Deep Convolutional Network Based on Attention Mechanism for Matching Optical and SAR Images

Haiqing He[1, 2,*], Shixun Yu[1], Fuyang Zhou[1], Hai Zhang[1], Longyu Chen[1]

[1] School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China (hyhqing@163.com)
[2] Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China

Keywords: Image matching, Deep learning, Deep convolutional network, Geometric distortion, Nonlinear radiation.

Abstract

Complex geometric distortions and nonlinear radiation differences between optical and synthetic aperture radar (SAR) images present challenges for the matching of sufficient and evenly distributed corresponding points. To address this problem, this paper proposes a deep convolutional network based on an attention mechanism for matching optical and SAR images. In order to obtain robust feature points, we employ phase consistency instead of image intensity and gradient information for feature detection. A deep convolutional network (DCN) is designed to extract high-level semantic features between optical and SAR images, providing robustness to geometric distortion and nonlinear radiation changes. Notably, incorporating multiple inverted residual structures in the DCN facilitates efficient extraction of local and global features, promoting feature reuse, and reducing the loss of key features. Furthermore, a dense feature fusion module based on coordinate attention is designed, focusing on the spatial positional information of effective features, integrating key features into deep descriptors to enhance the robustness of deep descriptors to nonlinear radiometric differences. A coarse-to-fine strategy is then employed to enhance accuracy by eliminating mismatches. Experimental results demonstrate that the proposed network performs better than the manually designed descriptors-based methods and the state-of-the-art deep learning networks in both matching effectiveness and accuracy. Specifically, the number of matches achieved is approximately 2 times greater than that of other methods, with a 10% improvement in F-measure.

1. Introduction

Synthetic Aperture Radar (SAR) is an active imaging sensor that utilizes microwaves to observe Earth targets. It possesses key characteristics such as all-weather capability, wide-area coverage, and strong penetrative ability, making it suitable for high-resolution Earth observation applications. The joint interpretation of optical images and SAR images is widely applied in image fusion (Yan and Kong, 2020), image registration (Quan et al., 2022), change detection (Liu et al., 2019), 3D reconstruction (Zhang et al., 2022a), and other fields, with optical and SAR image matching being one of the key technologies for these applications. However, significant nonlinear radiometric differences and geometric distortions exist between optical images and SAR images, along with speckle noise present in SAR images, making optical and SAR image matching challenging. In recent years, many researchers have proposed various matching methods for optical and SAR images, mainly divided into three categories: area-based matching, feature-based matching, and deep learning-based matching.

Among area-based matching methods, the most commonly used approaches are normalized cross-correlation and mutual information. These two methods directly utilize intensity information from images for computation. Ye et al. (Ye et al., 2017) employed phase consistency information instead of image intensity information for matching, proposing the histogram of oriented phase congruency (HOPC) algorithm, which effectively counters nonlinear radiometric differences and enhances matching performance. Li et al. (Li et al., 2020b) calculated multi-directional phase feature maps based on detected feature points and used them to construct descriptors. Building upon HOPC, Fan et al. (Fan et al., 2021) proposed angle-weighted orientation gradient descriptors, distributing gradient values to the two most correlated directions and employing three-dimensional phase information as a similarity measure, significantly improving matching performance. Despite the high accuracy of area-based matching methods, they suffer from high computational complexity and poor robustness to illumination changes and nonlinear radiometric differences.

Feature-based matching is a commonly used method in the field of image matching, which mainly consists of three steps: keypoint detection, descriptor construction, and feature matching, with the most famous being the scale-invariant feature transform (SIFT) algorithm (Yoo and Han, 2009). To meet the matching requirements of optical and SAR images, Dellinger et al. (Dellinger et al., 2015) proposed the SAR-SIFT method. This method utilizes directional gradients to construct descriptors, which are robust to speckle noise and finally combined with SIFT. Xiang et al. (Xiang et al., 2018) improved the method for detecting feature points in SAR images and constructed new descriptors for image registration. Li et al. (Li et al., 2020a) used phase consistency instead of image intensity for feature point detection and proposed the rotation invariant feature transform (RIFT) algorithm for multi-modal image matching. Feature-based matching has been widely applied and has made significant contributions. However, due to significant nonlinear radiometric differences between optical and SAR images, manually designed features have poor robustness and are difficult to produce highly repeatable features.

