of different methods. (a) Road. (b) Water. (c) Building. (d) Farmland. (e) Mountain. (f) Airport.

knowledge learned from the SEN1-2 dataset can be applied to other sensors. In order to explore this issue, we conducted registration experiments on data for five modalities, respectively. Fig. 11 showcases the matching results of the proposed algorithm. It can be seen that all image pairs are matched correctly. Fig. 12 compares the performance of other deep learning-based methods with the proposed model. The three generative

algorithms are invalid on more than two modalities, while the proposed model and ReDFeat model still work well with great accuracy. These two models achieve comparable accuracy on infrared-optical and day-night data, but the proposed method derives lower errors on other modalities. The results indicate that the proposed method is more adaptable to new data across multiple modalities.

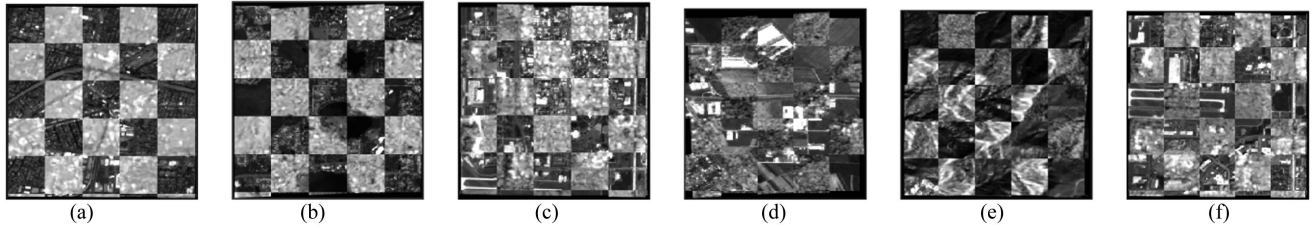


Fig. 9. Mosaic maps of the proposed method. (a) Road. (b) Water. (c) Building. (d) Farmland. (e) Mountain. (f) Airport.

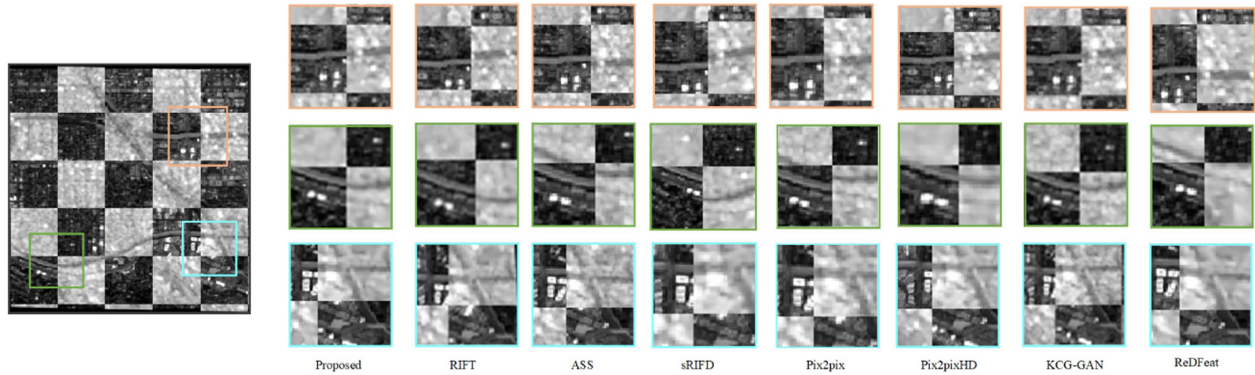


Fig. 10. Comparison of checkboard maps of different methods.

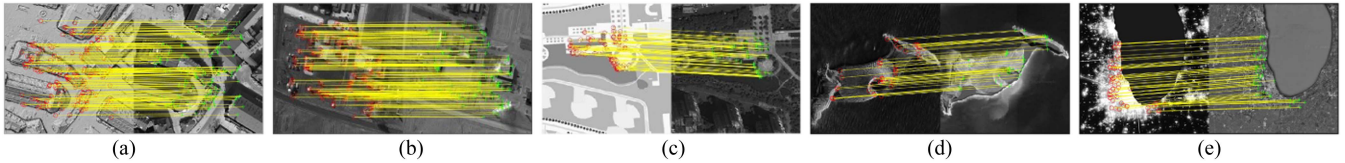


Fig. 11. Matching results on multimodal dataset. (a) Depth-optical. (b) Infrared-optical. (c) Map-optical. (d) Sar-optical. (e) Day-night.

