JARS Journal of Applied Remote Sensing

RESEARCH PAPER

Synthetic aperture radar and optical image registration using local and global feature learning by modality-shared attention network

Xin Hu,[a] Yan Wu,[a,*] Zhikang Li,[a] Xiaoru Zhao,[a] Xingyu Liu,[a] and Ming Li[b]
[a]Xidian University, School of Electronic Engineering, Remote Sensing Image Processing and Fusion Group, Xi'an, China
[b]Xidian University, National Key Laboratory of Radar Signal Processing, Xi'an, China

ABSTRACT. The registration of synthetic aperture radar (SAR) and optical images is a meaningful but challenging multimodal task. Due to the large radiometric differences between SAR and optical images, it is difficult to obtain discriminative features only by mining local features in the traditional Siamese convolutional networks. We propose a modality-shared attention network (MSA-Net) that introduces nonlocal attention (NLA) to the partially shared two-stream network to jointly exploit local and global features. First, a modality-specific feature learning module is designed to efficiently extract shallow modality-specific features from SAR and optical images. Subsequently, a modality-shared feature learning (MShFL) module is designed to extract deep modality-shared features. The local feature extraction module and the NLA module in MShFL extract deep local and global features to enrich feature representations. Furthermore, a triplet loss function with a cross-modality similarity constraint is constructed to learn modality-shared feature representations, thereby reducing nonlinear radiometric differences between the two modalities. The MSA-Net is trained on a public SAR and optical dataset and tested on five pairs of SAR and optical images. In the registration results of five pairs of test SAR and optical images, the matching rate of the MSA-Net is 5% to 15% higher than that of other compared methods, and the matching errors of the matched inliers are on average reduced by about 0.28. Several ablation experiments verify the effectiveness of the partially shared network structure, the MShFL module, and the cross-modality similarity constraint.

 [DOI: 10.1117/1.JRS.17.036504]

Keywords: synthetic aperture radar and optical images; image registration; local feature extraction; nonlocal attention; modality-shared feature learning; cross-modality similarity constraint

Paper 220577G received Oct. 5, 2022; revised May 21, 2023; accepted Aug. 8, 2023; published Aug. 30, 2023.

1 Introduction

Image registration is the process of matching two or more images of the same scene captured at different times, from different viewing angles, and using different sensors. This process achieves geometric alignment between the reference image and the image to be registered. Image registration is widely used in various fields, such as image fusion,[1–3] image mosaic,[4] image stitching,[5,6] and multitemporal image change detection.[7,8] With the rapid development of multisensor technology, many different modes of images can be obtained from the same scene. Synthetic

*Address all correspondence to Yan Wu, ywu@mail.xidian.edu.cn

aperture radar (SAR) and optical images are the two main ways of observing the earth. Because these two images have complementary information, they are widely used in multimodal remote image tasks.[1–3,8–12] For these multimodal tasks to be accurate, the registration results of SAR and optical images are crucial.

However, nonlinear radiation differences between SAR and optical images and coherent speckles in SAR images make registration an unsolved and challenging research issue. Both traditional handcrafted methods[13–19] and deep learning-based ones have been used to address these issues.[20–28]

The handcrafted image registration approaches mainly include intensity-based and feature-based methods. Due to the poor similarity of intensity-based descriptors, the former is less effective in multimodal image registration. The latter can deal with nonlinear radiation differences more effectively by mining the structural features of multimodal images to obtain descriptors. As a result, feature-based approaches are increasingly being used in image registration. The gradient-based descriptor algorithms, such as scale-invariant feature transform (SIFT),[13] SAR-SIFT,[15] and OS-SIFT,[16] are applied to optical image registration, SAR image registration, and optical and SAR image registration, respectively. Nonetheless, gradient-based descriptors extract less valuable information from darker images. Compared with gradients, changes in light and radiation have no effect on phase congruency. The registration algorithms based on phase congruency, such as the histogram of phase congruency (HOPC),[17] phase congruency structural descriptor (PCSD),[18] and radiation-invariant feature transform (RIFT),[19] have been successfully employed in SAR and optical image registration. Recently, Fan et al.[29] proposed a new nonlinear diffusion-based Harris-Laplace detector and a new structural descriptor based on multiscale adaptive binning phase congruency that is more robust to geometric and radiometric differences, improving the number of matching points and reducing matching errors.

The handcrafted methods based on gradient and phase congruency mentioned above mine modality-shared features, i.e., structural features. The handcrafted methods use a fixed feature extraction pipeline to align a small number of images. They do not have the ability to handle the resolution, scale, and rotations that occur in the alignment of a large number of SAR and optical images. With its powerful data mining capabilities, deep learning has been widely used in various tasks involving remote sensing images. In recent years, many deep-learning registration methods have been proposed. For example, L2-Net,[20] HardNet,[21] and Siamese fully convolutional network (SFcNet)[24] based on Siamese convolutional network have been proposed for single- or multimodal image matching or registration. MatchosNet[26] and CNet[27] based on pseudo-Siamese convolutional networks are proposed for SAR and optical image matching and registration.

The application of the above-mentioned Siamese or pseudo-Siamese convolutional networks to multimodal image registration has two problems. First, neither the Siamese nor pseudo-Siamese networks take into account both nonlinear radiation differences and mining modality-shared features. Since it is a two-stream network with completely shared parameters, the Siamese network ignores the radiation differences between SAR and optical images. The pseudo-Siamese network is a two-stream network that does not share parameters at all, which makes it unable to fully exploit the modality-shared features. Second, the local features extracted by convolutional networks are easily affected by radiation differences and noise, resulting in low similarity of extracted features.

For the first problem mentioned above, we propose a partially shared dual-stream convolutional neural network (CNN) to mine shared features that help improve registration accuracy. For the second problem mentioned above, we insert nonlocal attention (NLA) into the two-stream CNNs to extract both local and global features. Local features are limited due to receptive fields, which make it difficult to extract similar features from multimodal images. Global features have a larger range of receptive areas and are less affected by modal differences than local features. Therefore, considering global correlation can improve the robustness of features to a certain extent.

In this paper, we propose a modality-shared attention network (MSA-Net) to construct modality-shared feature descriptors for SAR and optical image registration. The main contributions of our method are summarized as follows.

(1) In MSA-Net, we built a partially shared network structure for mining the modality-specific and modality-shared features of SAR and optical images to overcome modal differences.
(2) In MSA-Net, a modality-specific feature learning (MSpFL) module is designed to extract shallow modality-specific features, and a modality-shared feature learning (MShFL) module is designed to extract deep modality-shared features. Specifically, the MShFL module consists of three local feature extraction (LFE) modules and three NLA modules, which can extract both local and global features.
(3) In MSA-Net, a triplet loss function with a cross-modality similarity constraint is designed to make the MSA-Net less susceptible to modal changes. The triplet loss can encourage the MSA-Net to learn shared features between SAR and optical images.

The reminder of this paper is organized as follows: Section 2 briefly introduces the related work of this paper. Section 3 elaborates on the proposed registration algorithm and the architecture of MSA-Net in detail. In Sec. 4, the experimental comparisons and analysis are performed on a public SAR and optical dataset. In addition, the comparisons between five test SAR and optical image pairs are carried out. Finally, Sec. 5 summarizes this paper.

2 Related Work

2.1 Deep Learning Registration Methods

Deep learning-based methods learn data-driven deep features through Siamese and pseudo-Siamese networks. L2Net[20] and HardNet[21] are methods applied to optical image matching. The similarity between the two is that they use a Siamese network with completely shared parameters to learn local feature descriptors. The difference is that L2-Net uses all positive and negative samples in the batch to optimize the network, whereas HardNet uses hard negative samples in a batch to complete the network optimization. Due to the successful application of HardNet in optical image matching, some papers[22,24] used its improved Siamese convolutional network to extract deep features and realize multimodal image matching or registration. Bürgmann et al.[22] input SAR and optical image patches of larger size into a modified Siamese network based on HardNet to obtain sparse feature descriptors and use the L2 distance of feature descriptors to obtain a similarity metric. Zhang et al.[24] adopted the SFcNet to learn the similarity score between two different kinds of image patches.