In the past decade, deep learning-based matching methods have garnered widespread attention in multi-modal image matching tasks due to their excellent generalization capabilities. Han et al. (Xufeng et al., 2015) introduced deep learning into image matching and proposed MatchNet, which adopts a dual-branch network structure to extract features from image patches and calculate feature similarity to obtain matching points. Reference (Merkle et al., 2017) proposed a siamese deep neural network (DNN) for extracting deep features from optical and SAR

https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-201-2024 |

images for image matching, where expanded convolutional layers are used to increase the receptive field and enhance image features. Li et al. (Li et al., 2022) utilized the feature learning network SARPointNet to obtain feature points and descriptors of images, improving matching performance. Zhang et al. (Zhang et al., 2022b) proposed the optical and SAR Image Matching Network (OSMNet), which adopts a multi-level feature fusion network architecture combined with a channel attention mechanism to extract better features from optical images for feature matching. Existing deep learning methods suffer from shallow network layers, making it difficult to capture higher-level semantic features of images; existing deep convolutional neural networks extract a multitude of features, which often contain noise and outliers, lacking robustness when facing significant nonlinear radiometric differences; the network models pay less attention to spatial positional information of features. To address these issues, this paper proposes a DCN based on an attention mechanism for optical and SAR image matching. The cardinal contributions of this work are itemized as follows:

1. The deep convolutional network (DCN) is composed of multiple inverted residual structures, which can significantly reduce the number of network parameters. It effectively integrates images' local and global semantic information, enabling feature reuse and demonstrating superior performance.
2. A dense feature fusion module based on spatial attention (DFFCA) is designed. The attention mechanism focuses more on the spatial positional information of dense features, highlighting key features within dense features. It maximizes integrating these key features into deep descriptors for feature matching.

2. Method

The method proposed in this paper comprises three parts: utilizing phase consistency (PC) instead of image intensity information for feature detection; constructing deep descriptors using a deep convolutional neural network based on the detected feature points; obtaining initial matching points through nearest neighbor matching, and then employing dynamic adaptive thresholding and the Random Sample Consensus (RANSAC) algorithm to removal mismatched points and obtain the final matching points. The flowchart of this paper is illustrated in Figure 1.

2.1 PC Detection

Classical image matching methods generally rely on the intensity and gradient information of images, which belong to spatial domain information. Apart from spatial domain information, frequency domain information (such as phase information) can also be employed to describe images. Oppenheim and Lim (Quan et al., 2023) first revealed the importance of phase information in preserving image features, as phase information exhibits robustness to changes in image contrast, illumination, scale, and so on. Therefore, this paper adopts PC instead of image intensity and gradient information (Li et al., 2020a), and then utilizes the Fast algorithm for feature point detection, ensuring the number of feature points, enhancing the repeatability of feature points, and exhibiting robustness to nonlinear radiometric differences. The formula for phase consistency calculation is as follows:

$$PC(x,y)=\frac{\sum_{m}\sum_{n}w_n(x,y)\lfloor A_{mn}(x,y)\Delta\phi_{mn}(x,y)-T\rfloor}{\sum_{m}\sum_{n}A_{mn}(x,y)+\varphi}, \tag{1}$$

where $w_n(x,y)$ = weight function

$A_{mn}(x,y)$ = amplitude of the wavelet scale n and orientation at (x, y)

T = noise threshold

φ = a small value

$\lfloor . \rfloor$ = the quantity contained is equal to itself when its value is positive, otherwise it is zero.

$\Delta\phi_{mn}(x,y)$ = phase deviation function

For the phase deviation function, the formula is:

$$\begin{aligned}A_{mn}(x,y)\Delta\phi_{mn}(x,y)&=\left(E_{mn}(x,y)\phi_E(x,y)+O_{mn}(x,y)\phi_O(x,y)\right)\\&\quad-\left|\left(E_{mn}(x,y)\phi_E(x,y)-O_{mn}(x,y)\phi_O(x,y)\right)\right|,\end{aligned} \tag{2}$$

where $\phi_E(x,y)=\sum_{m}\sum_{n}E_{mn}(x,y)/e(x,y)$

$$\phi_O(x,y)=\sum_{m}\sum_{n}O_{mn}(x,y)/e(x,y)$$

$$e(x,y)=\sqrt{\left(\sum_{m}\sum_{n}E_{mn}(x,y)\right)^2+\left(\sum_{m}\sum_{n}O_{mn}(x,y)\right)^2}$$

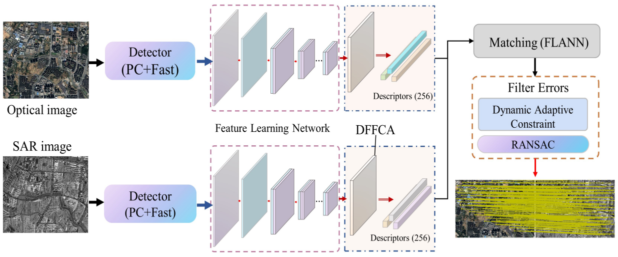


Figure 1. The framework of the proposed method.