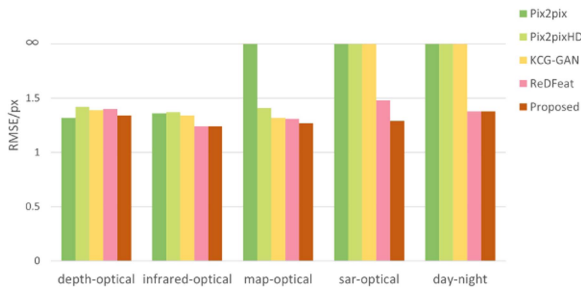


Fig. 12. Performance on multimodal dataset.

V. DISCUSSION

In this section, ablation experiments are conducted to better understand the necessity of each improvement in the proposed method. Then, we discuss the performance of the proposed method when facing challenging situations in remote sensing image registration.

A. Ablation Study

1) *Necessity of Spatial Attention:* For demonstrating the effect of the proposed feature representation network, we conducted experiments on the network with/without spatial

attention mechanism, respectively. Randomly selecting ten sets of test data while keeping the other parameters the same, the registration results are presented in Table III.

In terms of RMSE, the full model consistently reaches the lowest registration error for each pair. The inclusion of the attention module notably improved the NCM across most data. Especially in pair 3, the proposed representation network achieves successful matching with high accuracy and stability, whereas the network without spatial attention modules failed.

TABLE III
COMPARISON WITHOUT/WITH SPATIAL ATTENTION

Group	w/o attention		w/ attention	
	NCM	RMSE/px	NCM	RMSE/px
1	158	1.39	194	1.36
2	283	1.42	263	1.35
3	—	—	107	1.30
4	176	1.39	207	1.33
5	139	1.34	189	1.33
6	127	1.32	136	1.31
7	96	1.41	103	1.39
8	56	1.37	61	1.33
9	104	1.45	155	1.43
10	199	1.43	170	1.38

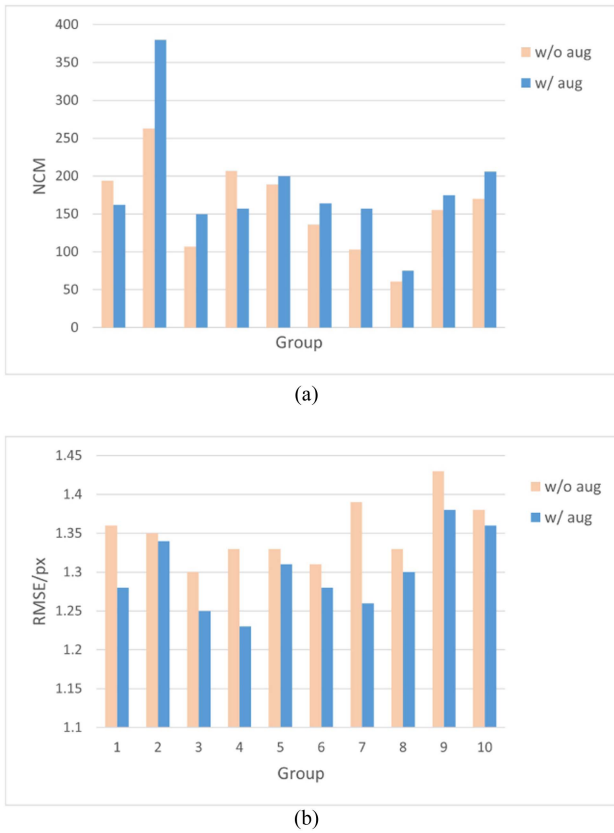


Fig. 13. Comparison without/with sample augmentation. (a) NCM. (b) RMSE.

On the whole, the proposed network outperforms the incomplete configurations.

2) *Necessity of Positive Sample Augmentation*: To ascertain the effectiveness of positive sample augmentation strategy on feature representation, the registration results with/without augmentation are compared in Fig. 13.

It can be seen that the errors by the augmented algorithm are further reduced, leading to a higher number of matching pairs on most data. These results indicate that the model with sample augmentation strategy effectively promotes registration robustness and accuracy, highlighting the efficacy of this approach in improving the performance of the feature representation network.