However, the Siamese network ignores the radiometric differences between SAR and optical images and the scattered noise in SAR images. Hughes et al.[23,24] explained that pseudo-Siamese networks are better suited for multimodal image matching. MatchosNet[26] and CNet,[27] based on pseudo-Siamese network structure, have achieved average results in SAR and optical image registration. MatchosNet is template matching, which can obtain better results only when there is no large translation or rotation transformation between the two images to be registered. However, the pseudo-Siamese network lacks the interaction between different modalities in the process of learning features, so it is more difficult to explore the shared features between different modalities. In addition, pseudo-Siamese networks increase the network parameters, which makes convergence more difficult. Therefore, we design a partially shared network to mine similar features in multimodal images while considering the differences between modalities.

2.2 Nonlocal Attention

Due to the limited perceptual field of CNN, it is difficult to capture the global correlations of images. Thus it does not cope well with multiple transformations and nonlinear radiometric differences in multimodal image registration. Due to its sequence (image block) modeling structure, the transformer[30] based on self-attention mechanism is particularly good at capturing global interactions between token embeddings. ViT[31] has successfully applied the transformer[30] from natural language processing to computer vision.[32–34] It demonstrates that transformer also has a strong ability to model spatial correlations of images. The main reason for the success of the transformer is its core component, the self-attention module that captures global information. The self-attention module computes the response at a position in a sequence (e.g., a sentence) by focusing on all positions and taking their weighted average in the embedding space. In the

registration task, the two images to be registered may have scale, rotation, translation, and other transformations. Due to various transformations, the local information of an image changes. However, the relative position between two pixels in one image is constant, which is equivalent to the relative position of two words in a sentence in natural language. Mathematically, self-attention can be understood as a nonlocal averaging operation[35] that captures long-distance dependencies. To fully exploit the features, we use a nonlocal operation similar to the self-attention module in the transformer to extract global features and CNN to extract local features.

3 Proposed Method

The proposed algorithm framework in this paper is depicted in Fig. 1. First, we use the phase congruency maximum moments (MMPC)[19] to extract keypoints from SAR and optical images. Then we cut patches centered on the keypoints to generate SAR and optical image patches. These patches are converted into 8-bit grayscale images and then fed into MSA-Net to learn deep features. The matching points are obtained by the nearest-neighbor matching strategy. Finally, the final registration results are obtained after removing the outliers by random sample consensus (RANSAC).

As shown in Fig. 2, the proposed MSA-Net learns modality-shared features between the SAR and optical images to improve multimodal image registration performance. First, the MSpFL module is proposed to extract specific shallow features from SAR and optical images. Second, the MShFL module is used to extract deep local and global features shared between the two modalities. The MShFL module includes three LFE modules and three NLA modules. Third, a triplet loss function with a cross-modality similarity constraint is used to match multimodal features. The structure of MSA-Net is shown in Table 1.

Given a set of SAR and optical patch pairs, $p_i^s \subseteq p^s$, $p_i^o \subseteq p^o$, $i \in 1, 2, \ldots, N$, where N is the number of the patches. When $i = j$, $p_i^s = p_j^o$ are the corresponding SAR and optical image patches. Thus $p_i^s = p_j^o$ are the noncorresponding SAR and optical image patches, when $i \neq j$. The goal of the MSA-Net is to make the distances between the corresponding patch pairs closer and the distances between the noncorresponding patch pairs farther. First, $\{(p_i^s, p_i^o), i = 1, 2, \ldots, N\}$ are separately fed into the $\text{MSpFL}_{\text{SAR}}$ and $\text{MSpFL}_{\text{OPT}}$ modules to obtain N pairs of modality-specific feature maps $\{(f_i^s, f_i^o) \in R^{c \times h \times w}, i = 1, 2, \ldots, N\}$, where h and w represent the spatial height and width of the feature map, and c is the channel dimension of each feature map. f_i^s and f_i^o represent two i'th shallow modality-specific features. Second, the MShFL module extracts deep modality-shared features. Each MShFL module includes three LFE modules and three NLA modules. The LFE module is used to extract local features, and the NLA module is used to extract global features. Third, the feature maps obtained by MShFL are fed into a kernel size with 8×8 convolution to obtain the 128-dimensional deep feature vectors $\{(e_i^s, e_i^o) \in R^{128 \times 1}, i = 1, 2, \ldots, N\}$. Finally, the triplet loss function with a cross-

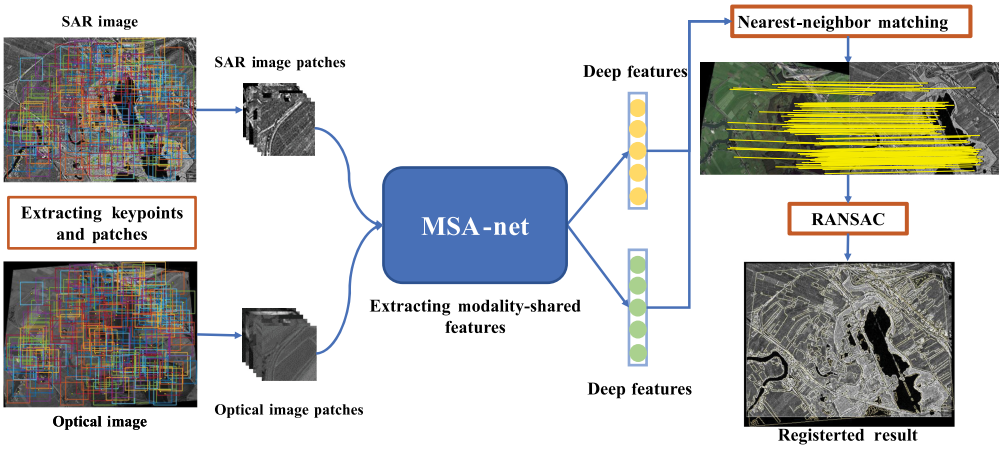

Fig. 1 Framework of the proposed registration algorithm.

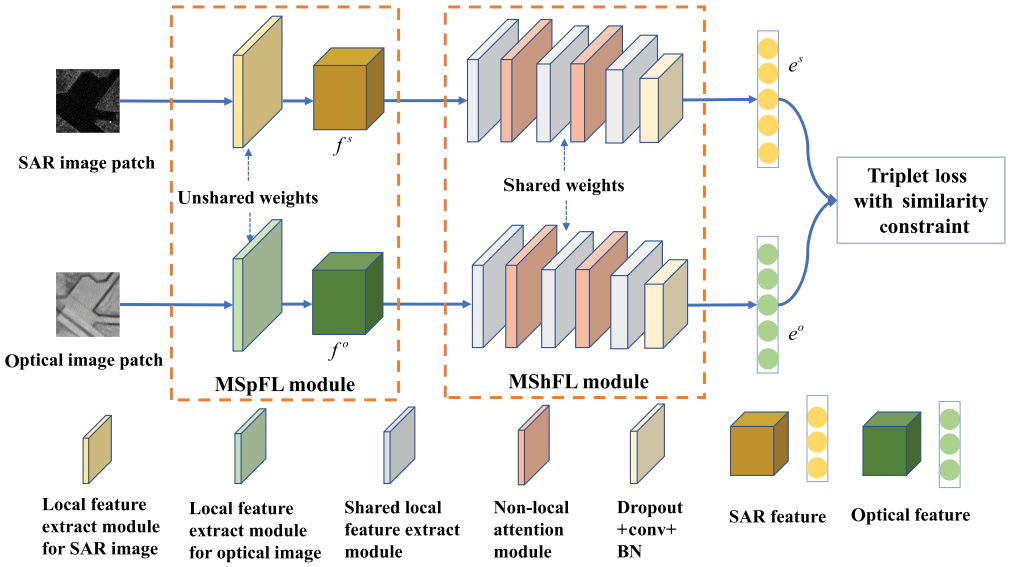

Fig. 2 Framework of the proposed MSA-Net.