2.2 Network Structure

Existing deep learning-based matching methods often utilize networks such as VGG, ResNet, etc., as the backbone network of the network model. These networks have a large number of parameters, high training complexity, and suffer from problems of insufficient local information and global semantic information. To address these issues, this paper adopts the lightweight network MobileNetV2 (Sandler et al., 2018) as the backbone network. Research has shown that using pooling layers in network models can degrade the performance of descriptors. The MobileNetV2 network utilizes fully convolutional layers without using pooling layers, ensuring the performance of descriptors while effectively reducing the loss of key features. The backbone network is composed of multiple inverted residual modules (IRM), as shown in Figure 2(a). The inverted residual module uses depthwise (DW) convolutional layers to reduce the resolution affected, which can decrease the number of parameters of the model. In this module, the use of skip connections effectively extracts local detail information and global contextual information of the image. The mathematical expression of the inverted residual module is:

$$\begin{aligned} &X_i = H_i \times X_{i-1} + X_{i-1}, \text{just } \boldsymbol{R}_i^{w\times h\times c} = \boldsymbol{R}_{i-1}^{w\times h\times c_1} \text{and } s = 1, \\ &\text{Or } \ X_i = H_i \times X_{i-1}, \text{ just } \boldsymbol{R}_i^{w\times h\times c} = \boldsymbol{R}_{i-1}^{w\times h\times c_2} \ \text{ and } s = 2, \end{aligned} \tag{3}$$

where X_i = output of the IRM

H_i = composite function composed of convolution, Relu6, and BN

$\boldsymbol{R}_i^{w\times h\times c}$ = the 3D tensor of the output feature map of the i IRM

s = step size of the DW convolution layer in the i IRM

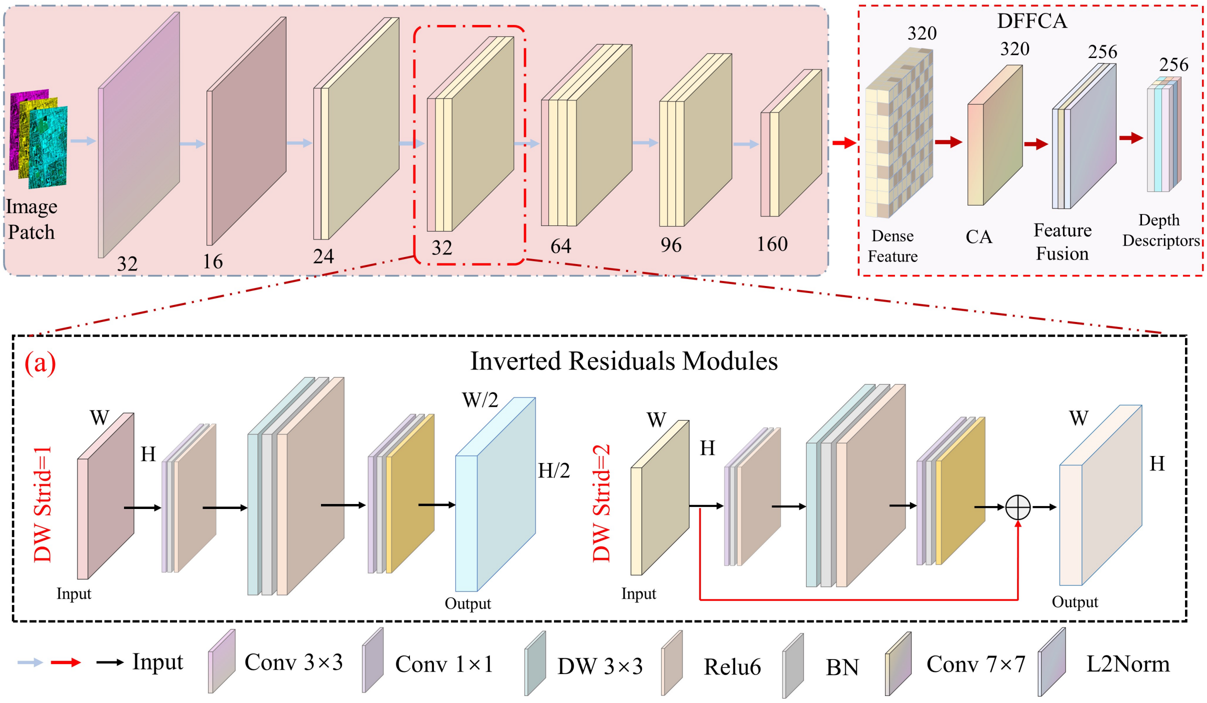


Figure 2. Structure of the proposed network.