B. Challenging Issues

1) *Translation, Rotation, and Scale Variations*: The matching results of image pairs with translation, scale, and rotation variations are shown in Fig. 14. For the image pair with only translation differences, the proposed method can derive considerable matching pairs. Then the SAR image is taken as the reference image, and the optical image is rotated anticlockwise by 10° , as depicted in (b), our method still works well in this scene. In (c), the size of the optical image is reduced from $500\text{px} \times 500\text{px}$ to $400\text{px} \times 400\text{px}$, while the size of the SAR image remains unchanged. It can be found that the proposed method performs poorly on images with scale variations. Therefore,

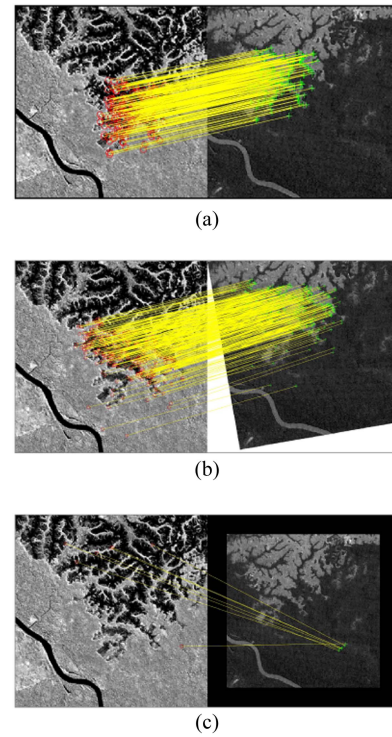


Fig. 14. Performance on translation, rotation, and scale variations. (a) Translation. (b) Rotation. (c) Scale.

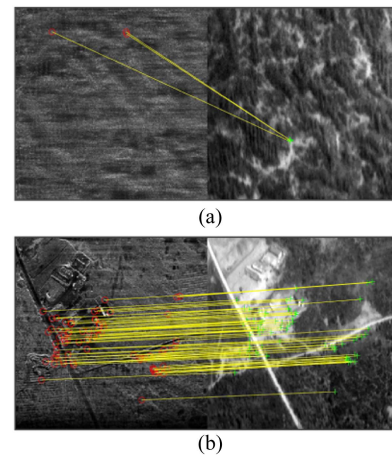


Fig. 15. Performance on unstructured and structure terrains. (a) Unstructured. (b) Structure.

the proposed method can handle images with translation and rotation differences, but shows unsatisfactory scale invariance.

2) *Homogeneous Terrains*: Differences in imaging modes result in various features on the same landform. The texture information is more prone to instability between multimodal images, while the structural information presents the superior performance. Therefore, the proposed method probably fails on terrains lacking apparent structure. As shown in Fig. 15, the SAR and optical images exhibit significantly different properties in the forest, rendering the algorithm invalid in (a). The area with roads and artificial structures provides references for corresponding feature detection, enabling our method to achieve

successful matching in (b). The vulnerability to texture-rich regions is a consistent issue in current registration algorithms, and we will focus on improving the robustness of texture information extraction in the future work.

VI. CONCLUSION

In this article, we propose a contrastive learning-based feature representation method for multimodal remote sensing image registration. A consistency feature representation network is proposed with a spatial attention module, significantly improving the capability of the network to focus on important spatial information. In addition, a sample augmentation strategy is designed for contrastive learning networks, enhancing the adaptability of the method in various scenarios with more diversity and variation. The extended contrastive loss can better constrain the model to efficiently learn latent features. Moreover, a multimodal remote sensing image registration framework is developed based on the proposed feature representation network to fully extract and describe image features. Accurate alignment facilitates the implementation of other downstream remote sensing tasks. The framework is efficient for images with significant modal variations, and is expected to be extended to multiple modalities. However, the proposed method is not an end-to-end process. Our future work will concentrate on learning spatial mapping directly with semantic embedding, thereby achieve more efficient registration.