Table 1 Detailed structure of the proposed MSA-Net.

Structure	Details	Output size
Input layer	Optical image patch	$1 \times 64 \times 64$
	SAR image patch	$1 \times 64 \times 64$
$\text{MSpFL}_{\text{SAR}}$	$\text{Conv}(32 \times 3 \times 3)/\text{BN}/\text{ReLU}$	$32 \times 32 \times 32$
	$\text{Conv}(32 \times 3 \times 3)/\text{BN}$	$32 \times 32 \times 32$
$\text{MSpFL}_{\text{OPT}}$	$\text{Conv}(32 \times 3 \times 3)/\text{BN}/\text{ReLU}$	$32 \times 32 \times 32$
	$\text{Conv}(32 \times 3 \times 3)/\text{BN}$	$32 \times 32 \times 32$
MShFL	LFE_1	$32 \times 32 \times 32$
	NLA_1	$32 \times 32 \times 32$
	LFE_2	$64 \times 16 \times 16$
	NLA_2	$64 \times 16 \times 16$
	LFE_3	$128 \times 8 \times 8$
	NLA_3	$128 \times 8 \times 8$
Last layer	$\text{Dropout}(0.1)/\text{Conv}(128 \times 8 \times 8)$	128×1

modality similarity constraint motivates the MSA-Net to learn more similar feature representations between corresponding SAR and optical image patches.

3.1 Modality-Specific Feature Learning Module

Considering the nonlinear radiation differences between SAR and optical images, we design MSpFL modules that do not share parameters to extract modality-specific features, such as texture, in the shallow network. There are two reasons to propose MSpFL modules. On the one hand, since the difference between modalities is not considered, it is difficult for the Siamese network to extract effective features that are favorable for multimodal image registration. On the other hand, if a pseudo-Siamese network is used, it will increase the number of network parameters and the difficulty of convergence.

Based on the above analysis, this paper proposes a partially shared feature extraction network. The MSpFL module is used to extract the specific shallow features of SAR and optical images. Suppose that the SAR and optical image patch pairs are $\{(p_i^s, p_i^o), i = 1, 2, \ldots, N\}$, p_i^s is the i'th SAR image patch, and p_i^o is the corresponding optical image patch of p_i^s. When they are fed into the MSpFL module, the modality-specific features learned from SAR and optical images can be expressed as follows:

$$f_i^s = \mathrm{MSpFL}_{\mathrm{SAR}}(p_i^s), \quad f_i^o = \mathrm{MSpFL}_{\mathrm{OPT}}(p_i^o), \tag{1}$$

where $\mathrm{MSpFL}_{\mathrm{SAR}}$ and $\mathrm{MSpFL}_{\mathrm{OPT}}$ represent the branches that extract modality-specific features of SAR and optical images, resepctively. f_i^s is the learned SAR feature of p_i^s, and f_i^o is the learned optical feature of p_i^o.

$\mathrm{MSpFL}_{\mathrm{SAR}}$ and $\mathrm{MSpFL}_{\mathrm{OPT}}$ are two networks that do not share parameters but have the same structure. As shown in Table 1, $\mathrm{MSpFL}_{\mathrm{SAR}}$ and $\mathrm{MSpFL}_{\mathrm{OPT}}$ contain two convolutional layers with a kernel size of 3×3 and a number of convolutional filters of 32. The strides of the two convolutional layers are 2 and 1, respectively. The first convolutional layer is followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) layer, and the second convolutional layer is followed by a BN layer.

3.2 Modality-Shared Feature Learning Module

After extracting the shallow specific features from the two modalities, we need to further extract the shared deep features. Here we design the MShFL module to learn the shared features from SAR and optical images, which includes three LFE modules and three NLA modules:

$$e_i^s = \mathrm{MShFL}(f_i^s), \quad e_i^o = \mathrm{MShFL}(f_i^o), \tag{2}$$

where e_i^s is the shared feature of f_i^s, e_i^o is the shared feature of f_i^o, and MShFL is the MShFL module.

3.2.1 *Local feature extract module*

We design three LFE modules in the MShFL module to extract local features shared between SAR and optical images. Each LFE module includes a residual structure of two convolution layers with a kernel size of 3×3. The strides of the two convolution layers are 2 and 1, respectively. The channels of the two convolutional layers in the three modules are 32, 64, and 128. Residual connections are added to prevent the network from overfitting. The learned modality-shared local features can be represented as follows:

$$f_{i,k+1}^s = \mathrm{LFE}_k(f_{i,k}^s), \quad f_{i,k+1}^o = \mathrm{LFE}_k(f_{i,k}^o), \tag{3}$$

where LFE_k represents k'th shared LFE module, $k = 1, 2, 3$. $f_{i,k}^s$ and $f_{i,k}^o$ represent the SAR and optical features input to LFE_k, respectively. $f_{i,k+1}^s$ and $f_{i,k+1}^o$ represent the SAR and optical features output from LFE_k, respectively.

3.2.2 *Nonlocal attention module*

Although CNN has good perception abilities for local regions, it lacks the modeling of global information. For the registration task, the information of the local area is important, but the global information can better describe the long-distance correlation and has good robustness to changes in scale, rotation, etc. For example, in two images with the same scene, the content in the images remains the same even after a slight rotation and translation transformation. Local information cannot capture unchanged information due to its limited perceptual range, while global information can still capture relevant information due to its large perceptual region. The NLA module aims at strengthening the features of the current pixel position by aggregating the information from all other positions in the feature map.

As shown in Fig. 3, $x \in R^{c \times h \times w}$ is transformed into $x' \in R^{c \times hw}$ by the reshape operation R. Let x' denotes the input feature map of the NLA module, where c is the number of channels, and hw is the number of positions in the feature map. We use three 1×1 convolutions that form three functions called α, g, and v, reducing the number of channels from c to c/r for the input x. In our

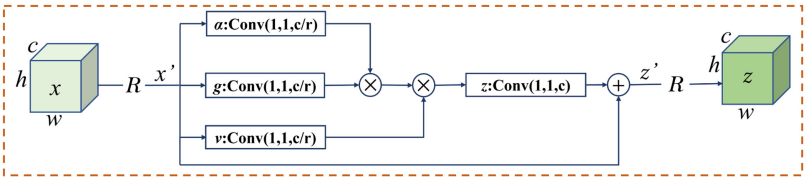

Fig. 3 NLA module extracts global features. "*R*" represents reshape operation, "⊗" denotes matrix multiplication, and "⊕" denotes element-wise sum.

experiments, we set the reduction factor r to 2. We use dot-product similarity[30] to define the normalized pairwise relationship between positions i and j in a feature map:

$$w_{ij} = \frac{f(x_i', x_j')}{C(x)} = \frac{\langle W_\alpha x', W_g x_j' \rangle}{hw}, \tag{4}$$

where $w_{ij} \in R^{hw \times hw}$ is a normalized similarity matrix, $f(x_i', x_j')$ is the relationship between positions i and j, W_α and W_g are the weights of two convolution functions α and g. $C(x)$ is a normalization factor, in this case $C(x) = hw$.

Then the similarity matrix w_{ij} is multiplied by the value matrix W_v to obtain the output matrix:

$$y_i = \sum_{j=1}^{hw} w_{ij}(W_v \cdot x_j'), \tag{5}$$

where y_i is the attention value of position i, and W_v is the weight of the convolution function v.

The output of the NLA module is defined as

$$z_i' = W_z y_i + x_i', \tag{6}$$

where W_z is a 1×1 convolution with the channel number of c, "+" in the formula indicates residual connection, z_i' represents the output of the NLA module at position i. $z' \in R^{c \times hw}$ is transformed into $z \in R^{c \times h \times w}$ by the reshape operation R. The input feature dimension of the NLA module is the same as the output feature dimension, which allows us to insert it anywhere in the network.