The MobileNetV2 network is an image classification network that cannot be directly used for extracting deep descriptors from images. In this paper, the network model is improved by replacing the last two layers of the backbone network with a dense feature fusion module. The proposed network model is illustrated in Figure 2, consisting of convolutional layers, inverted residual structures, and dense feature fusion modules, with detailed information provided in Table 1. To integrate the dense features extracted by the backbone network into the deep descriptors, we designed a dense feature fusion module based on coordinate attention (DFFCA), as shown in Figure 3. The coordinate attention (CA) mechanism (Hou et al., 2021), with almost no additional computational cost, effectively integrates spatial positional information into dense features, emphasizing key features and suppressing the contribution of non-significant features. The DFFCA effectively incorporates key features into deep descriptors, thereby improving their matching accuracy and robustness.

2.3 Loss Function

To train the network, the loss function adopts the "hardest example" mining strategy. Mishchuk et al. (Mishchuk et al., 2017) proposed the HardNet loss function, which requires that the distance between each row and each column with the correct match be minimized. For each positive sample, n negative samples are generated, and the one with the smallest L_2 distance to the correct match is selected for optimization of the model. Using the distance formula of L_2 : $D\left(o_i, s_j\right) = \sqrt{2 - 2o_i s_j}, i = 1\cdots n, j = 1\cdots n$, where o_i represents the i deep descriptor of the optical image, and s_j represents the j deep descriptor of the SAR image. For each pair of matched deep descriptors $\left(o_i, s_i\right)$, find the nearest non-matching deep descriptor $s_{j_{\min}}$ to the deep descriptor o_i, and the nearest non-matching deep descriptor $o_{k_{\min}}$ to the deep descriptor s_j, forming a quadruplet $(o_i, s_i, s_{j_{\min}}, o_{k_{\min}})$, where $j \neq i$ and $k \neq i$. Then, a triplet $(o_i, s_i, s_{j_{\min}}, o_{k_{\min}})$ is formed from the quadruplet $(o_i, s_i, s_{j_{\min}})$, if $D\left(o_i, s_{j_{\min}}\right) < D\left(o_{k_{\min}}, s_i\right)$ forms another triplet $(o_i, s_i, o_{k_{\min}})$. The objective of the loss function is to minimize the distance between matching pairs of depth descriptors and non-matching depth descriptors. The loss function continuously reduces the distance between matching pairs and increases the distance between non-matching pairs, thus optimizing the network model during the backpropagation process, and completing the model training. The formula for the loss function is as follows:

$$L = \frac{1}{n}\sum_{i=1,n} \max\left(0, 1 + D\left(o_i, s_i\right) - \min\left(D\left(o_i, s_{j_{\min}}\right), D\left(o_{k_{\min}}, s_i\right)\right)\right). \tag{4}$$

Input	Operator	c	n	s
$224\times224\times3$	Conv2d 3×3	32	1	2
$112\times112\times32$	IRM	16	1	1
$112\times112\times16$	IRM	24	2	2
$56\times56\times24$	IRM	32	3	2
$28\times28\times32$	IRM	64	4	2
$14\times14\times64$	IRM	96	3	1
$14\times14\times96$	IRM	160	3	2
$7\times7\times160$	IRM	320	1	1
$7\times7\times320$	**CA**	320	-	-
$7\times7\times320$	Conv2d7×7+BN	256	-	-

Table 1. Details of the proposed network.

In the table, c represents the depth channel of the output feature map, n represents the number of repetitions of IRM, s represents the step size of the first layer input for each sequence, and all other steps are 1.

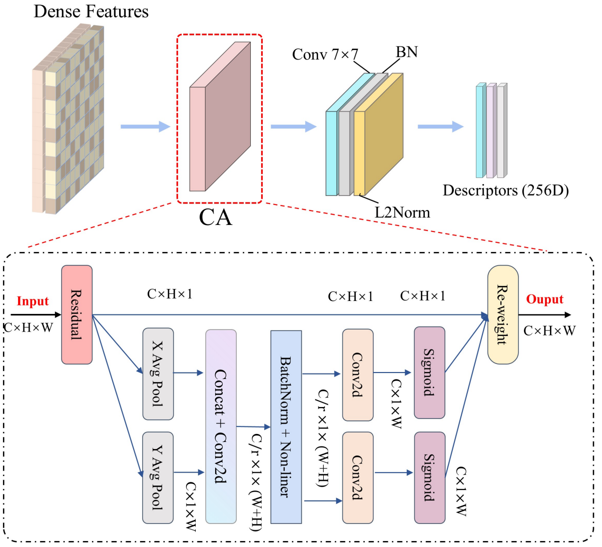


Figure 3. Structure of DFFCA.