REFERENCES

- [1] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.
- [2] J. Jiang et al., "Heterogeneous dynamic graph convolutional networks for enhanced spatiotemporal flood forecasting by remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3108–3122, 2024.
- [3] C. Chen et al., "A safe-distance control scheme to avoid new infection like COVID-19 virus using millimeter-wave radar," *IEEE Sensors J.*, vol. 24, no. 5, pp. 5687–5703, Mar. 2024.
- [4] 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.
- [5] 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.
- [6] 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. Photogrammetry Remote Sens.*, vol. 188, pp. 331–350, 2022.
- [7] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-Like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.
- [8] 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.
- [9] X. Xiong, G. Jin, Q. Xu, H. Zhang, L. Wang, and K. Wu, "Robust registration algorithm for optical and SAR images based on adjacent self-similarity feature," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
- [10] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8047–8057.
- [11] C. Lang, G. Cheng, B. Tu, and J. Han, "Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.
- [12] G. Cheng, C. Lang, and J. Han, "Holistic prototype activation for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4650–4666, Apr. 2023.
- [13] H. Zhang et al., "Optical and SAR image matching using pixelwise deep dense features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [14] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10669–10686, Sep. 2023.
- [15] C. Lang, J. Wang, G. Cheng, B. Tu, and J. Han, "Progressive parsing and commonality distillation for few-shot remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–10, 2023.
- [16] W. Xu, X. Yuan, Q. Hu, and J. Li, "SAR-optical feature matching: A large-scale patch dataset and a deep local descriptor," *Int. J. Appl. Earth Observation Geoinformation*, vol. 122, Aug. 2023, Art. no. 103433.
- [17] R. Murugan, M. Srivastava, and G. K. Rajput, "Robust image registration using deep reinforcement learning," in *Proc. Int. Conf. Optim. Comput. Wireless Commun.*, 2024, pp. 1–7.
- [18] T. Xiao, C. Chen, M. Dong, K. Ota, L. Liu, and S. Dustdar, "Multi-agent reinforcement learning-based trading decision-making in platooning-assisted vehicular networks," *IEEE/ACM Trans. Netw.*, pp. 1–16, early access, 2023, doi: [10.1109/TNET.2023.3342020](https://doi.org/10.1109/TNET.2023.3342020).
- [19] W.-L. Du, Y. Zhou, J. Zhao, and X. Tian, "K-means clustering guided generative adversarial networks for SAR-Optical image matching," *IEEE Access*, vol. 8, pp. 217554–217572, 2020.
- [20] Y. Liu, Y. Liu, and J. Wang, "RFM-GAN: Robust feature matching with GAN-Based neighborhood representation for agricultural remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [21] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [22] N. Dey, J. Schlemper, S. S. M. Salehi, B. Zhou, G. Gerig, and M. Sofka, "ContraReg: Contrastive learning of multi-modality unsupervised deformable image registration," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, pp. 6677–2022, 2022.
- [23] B. Li, L. Y. Wu, D. Liu, H. Chen, Y. Ye, and X. Xie, "Image template matching via dense and consistent contrastive learning," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2023, pp. 1319–1324.
- [24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. Mach. Learn.*, pp. 1597–1607, 2020.
- [25] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.
- [26] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6894–6910.
- [27] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15745–15753.
- [28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput.-Assist. Interv.—MICCAI 2015: 18th Int. Conf., Munich, Germany, Oct. 5–9, 2015, Proc., Part III 18*, pp. 234–241, 2015.
- [29] J. Schlemper et al., "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.
- [30] Y. Hang, B. Xia, W. Yang, and Q. Liao, "SCS-Co: Self-consistent style contrastive learning for image harmonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19678–19687.
- [31] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024.
- [32] Z. Huang et al., "Contrastive masked autoencoders are stronger vision learners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2506–2517, Apr. 2024.
- [33] J. Li, P. Zhou, C. Xiong, and S. C. H. Hoi, "Prototypical contrastive learning of unsupervised representations," in *Int. Conf. Learn. Representations*, 2021.
- [34] P. Khosla et al., "Supervised contrastive learning," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 18661–18673, 2020.
- [35] T. Xiao, C. Chen, Q. Pei, Z. Jiang, and S. Xu, "SFO: An adaptive task scheduling based on incentive fleet formation and metrizable resource orchestration for autonomous vehicle platooning," *IEEE Trans. Mobile Comput.*, early access, Nov. 28, 2023, doi: [10.1109/TMC.2023.3337246](https://doi.org/10.1109/TMC.2023.3337246).