The NLA module can be regarded as a global information construction module. The specific operation is to obtain the global contextual features of the current pixel by the weighted average of all positions.

3.3 Triplet Loss with Similarity Constraint

In order to learn modality-shared features between the SAR and the corresponding optical patch pairs, the triple loss function is used to minimize the distances of the features of corresponding patch pairs and maximize the distances of the features of noncorresponding patch pairs. The margin m is a hyperparameter that aims to keep a certain distance between the corresponding and noncorresponding features. The triple loss L_t is calculated as follows:

$$L_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \max(d(e_i^s, e_i^o) - d(e_i^s, e_j^o) + m, 0), \tag{7}$$

where $d(e_i^s, e_i^o) = \sqrt{2 - 2e_i^s e_i^o}$ represents the L2 pairwise distance of the corresponding features, $d(e_i^s, e_j^o) = \sqrt{2 - 2e_i^s e_j^o}$, $i \neq j$, represents the L2 pairwise distance of the noncorresponding features, and N_t represents the number of samples in a batch, $m = 1$.

Negative sampling is usually required when calculating the triple loss function, which usually includes random sampling in all samples[20] and hard negative sample sampling in batches.[21] We adopt the latter to improve the optimization performance of the MSA-Net. Under hard negative sample mining, the triple loss is rewritten as

$$L_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\max(d(e_i^s, e_i^o) - d(e_i^s, e_{\text{hard}}^o) + m, 0), \tag{8}$$

where e_{hard}^o is the hard negative sample in the minibatch of e_i^s.

In order to reduce the modal differences between corresponding patch pairs of SAR and optical images, it is necessary to add similarity constraints between them. In this way, the network can learn more compact feature representations between corresponding patches. The cross-modality similarity constraint loss function L_s is defined as

$$L_s = \frac{1}{N_t}\sum_{i=1}^{N_t}\log(1 + exp(d(e_i^s, e_i^o))), \tag{9}$$

where log and exp represent a logarithmic and exponential function, respectively.

Therefore, the overall loss function is expressed as

$$L = L_t + \lambda L_s, \tag{10}$$

where λ is the weighting coefficient of the loss function.

4 Experiments and Discussion

To validate the performance of the proposed MSA-Net, the experiments are performed on a public SAR and optical image dataset SEN1-2[36] and five pairs of SAR and optical images. We compare our method with three handcrafted and three deep learning methods, including OS-SIFT,[16] PCSD,[18] RIFT,[19] HardNet,[21] MatchosNet,[26] and CNet.[27] Section 4.1 describes a data description. Section 4.2 introduces implementation details for the training stage, testing stage, and matching stage. Section 4.3 describes the evaluation criteria used to evaluate algorithm performance. Section 4.4 presents the results and analysis of the experiments. Section 4.4.1 analyzes the influence of three weight-sharing strategies on the model's performance. In Secs. 4.4.2 and 4.4.3, we verify the effectiveness of the NLA module and the cross-modality similarity constraint. In Sec. 4.4.4, we conduct a parameter sensitivity analysis to analyze the influence of parameter λ on the matching accuracy. In Sec. 4.4.5, we analyze the influence of fine tuning on registration results. In Sec. 4.4.6, we compare the registration results of seven algorithms on five pairs of SAR and optical images, including qualitative analysis and quantitative evaluation. The computational times of seven methods are shown in Sec. 4.4.7.

4.1 Data Description

The proposed MSA-Net is trained on SEN1-2,[36] which is a public SAR and optical image dataset. The SEN1-2 includes 282,384 co-registered SAR and optical images, each with a size of 256×256 and a resolution of 10 m. The SEN1-2 dataset contains four folders: ROIs1158_spring, ROIs1868_summer, ROIs1970_fall, and ROIs2017_winter. In order to ensure that the training samples and test samples do not overlap, we select the data in ROIs1868_summer as the training samples and the data in ROIs1970_fall as the test samples and filter out fuzzy images without obvious ground objects, as shown in Fig. 4. We uniformly sample the images selected from SEN1-2 with a stride of 32 to obtain a patch of size 64×64. After rotation and zooming, we obtain 70,193 training sample pairs and 11,147 validation sample pairs. The information about five pairs of SAR and optical images used for testing is shown in Table 2.

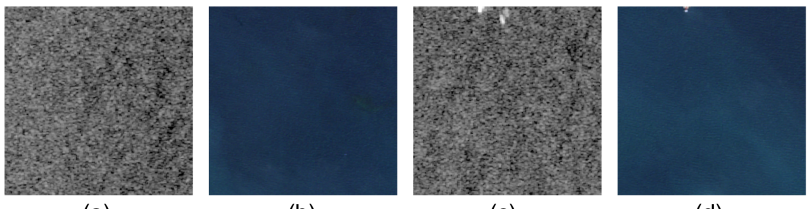

Fig. 4 Examples of SAR and optical images with no obvious objects filtered out from SEN1-2. A pair of corresponding (a), (c) SAR and (b), (d) optical images.

Table 2 Information about five pairs of SAR and optical images used for testing.

Pair	Modality	Sensor	Size (m)	Resolution (m)
1	SAR	TerraSAR-X	550×512	*
	Optical	Google Earth	550×512	*
1	SAR	TerraSAR-X	550×512	*
	Optical	Google Earth	550×512	*
1	SAR	TerraSAR-X	550×512	*
	Optical	Google Earth	550×512	*
4	SAR	GaoFen-3	512×512	1
	Optical	Google Earth	512×512	1
4	SAR	GaoFen-3	512×512	1
	Optical	Google Earth	512×512	1

The "*" in Table 2 represents unknown information. The SAR and optical images in pairs 1 to 3 are from TerraSAR-X and Google Earth, and the SAR and optical images in pairs 4 to 5 are from GaoFen-3 and Google Earth. In this paper, all experiments are conducted on a desktop computer with Windows 10, an RTX 3060 GPU, and 24 GB RAM. The MSA-Net proposed in this paper is built using the PyTorch[37] deep learning framework.

4.2 Implementation Details

4.2.1 *Training stage*

We first train our network from scratch on 70,193 pairs of training samples in SEN1-2 and validate it on 11,147 pairs of validation samples. The batch size is 300. The SGD optimizer with an initial learning rate of 0.1 is adopted. The training epoch is 20. Next, the network is fine-tuned with the five pairs of images before testing. Specifically, we manually select four points on each SAR and optical image and use HOPC[17] to get registered images. According to the stride of 16, we cut the registered image into patches with a size of 64×64. Some blank patches without content are removed. The reason for using a stride of 16 instead of 32 for the test image is that the number of test data is small, and reducing the stride can generate more patches. Meanwhile, reducing the stride results in more overlap between adjacent patches, leading to increased similarity between their contents. This forces the network to learn more discriminative features, so that the feature distance between different patches is as large as possible. These patches are amplified using the same data enhancement techniques as SEN1-2. The purpose of using overlapping blocks instead of nonoverlapping blocks is to make MSA-Net learn more discriminative features from certain similar samples. When one pair of images is tested, the remaining pairs of images are used to fine-tune the network. The parameters m and λ in the proposed loss function are 1 and 0.1, respectively.

4.2.2 *Testing stage*

The feature points are extracted using the MMPC.[19] Next, a local patch with a size of 64×64 is cropped around each feature point. Then the local patch is fed into the network to obtain the 128-dimensional feature vector, that is, the proposed modality-shared feature descriptor.

4.2.3 *Matching stage*

We match the modality-shared feature descriptors corresponding to each feature point through nearest-neighbor matching to obtain the initial matching points. The affine transformation parameters from the image to be registered (optical image) to the reference image (SAR image) are

calculated by using the initial matching points. Finally, outliers are deleted by RANSAC to obtain the final registration result.