2.4 Filter Errors

This paper employs a strategy of coarse-to-fine to remove mismatched points. Traditionally, to assess the quality of the i pair of matches, a fixed multiplier factor t is commonly used as a threshold. When $d < t \times d'$ it is considered that this pair of matches has good quality. Due to significant differences between optical and SAR images, it is challenging to determine the multiplier factor t. In this paper, an adaptive threshold constraint is used instead of the multiplier factor, eliminating the need for manual threshold adjustment and demonstrating good adaptability. The average difference between the nearest neighbor point and the second nearest neighbor point is calculated as the basis for judgment, with the calculation formula as follows:

$$avgd = \frac{1}{N}\sum_{i=1}^{N}\left(d' - d\right), \tag{5}$$

where N = number of reference image feature points
d = image coordinates
d' = distance of the second nearest neighbor point
$avgd$ = mean distance

When the matching points meet the condition $d' > avgd + d$, it is considered that the quality of this matching point is good.

After adaptive thresholding coarse screening, Numerous mismatched points can be removed, significantly improving the inlier ratio, but there are still some mismatched points. This paper uses the RANSAC algorithm to refine the coarse-screened matching points. Due to the complex geometric distortion and significant nonlinear radiation differences between optical and SAR images, using a single geometric model as the estimation model cannot eliminate mismatched points. Therefore, this paper adopts an affine transformation model and a homography matrix as the RANSAC algorithm estimation model, which effectively improves the computational efficiency of RANSAC random sampling and geometric consistency verification, and enhances the robustness of the matching results.

3. Experiments and Results

3.1 Experimental Dataset and Implementation Details

To train the network models, this paper utilizes two publicly available datasets containing optical and SAR imagery. The QXS-SROPT dataset was proposed by Huang et al. (Huang et al., 2021) in 2021, comprising 20,000 images obtained from Gaofen-3 synthetic aperture radar satellite imagery and Google Earth imagery. The SEN1-2 dataset was introduced by Schmitt et al. (Schmitt et al., 2018) in 2018. It includes 282,384 pairs of SAR and optical image patches from various regions worldwide and all meteorological seasons. For the training process, this paper utilizes the summer subset of the SEN1-2 dataset. The images in both datasets are sized 256×256, and during network training, they are randomly cropped into 224×224 image patches.

The test data consists of real optical and SAR images, as shown in Table 2. The test data are optical and SAR images obtained by different sensors. They vary in resolution, imaging time, and ground coverage. Due to the presence of complex geometric distortions and significant nonlinear radiometric differences among the images, they are particularly suitable for evaluating the proposed methods.

During the training process, the MobileNetV2 backbone network is trained using transfer learning. This paper employs the Adam optimizer for training, with an initial learning rate of 0.001 and a batch size of 256. Training is conducted using a single NVIDIA RTX 4060Ti GPU to improve training efficiency.

Pair	Sensor(SAR/Optical)	Size	Resolution
A	Sentinel-1/Sentinel-2	1000×800	10
B	GF3/GF2	1000×800	2
C	Sentinel-1/Google Earth	900×700	10
D	GF3/ Google Earth	900×700	2
E	TerraSAR-X/Google Earth	500×500	3
F	TerraSAR-X/Google Earth	600×600	2

Table 2. Test image pairs and their characteristics.

3.2 Experiment and Discussion

To evaluate the matching performance of the proposed method, this paper adopts the number of correct matching points (NCM), *F-measure*, and root mean square error (RMSE) as evaluation metrics. To assess different methods, a matching point is considered correct if its error rate is less than 3 pixels. The *F-measure* represents the matching performance and is defined as the harmonic mean of precision and recall, defined as:

$$F\text{-}measure = \frac{2 \times \mathrm{MP} \times \mathrm{Recall}}{\mathrm{MP+Recall}}, \tag{6}$$

where MP = accuracy of matching points, Ratio of NCM to total matching points

Recall = recall rate, Ratio of NCM to keypoints

Root mean square error (RMSE) reflects the matching accuracy of matching point pairs, defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{NCM}} \sum_{i=1}^{\mathrm{NCM}} \left[(x_i' - x_i)^2 + (y_i' - y_i)^2 \right]}, \tag{7}$$