- [36] K. Cha, J. Seo, and Y. Choi, "Contrastive multiview coding with electro-optics for SAR semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [37] N. Pielawski et al., "COMIR: Contrastive multimodal image representation for registration," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18433–18444.
- [38] Y. Wang, C. M. Albrecht, and X. X. Zhu, "Self-supervised vision transformers for joint SAR-optical representation learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 139–142.
- [39] S. Mao et al., "Adaptive self-supervised SAR image registration with modifications of alignment transformation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
- [40] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.
- [41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
- [42] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. IV-1, pp. 141–146, Sep. 2018.
- [43] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 60–65.
- [44] Y. Li, B. Li, G. Zhang, Z. Chen, and Z. Lu, "sRIFD: A shift rotation invariant feature descriptor for multi-sensor image matching," *Infrared Phys. Technol.*, vol. 135, Dec. 2023, Art. no. 104970.
- [45] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.
- [46] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.
- [47] Y. Deng and J. Ma, "ReDFeat: Recoupling detection and description for multimodal feature learning," *IEEE Trans. Image Process.*, vol. 32, pp. 591–602, 2023.



Zhen Han received the B.Eng. degree in electronic and information engineering from Yanshan University, Qinhuangdao, China, in 2020. She is currently working toward the Ph.D. degree in information and communication engineering with the School of Telecommunication Engineering, Xidian University, Xi'an, China.

Her research interests include image matching and image stitching.



Ning Lv received the B.Eng. degree in electrical engineering and automation from Xi'an Jiaotong University, Xi'an, China, in 2000, and the M.Sc. degree in circuit and system engineering from Xidian University, Xi'an, China, in 2004. He is currently working toward the Ph.D. degree in information and communication engineering with the School of Electronic Engineering, Xidian University.

His research interests include image registration and image stitching.



Zhiyi Wang received master's degree in business management from ChangChun University of Science and Technology, Changchun, China, in 2009.

He is a Senior Engineer of information technology. He has been engaged in power information and communication work for a long time. He has led the construction of multiple large-scale power information and communication projects and has made outstanding contributions in areas such as Beidou, 5G, and integrated communication.



Wei Han received the B.S. degree in electronic information engineering from the Rocket Force University of Engineering, Xi'an, China, in 2010.

He is deeply engaged in the commercialization of scientific and research findings, promoting technological-enabled economic development. He applies the innovative operation model practice to the incubation of high-tech enterprises, and familiar with the industry and specializes in attracting investment, introducing intelligence, service for high-tech innovation, to promote more than 100 startups have been

successfully incubated for the region. He takes technology and finance as a "New engine" to promote high-quality development of science and technology enterprises and led the team to cultivate the technology financial ecosystem, setup a fund group focusing on hard technology investment, and initiated the first equity investment fund in Shaanxi province focusing on the proof-of-concept phase of science and technology.



Li Cong received the Ph.D. degree in telecommunication from Xidian University, Xi'an, China, in 2011.

She currently a Deputy Director with State Grid Jilin Province Electric Power Company Limited Information Communication Company, Changchun, China. She is a Senior Engineer of the power engineering technology. She is also a Standing Director of the IEEE PES Wire Communications Subcommittee. She has authored or coauthored one book, and more than 30 scientific papers in international journals and conference proceedings. She has contributed to the

development of two copyrighted software systems and invented more than 20 patents.

Dr. Cong was the recipient of more than 70 honors and awards in the area of State Grid system.



Shaohua Wan (Senior Member, IEEE) received the Ph.D. degree in edge intelligence from the School of Computer, Wuhan University, Wuhan, China, in 2010.

He is currently a Full Professor with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China. From 2016 to 2017, he was a Visiting Professor with the Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany. He has authored more than 150 peer-reviewed

research papers and books, including more than 40 IEEE/ACM Transactions papers and many top conference papers in the fields of edge intelligence. His research focuses on deep learning for Internet of Things.



Chen Chen (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in telecommunication from Xidian University, Xi'an, China, in 2000, 2006, and 2008, respectively.

He is currently a Professor with the School of Telecommunications Engineering, Xidian University, where he is also the Director of the Xi'an Key Laboratory of Mobile Edge Computing and Security and the Intelligent Transportation Research Laboratory. He was a Visiting Professor with the Department of EECS, University of Tennessee, Knoxville, TN,

USA, and the Department of CS, University of California at Los Angeles, Los Angeles, CA, USA. He has authored or coauthored four books, and more than 150 scientific papers in international journals and conference proceedings. He has contributed to the development of 11 copyrighted software systems and invented more than 100 patents.

Dr. Chen is a Distinguished Member of China Computer Federation (CCF) and a Senior Member of the China Institute of Communications (CIC). He serves as the General Chair, the PC Chair, the Workshop Chair, or a TPC Member for a number of conferences.