4.3 Evaluation Criteria

We use two metrics to evaluate the matching performance, including FPR95 and accuracy. The FPR95 refers to the false positive rate in which the true positive rate equals 95%, the accuracy refers to the probability that the network correctly predicts the matching label of two patches. Therefore, the smaller the FPR95 is, the higher the accuracy is, and the better the matching performance will be.

We quantitatively evaluate different registration methods using the number of correct matches (NCM), the ratio of correct matches (RCM), and the root-mean-square error (RMSE) as evaluation metrics:

$$\mathrm{RCM} = \frac{\mathrm{NCM}}{\mathrm{NM}}, \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{NCM}} \sum_{i=1}^{\mathrm{NCM}} [(m_i' - m_i)^2 + (n_i' - n_i)^2]}, \tag{12}$$

where NM is the total of matches, (m_i', n_i') is the coordinate of the image to be registered after the transformation matrix, and (m_i, n_i) is the coordinate of the reference image.

4.4 Experimental Results and Analysis

4.4.1 *Influence of three weight-sharing strategies*

As the analysis in Sec. 2.1 shows, although the Siamese network with completely shared weights can mine shared features, it may ignore modal differences. The pseudo-Siamese network with completely unshared weights can extract modality-shared features, but it does not take modality-shared features into account. Therefore, this paper discusses three weight-sharing strategies, including completely unshared, completely shared, and partially shared weights. The experiment is carried out on the SEN1-2 dataset, and the experimental results are shown in Table 3. The "Dist_np" in Table 3 represents the L2 pairwise distance between positive samples and negative samples. The larger the distance is, the better the network's ability to distinguish between positive and negative samples will be.

From Table 3, we can see that the network structure with completely shared weights achieves the lowest matching accuracy. The FPR95 is the highest and Dist_np is the lowest among the three weight-sharing strategies, which indicates that the network with completely shared weights does not cope well with differences of multimodal images. The third and fourth rows in Table 3 are the experimental results of the completely unshared weight network and the partially shared weight network, respectively. The latter achieves a higher matching accuracy, about 1.1% higher than the former, and the FPR95 of the latter is 0.1769, about 0.06 lower than the former. The Dist_np of the partially shared weight network is 0.2299, about 0.03 higher than the completely unshared weight network. Therefore, we propose a dual-stream network with partially shared weights can better mine the modality-shared features of SAR and optical images and improve matching accuracy.

Table 3 Matching results of different weight-sharing strategies.

Weight-sharing strategy	Accuracy	FPR95	Dist_np
Completely unshared	0.9539	0.2356	0.2035
Completely shared	0.9346	0.2497	0.1943
Partially shared	0.9669	0.1769	0.2299

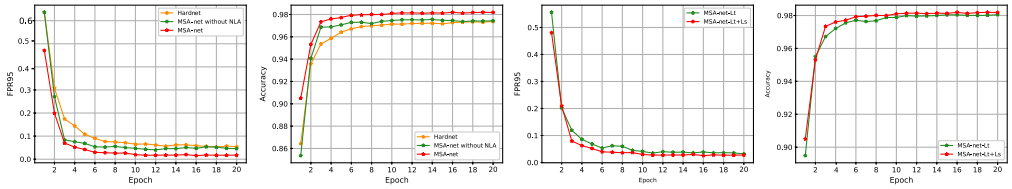

Fig. 5 Changing trend of (a) FPR95 and (b) accuracy with epoch under different models. Changing trend of (c) FPR95 and (d) accuracy with epoch under different loss functions.

4.4.2 *Effectiveness of the NLA module*

Figures 5(a) and 5(b) show the changing trend of FPR95 and accuracy under three network structures, which are HardNet, MSA-Net without NLA, and the proposed MSA-Net. MSA-Net without NLA represents the network structure after removing the NLA module from MSA-Net. It can be seen that the matching accuracy of MSA-Net is the highest and the FPR95 of MSA-Net is the lowest among the three models. The NLA module mines global features with a larger receptive field. This is so that MSA-Net can cope with radiation differences to a certain extent, thus improving matching performance. We can also see that MSA-Net without NLA achieves higher matching accuracy and lower FPR95 in fewer epochs compared to HardNet. The difference between the two models lies in whether the weights of the feature network are completely shared and whether residual connections are used. HardNet learns features from SAR and optical images using a network with completely shared weights and no residual connections, both of which make HardNet less effective than MSA-Net without NLA in handling multimodal data.

4.4.3 *Influence of loss function*

Figures 5(c) and 5(d) show the FPR95 and accuracy change curves of different loss functions. After adding cross-modality similarity constraints to the loss function, the matching accuracy is further improved and the FPR95 is further reduced. It proves that the combination of triplet loss and cross-modality similarity constraint loss allows the MSA-Net to further explore the similarity between SAR and optical images.

4.4.4 *Parameter sensitivity analysis*

In this section, we analyze the influence of the weight coefficient λ on matching accuracy. To observe the changing trend of accuracy in the test set of SEN1-2, we gradually increase the value of λ from 0 to 1 with an interval of 0.1. Figure 6 reports the average of three experiments. When

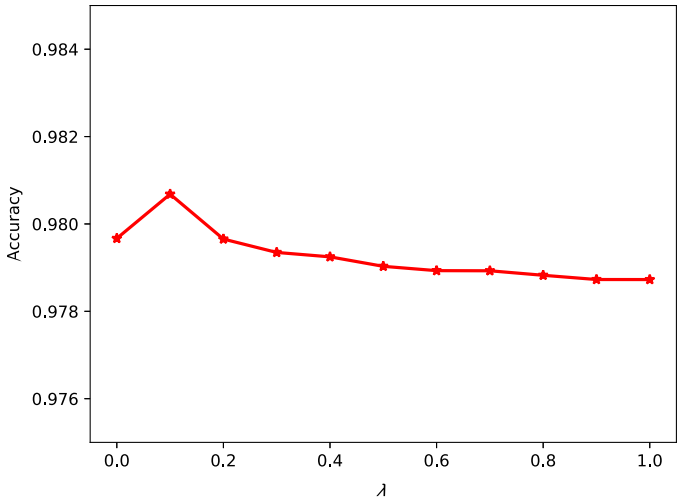

Fig. 6 Changing trend of accuracy with λ.

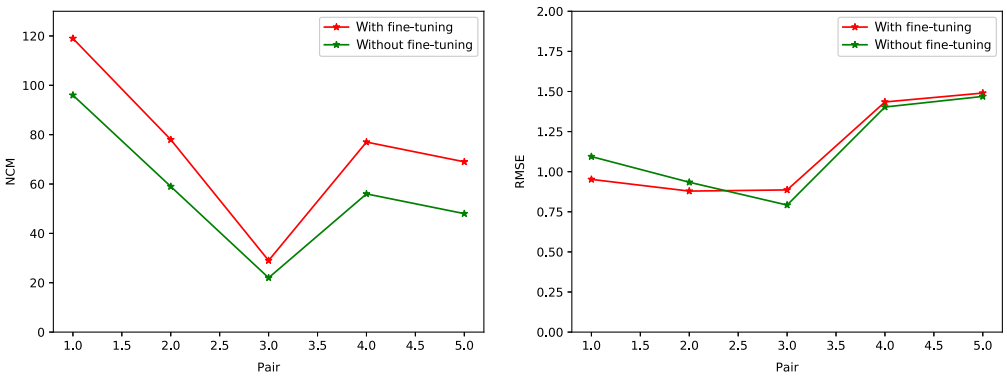

Fig. 7 Influence of fine-tuning on the registration results of five test image pairs. (a) NCM obtained by MSA-Net in five pairs of test images with or without fine-tuning. (b) RMSE obtained by MSA-Net in five pairs of test images with or without fine-tuning.

the value of λ is between 0 and 1, the value of accuracy is between 0.978 and 0.982. It indicates that adding cross-modality similarity constraints to the loss function will not affect the stability of the network. When λ is 0.1, accuracy reaches its maximum value. It can be seen that a small weight for L_s can make the network learn well. When λ increases to 1, accuracy gradually decreases. We conclude that giving a larger λ to cross-modal similarity constraints is not conducive to learning features of multimodal images. We should not let the network pay too much attention to learning similar features without considering that the multimodal data itself is different. So it is necessary to obtain an appropriate weight coefficient through experiments.