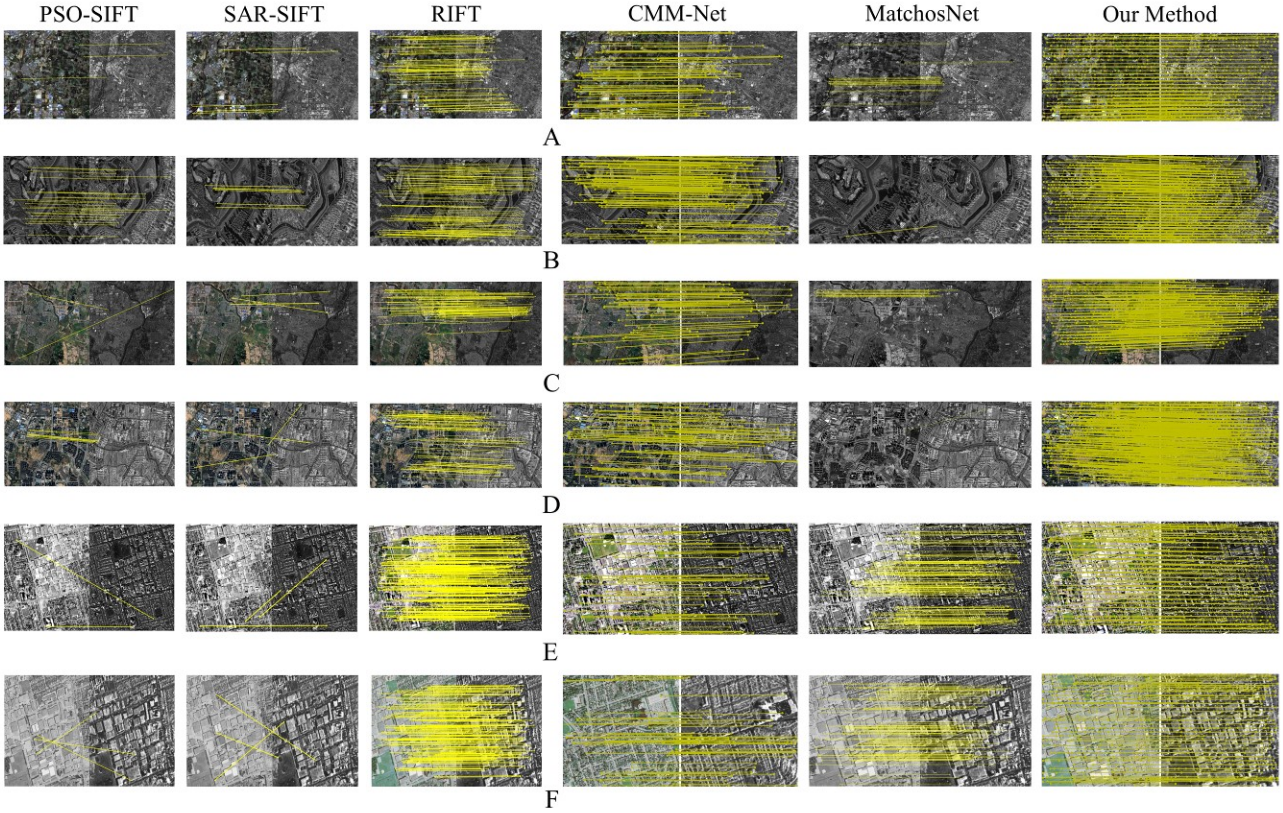


Figure 4. Matching results of six methods on test data.

3.2.1 Qualitative Comparisons: Figure 4 illustrates the matching results of the proposed method compared to five other matching algorithms: PSO-SIFT (Ma et al., 2017) SAR-SIFT (Dellinger et al., 2015), RIFT (Li et al., 2020a), CMM-Net (Lan et al., 2021), and MatchosNet (Liao et al., 2022). The figure demonstrates that the proposed method obtains a large number of matching points with a uniform distribution of keypoints.

The proposed method fully utilizes the nonlinear modeling capability of deep learning to extract more robust, stable, and reproducible feature points. Compared to two traditional manually designed descriptors (PSO-SIFT and SAR-SIFT), the deep learning-based method demonstrates superior performance on test data. The RIFT algorithm uses phase congruency information for matching, and transforming images from the spatial domain to the frequency domain. This method achieves successful matching on all test data, but the NCM is much lower than that of the proposed method. The CMM-Net method extracts advanced semantic information from images, resulting in a higher NCM than that obtained by MatchosNet. MatchosNet builds descriptors based on local information in graphics, leading to matching failure on significantly different images (like pairs B and F). Qualitative results demonstrate that the proposed method not only achieves a greater number of NCM but also exhibits a more even distribution of matching points. This is attributed to the utilization of PC for feature point detection in the proposed method, which remains invariant to nonlinear radiometric differences. The network simultaneously considers both low-level and high-level semantic information of images. The designed DFFCA focuses on spatial positional information of features, integrating key features into deep descriptors, thereby enhancing descriptor performance. Consequently, both the quantity and distribution of NCM surpass those of other methods.