4.4.5 *Influence of fine-tuning on registration results*

In Fig. 7, we show the influence of fine-tuning on the registration results of five test images, including NCM and RMSE. Figure 7(a) shows that the NCM increases significantly after fine-tuning, especially in pairs 1, 2, 4, and 5. As shown in Fig. 7(b), the RMSEs of pairs 1 and 2 decrease after fine-tuning. The RMSE of pair 3 is higher after fine-tuning. The RMSEs of pairs 4 and 5 do not change significantly before and after fine-tuning. It can be seen that registration results can be improved to some extent through fine-tuning. Even without fine-tuning, relatively good results can be obtained, indicating that the model trained with SEN1-2 data has transfer ability. Considering that different data are distributed differently, fine-tuning can improve the adaptability of the network and obtain better registration results.

4.4.6 *Comparison with the other registration methods*

We select five pairs of SAR and optical images for evaluating seven registration methods. In pairs 1 to 3, there are not only radiation differences between SAR and optical images but also rotation and translation transformations. Pairs 4 and 5 have obvious radiation differences, noise, slight translation transformations, and no rotation transformation. Figures 8(a)–8(g)–12(a)–12(g) show the matching points obtained by seven methods in five pairs of SAR and optical images. Figures 8(h)–12(h) show the registration results of MSA-Net. We superimpose the outlines (as shown by the yellow curve) of transformed optical images on the original SAR images and use several red rectangles to highlight the registration results for certain regions.

From Figs. 8(a)–8(c)–12(a)–12(c), we can see that the three handcrafted algorithms have achieved relatively good registration results for the five pairs of images. RIFT obtains more matching points than OS-SIFT and PCSD thanks to its use of radiation-invariant phase congruency to extract keypoints and construct descriptors. OS-SIFT has the fewest matching points among the three handcrafted algorithms because the descriptor based on gradient is worse than the one based on phase congruency in dealing with radiation differences. PCSD matches descriptors in the local window, resulting in fewer matching points obtained.

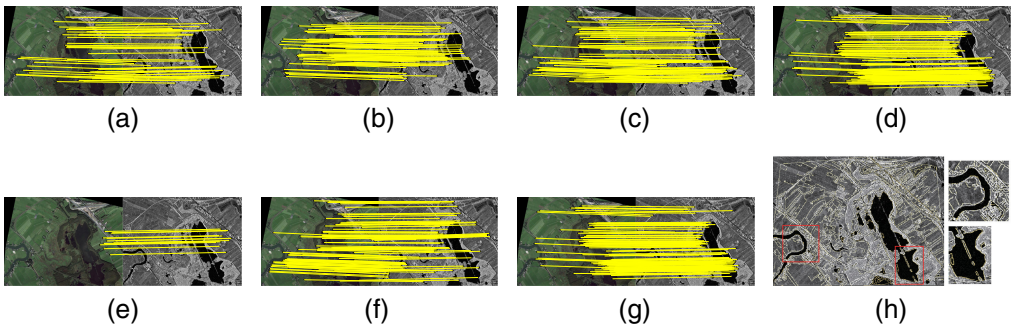

Fig. 8 Registration results of seven methods in image pair 1. (a)–(g) Matches found using OS-SIFT, PCSD, RIFT, HardNet, MatchosNet, CNet, and MSA-Net, respectively. (h) Registration result of our proposed MSA-Net.

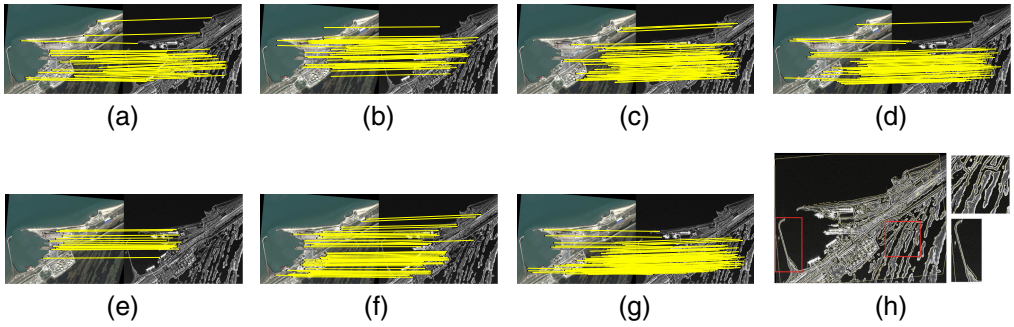

Fig. 9 Registration results of seven methods in image pair 2. (a)–(g) Matches found using OS-SIFT, PCSD, RIFT, HardNet, MatchosNet, CNet, and MSA-Net, respectively. (h) Registration result of our proposed MSA-Net.

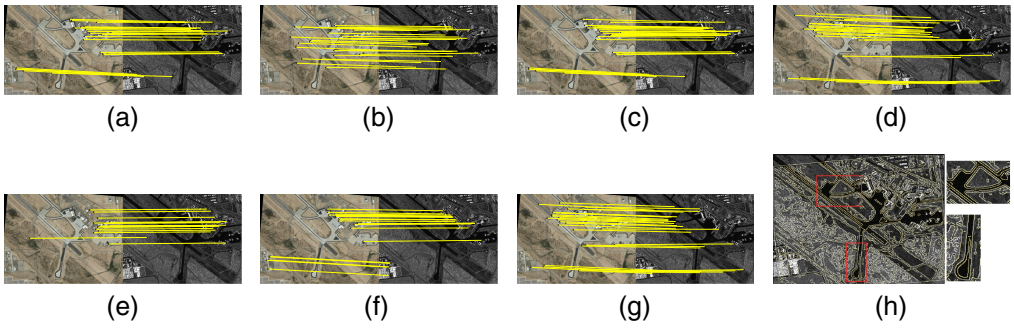

Fig. 10 Registration results of seven methods in image pair 3. (a)–(g) Matches found using OS-SIFT, PCSD, RIFT, HardNet, MatchosNet, CNet, and MSA-Net, respectively. (h) Registration result of our proposed MSA-Net.

Among the four deep learning methods, MatchosNet obtains few matching points with obvious errors in pairs 1 to 3 [as shown in Figs. 8(e)–10(e)], but a certain number of correct matching points can be found in pairs 4 and 5 [as shown in Figs. 11(e) and 12(e)]. The main reason is that MatchosNet is a position-matching algorithm, which needs to know the approximate offset of the two images to be registered in advance. Therefore, it is often difficult for MatchosNet to find the correct number of matching points in a pair of images with rotation and translation transformations. HardNet, CNet, and MSA-Net are patch-based feature extraction networks. The difference between the three is the weight-sharing strategy of the networks. Specifically, HardNet is a Siamese network with weight sharing, CNet is a pseudo-Siamese network without weight sharing completely, and MSA-Net is a network structure with partial weight

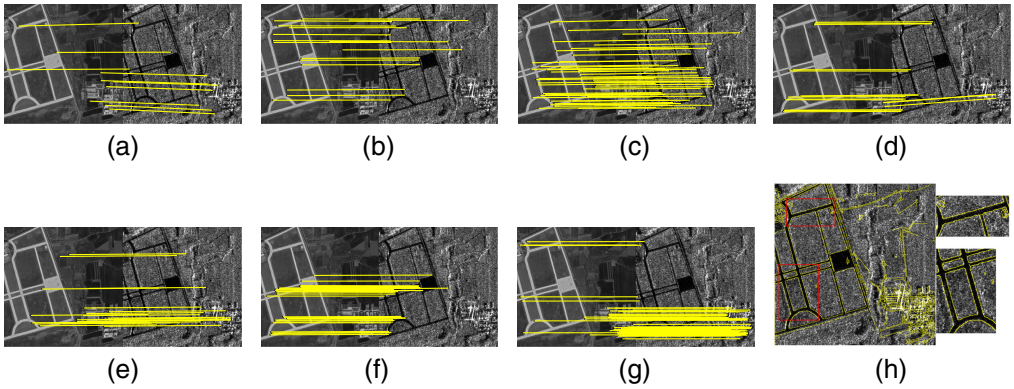

Fig. 11 Registration results of seven methods in image pair 4. (a)–(g) Matches found using OS-SIFT, PCSD, RIFT, HardNet, MatchosNet, CNet, and MSA-Net, respectively. (h) Registration result of our proposed MSA-Net.

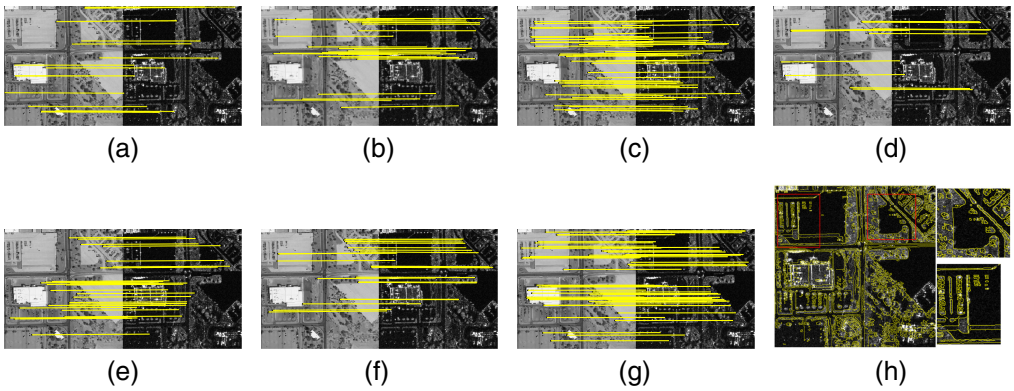

Fig. 12 Registration results of seven methods in image pair 5. (a)–(g) Matches found using OS-SIFT, PCSD, RIFT, HardNet, MatchosNet, CNet, and MSA-Net, respectively. (h) Registration result of our proposed MSA-Net.

sharing. HardNet does not consider the modal differences, and CNet does not take the mining of the modality-shared features into account, so the matching points obtained by both [as shown in Figs. 8(d) and 8(f)–12(d) and 12(f)] are not as much as those obtained by MSA-Net. As shown in Figs. 8(g)–12(g), our proposed MSA-Net achieves the highest number of matching points among the seven registration methods. MSA-Net combines two-stream CNNs and NLA to extract both local and global features and obtains features with different perceptual ranges to cope with radiation differences between SAR and optical images. In Figs. 8(h)–12(h), after carefully observing the registration results, we find that the outlines of the transformed optical image and the corresponding SAR images almost overlap each other. Meanwhile, the regions circled by the red rectangles are well-matched. In summary, compared with the other six methods, the proposed MSA-Net achieves the best registration results and the most matches on the five pairs of images, which can prove the effectiveness of the proposed method.

Table 4 shows the quantitative comparison of the seven methods on five pairs of test images, including NCM, RCM, and RMSE. The higher the NCM and RCM and the lower the RMSE, the better the performance of the registration algorithm. The "*" in Table 4 indicates that the method is invalid for this pair of images. Even though the algorithm finds some matching points, the error of these matching points is relatively large.

Our proposed MSA-Net obtains the highest NCM and RCM and the lowest RMSE among the seven algorithms. This is mainly due to the fact that our algorithm not only considers the radiation differences between SAR and optical images but also mines the shared features between the two modalities to improve the similarity of features. In order to achieve the above, MSA-Net

Table 4 Quantitative comparison of the proposed MSA-Net with OS-SIFT, PCSD, RIFT, HardNet, MatchosNet, and CNet. The bold values represent the best results among all compared methods.

Pair	Metric	OS-SIFT	PCSD	RIFT	HardNet	MatchosNet	CNet	MSA-Net
1	NCM	50	73	85	110	13	101	**119**
	RCM (%)	33.33	48.67	56.67	73.33	*	67.33	**79.33**
	RMSE	1.3516	1.2018	1.0523	0.9816	*	1.1243	**0.9520**
2	NCM	40	54	65	67	17	57	**78**
	RCM (%)	26.67	36	43.33	44.66	*	38	**52**
	RMSE	1.0425	1.1443	0.9795	0.9589	*	1.1790	**0.8790**
3	NCM	21	23	25	26	16	21	**29**
	RCM (%)	42	46	50	42	*	42	**58**
	RMSE	1.3659	1.1039	1.0301	0.9659	*	1.1414	**0.8863**
4	NCM	10	16	47	16	22	34	**69**
	RCM (%)	*	5.93	18.57	6.31	8.69	13.43	**29.99**
	RMSE	*	1.7755	1.6657	1.5570	1.5085	1.6414	**1.4346**
5	NCM	18	25	48	13	27	17	**62**
	RCM (%)	8.87	10.34	24.13	6.40	13.30	8.37	**40.78**
	RMSE	2.0426	1.8089	1.6524	1.5198	1.6484	1.4634	**1.4201**

constructs two MSpFL modules and three MShFL modules to extract local and global features and designs a loss function that considers similarity constraints. PCSD, RIFT, HardNet, and CNet all find some correct matching points in the five pairs of test images. PCSD and RIFT construct descriptors based on phase congruency, which can cope with radiation differences. HardNet and CNet are patch-based deep feature extraction networks, and the network can mine similar features of the two modalities through data learning. However, as the radiation difference increases, the matching points will decrease accordingly. PCSD, RIFT, HardNet, and CNet obtain more matching points in pairs 1 and 2 than in pairs 3 to 5. OS-SIFT and MatchosNet fail on some image pairs. OS-SIFT is a gradient-based descriptor that is susceptible to radiation differences. MatchosNet is an algorithm based on position matching, which fails in image pairs with rotation and translation transformations.

4.4.7 *Computational time of the seven methods*

The computational time of the seven algorithms on the five pairs of images is shown in Table 5. Among the seven algorithms, RIFT has the fastest computational efficiency. This is because it

Table 5 Computational time of the seven methods.

Pair	OS-SIFT (s)	PCSD (s)	RIFT (s)	HardNet (s)	MatchosNet (s)	CNet (s)	MSA-Net (s)
1	31.32	44.45	12.73	13.79	17.16	16.05	19.33
2	25.32	41.26	9.57	11.86	14.28	17.89	14.34
3	24.92	45.26	7.71	8.82	10.65	13.14	13.34
4	19.73	42.24	8.63	9.65	10.65	16.14	18.47
5	20.92	43.10	7.71	9.78	10.43	15.42	16.87

uses the FAST detector to extract key points, the phase consistency algorithm to extract descriptors, and the ratio threshold for matching. The calculation time of OS-SIFT is much longer than that of RIFT. Because OS-SIFT extracts keypoints and constructs descriptors in the multiscale space, and it takes a certain amount of time to construct a multiscale space. PCSD takes the longest time, mainly because it uses template matching. Each keypoint calculates its similarity to all keypoints in the local window. The larger the window is, the more the keypoints are, and the longer the time will be. The computational time of the four deep learning algorithms is relatively close. HardNet takes the least time because it is a Siamese network with the least network parameters. MatchosNet and CNet are pseudo-Siamese networks with more parameters than HardNet. MSA-Net has a partially shared network structure and three NLA modules. The computational complexity of the NLA module is the square of the number of samples. So the computation time of MSA-Net is longer than that of MatchosNet and CNet.