Pair		**Method**					
		PSO-SIFT	**SAR-SIFT**	**RIFT**	**CMM-Net**	**MatchosNet**	**our method**
A	NCM	3	5	84	58	29	**389**
	F-measure	0.001	0.003	0.027	0.020	0.001	**0.105**
	RMSE	2.325	1.861	1.665	2.831	2.105	**1.344**
B	NCM	24	6	112	137	-	**410**
	F-measure	0.005	0.003	0.322	0.058	-	**0.106**
	RMSE	1.676	1.7087	1.512	2.81	-	**1.345**
C	NCM	-	-	91	60	20	**350**
	F-measure	-	-	0.028	0.028	0.001	**0.11259**
	RMSE	-	-	1.490	3.010	1.728	**1.333**
D	NCM	-	-	79	70	-	**372**
	F-measure	-	-	0.022	0.030	-	**0.117**
	RMSE	-	-	1.521	2.974	-	**1.345**
E	NCM	-	-	241	49	116	**282**
	F-measure	-	-	0.076	0.059	0.018	**0.21092**
	RMSE	-	-	1.640	2.853	1.677	**1.323**
F	NCM	-	-	183	60	109	**308**
	F-measure	-	-	0.056	0.042	0.014	**0.171**
	RMSE	-	-	1.457	2.771	1.711	**1.342**

Table 3. Quantitative matching results.

3.2.2 Quantitative Comparisons: In quantitative experiments, this paper employs three evaluation metrics to analyze the matching performance of each method. The threshold for NCM and RMSE is set to 3 pixels. The experimental results are presented in Table 3, where the best results are highlighted in bold, representing the average of 10 trials.

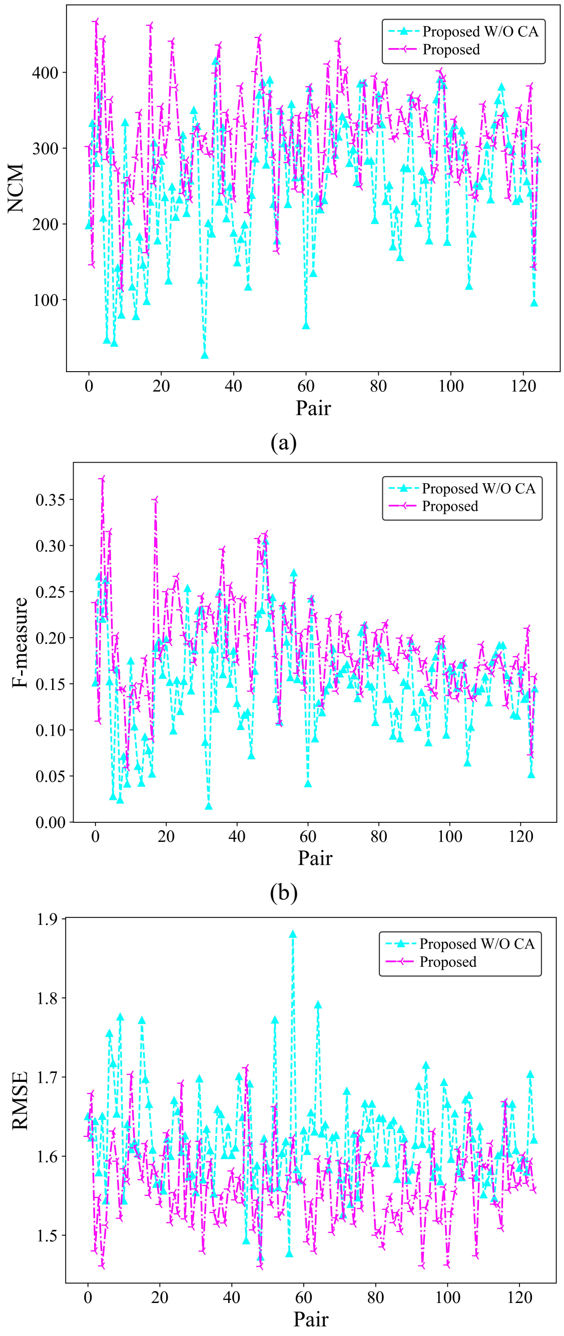
The results indicate that the matching performance of the proposed method is significantly better than other methods. Both PSO-SIFT and SAR-SIFT algorithms exhibit poor matching performance on the test data. Manually designed descriptors are limited to low-level semantic information of images and lack robustness against significant nonlinear radiometric differences. RIFT utilizes PC to construct descriptors, which have a certain invariance to nonlinear radiometric differences. However, the NCM of RIFT is much lower than that of the method proposed in this paper. Qualitative experiments show that the distribution of matching points by the RIFT algorithm is uneven, as seen for instance in image pairs A and C. CMM-Net is a method for matching heterogeneous images. Unlike traditional methods, feature point detection in CMM-Net is conducted after feature description. Both feature point detection and description in CMM-Net are performed on feature maps, extracting features containing high-level semantic information, which are more suitable for matching heterogeneous images. However, its drawback lies in the poor localization accuracy of matching points. MatchosNet is specifically designed for optical and SAR image matching. MatchosNet fails to match in pairs B and D. MatchosNet utilizes DOG for feature point detection, which lacks robustness against nonlinear radiometric differences. Additionally, the construction of descriptors in MatchosNet does not fully consider high-level semantic information of images, resulting in poor matching performance.