5 Conclusion

This paper proposes a deep learning algorithm, called MSA-Net, for accomplishing SAR and optical image registration. The algorithm combines a dual-stream network with NLA to construct modality-shared features descriptors for SAR and optical images. In MSA-Net, the MSpFL module extracts shallow features from SAR and optical images to obtain modality-specific features, whereas the MShFL module combines the LFE module and NLA module to obtain richer deep modality-shared features, including both local and global features. A loss function that combines triplet loss with cross-modality similarity constraint is also proposed to further improve matching accuracy by constraining the network to be immune to radiometric differences. The experiments are conducted on a public dataset and five pairs of test SAR and optical images. Specifically, the validation experiments verify that the partially shared weight strategy, the NLA, and cross-modality similarity constraints in MSA-Net help to improve the matching accuracy of SAR and optical images. In addition, the comparative experimental results demonstrate that our proposed method achieves the best registration results among the seven registration methods. In future work, we would like to study more diverse attention mechanisms and feature fusion techniques to further enhance the capabilities of the network for SAR and optical image registration.

Acknowledgements

The authors would like to thank all the anonymous reviewers for their valuable comments that helped to improve the quality of this paper. No potential conflicts of interest were reported by the author(s). This work was supported by the National Natural Science Foundation of China (Grant No. 62172321) and the Civil Space Thirteen Five Years Pre-Research Project (Grant No. D040114).

References

1. Y. Wu et al., "Fusion of synthetic aperture radar and visible images based on variational multiscale image decomposition," *J. Appl. Remote Sens* **11**(2), 025006 (2017).
2. A. Shakya, M. Biswas, and M. Pal, "CNN-based fusion and classification of SAR and Optical data," *Int. J. Remote Sens.* **41**(22), 8839–8861 (2020).
3. X. Jiang et al., "Building damage detection via superpixel-based belief fusion of space-borne SAR and optical images," *IEEE Sens. J.* **20**(4), 2008–2022 (2020).
4. Z. Zhang et al., "Endoscope image mosaic based on pyramid ORB," *Biomed. Signal Process. Control* **71**, 103261 (2022).
5. S. Eken et al., "Resource-and content-aware, scalable stitching framework for remote sensing images," *Arab. J. Geosci.* **12**, 1–13 (2019).
6. S. Eken and A. Sayar, "A MapReduce-based distributed and scalable framework for stitching of satellite mosaic images," *Arab. J. Geosci.* **14**(18), 1–16 (2021).
7. G. N. Vivekananda, R. Swathi, and A. Sujith, "Multi-temporal image analysis for LULC classification and change detection," *Eur. J. Remote Sens.* **54**(suppl 2), 189–199 (2021).
8. Y. Wu et al., "Commonality autoencoder: learning common features for change detection from heterogeneous images," *IEEE Trans. Neural Networks Learn. Syst.* **33**, 4257–4270 (2021)

9. X. Li et al., "Collaborative attention-based heterogeneous gated fusion network for land cover classification," *IEEE Trans. Geosci. Remote Sens.* **59**(5), 3829–3845 (2020).
10. W. Kang et al., "CFNet: a cross fusion network for joint land cover classification using optical and SAR images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **15**, 1562–1574 (2022).
11. X. Li et al., "Dense adaptive grouping distillation network for multimodal land cover classification with privileged modality," *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022).
12. X. Li, L. Lei, and G. Kuang, "Locality-constrained bilinear network for land cover classification using heterogeneous images," *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
13. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
14. J. Fan et al., "SAR image registration using phase congruency and nonlinear diffusion-based SIFT," *IEEE Geosci. Remote Sens. Lett.* **12**(3), 562–566 (2015).
15. F. Dellinger et al., "SAR-SIFT: a SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.* **53**(1), 453–466 (2015).
16. 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.* **56**(6), 3078–3090 (2018).
17. Y. Ye et al., "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.* **55**(5), 2941–2958 (2017).
18. J. Fan et al., "SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor," *IEEE Trans. Geosci. Remote Sens.* **56**(9), 5368–5379 (2018).
19. J. Li, Q. Hu, and M. Ai, "RIFT: multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.* **29**, 3296–3310 (2020).
20. Y. Tian, B. Fan, and F. Wu, "L2-Net: deep learning of discriminative patch descriptor in Euclidean space," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 661–669 (2017).
21. A. Mishchuk et al., "Working hard to know your neighbor's margins: local descriptor learning loss," in *Adv. Neural Inf. Process. Syst.* (2017).
22. 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.* **158**, 241–248 (2019).
23. L. H. Hughes et al. "Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN," *IEEE Geosci. Remote Sens. Lett.* **15**(5), 784–788 (2018).
24. H. Zhang et al., "Registration of multimodal remote sensing image based on deep fully convolutional neural network," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **12**(8), 3028–3042 (2019).
25. L. H. Hughes et al., "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogramm. Remote Sens.* **169**, 166–179 (2020).
26. Y. Liao et al., "Feature matching and position matching between optical and SAR with local deep feature descriptor," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **15**, 448–462 (2021).
27. D. Quan et al., "Deep feature correlation learning for multi-modal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2022).
28. H. Han, C. Li, and X. Qiu, "Multi-modal remote sensing image matching method based on deep learning technology," *J. Phys. Conf. Ser.* **2083**(3), 032093 (2021).
29. J. Fan et al., "A novel multiscale adaptive binning phase congruency feature for SAR and optical image registration," *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2022).
30. A. Dosovitskiy et al., "An image is worth 16 × 16 words: transformers for image recognition at scale," https://doi.org/10.48550/arXiv.2010.11929 (2020).
31. A. Vaswani et al., "Attention is all you need," https://doi.org/10.48550/arXiv.1706.03762 (2017).
32. H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2022).
33. H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022).
34. X. Liu et al., "High resolution SAR image classification using global-local network structure based on vision transformer and CNN," *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
35. X. Wang et al., "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 7794–7803 (2018).
36. M. Schmitt, L. H. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," in *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-1*, pp. 141–146 (2018).
37. A. Paszke et al., "Automatic differentiation in PyTorch," in *NIPS Workshops* (2017).

Xin Hu received her BE degree from Xi'an University of Technology in 2019. She is currently pursuing her PhD with the Remote Sensing Image Processing and Fusion Group of the School of Electronic Engineering at Xidian University. Her main research direction is remote sensing image registration.

Yan Wu received her BS degree in information processing and her MA and PhD degrees in signal and information processing from Xidian University, Xi'an, China, in 1987, 1998, and 2003, respectively. Since 2005, she has been a professor in the Department of Electronic Engineering, Xidian University. Her broad research interests include remote sensing image analysis and interpretation, data fusion of multisensory images, synthetic aperture radar auto-target recognition, and statistical learning theory and application.

Zhikang Li received his BS degree from the Xidian University, Xi'an, China, in 2021. He is currently pursuing his PhD with the Remote Sensing Image Processing and Fusion Group, School of Electronic Engineering, Xidian University, Xi'an, China. His research interests include synthetic aperture radar image analysis and feature extraction.

Xiaoru Zhao received his BS degree from the Taiyuan University of Technology, Taiyuan, China, in 2020. He is currently pursuing his MA degree with the Remote Sensing Image Processing and Fusion Group, School of Electronic Engineering, Xidian University, Xi'an, China. His research interests include synthetic aperture radar image analysis and feature extraction.

Xingyu Liu received her BS degree in engineering from Xidian University, Shaanxi, China, in 2020, majoring in measurement and control technology and instrumentation program. Currently, she continues to pursue her MS degree with the Remote Sensing Image Processing and Fusion Group, Xidian University. Her main research direction is synthetic aperture radar image feature extraction and classification.

Ming Li received his BS degree in electrical engineering and his MS and PhD degrees in signal processing from Xidian University, Xi'an, China, in 1987, 1990, and 2007, respectively. In 1987, he joined the Department of Electronic Engineering, Xidian University, where he is currently a professor with the National Key Laboratory of Radar Signal Processing. His research interests include adaptive signal processing, detection theory, ultrawideband, and synthetic aperture radar image processing.