As shown in Table 3, the proposed method achieves the highest number of NCM on the test images. Through F-measure comparison, the *F-measure* of the proposed method significantly outperforms other methods, overall obtaining the best matching results. This also indicates that the proposed method can extract robust matching points from optical and SAR images. Additionally, the RMSE on the test data is less than 2 pixels for the proposed method. The proposed method utilizes DFFCA to integrate key features containing high-level semantic information into deep descriptors. The network constructs deep descriptors with strong robustness against nonlinear radiometric differences. The proposed method achieves good matching accuracy on optical and SAR images with noise interference and significant nonlinear radiometric differences.

3.3 Ablation Experiment

To further validate the contribution of the attention mechanism to the matching task, this paper conducts ablative experiments to test the impact of the coordinate attention mechanism on matching performance. Using the same dataset to train the network models, 125 pairs of images are selected from the SAR2Opt (Zhao et al., 2022) dataset for testing. The dataset's image pairs size is 600×600, which was not used during training. The experimental results are shown in Figure 5.

The proposed method adds a CA module to the network, as shown in Figure 4. After adding CA, the accuracy of NCM and matching points is further improved, enhancing the matching performance of the algorithm. The CA module pays more attention to the spatial positional information of effective features, integrating key features into attention maps, and suppressing the expression of non-key features. The designed DFFCA integrates attention maps containing key features into deep descriptors, improving the robustness and stability of descriptors to nonlinear radiometric differences.



(a)

(b)

(c)

Figure 5. CA ablation experiment results: (a) NCM, (B) F-measure, (c) RMSE.

4. Conclusion

Due to complex geometric distortions and significant nonlinear radiometric differences between optical and SAR images, traditional matching methods have difficulty obtaining a sufficient and uniformly distributed set of matching points. To tackle this issue, this paper proposes a DCN based on attention mechanisms for matching optical and SAR images. Experimental results validate the accuracy and robustness of the proposed method. Compared with five other methods, the proposed method achieves accurate and stable matching in different scenarios, outperforming other methods. Firstly, the paper uses PC instead of intensity information of images for feature detection, obtaining feature points invariant to nonlinear radiometric differences. Secondly, a DCN is employed to extract both local and global semantic information from images. The network effectively reduces feature loss and achieves feature reuse using an IRM structure. The DFFCA is designed to pay more attention to the spatial positional information of effective features, merging key features from dense features into deep descriptors. The constructed deep descriptors exhibit robustness to nonlinear radiometric differences. Finally, adaptive thresholds and RANSAC are utilized to improve the quantity and accuracy of matching points.

Meanwhile, ablation experiment results confirm the performance of the proposed method CA structure in optical and SAR image matching. The utilization of the CA structure has improved the NCM and matching accuracy. Future work will focus on improving the performance of the network model in constructing deep descriptors, further enhancing the model's generalization ability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 42261075 and 41861062, in part by the Jiangxi Provincial Natural Science Foundation under Grant 20224ACB212003, in part by the Jiangxi Provincial Training Project of Disciplinary, Academic, and Technical Leader (20232BCJ22002), and in part by the State Key Laboratory of Geo-Information Engineering and Key Laboratory of Surveying and Mapping Science and Geospatial Information Technology of MNR, Chinese Academy of Surveying and Mapping under Grant 2022-02-04.

References

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., and Tupin, F., 2015: SAR-SIFT: A SIFT-Like Algorithm for SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 53, 453-466. doi.org/10.1109/TGRS.2014.2323552.

Fan, Z., Zhang, L., Liu, Y., Wang, Q., and Zlatanova, S., 2021: Exploiting High Geopositioning Accuracy of SAR Data to Obtain Accurate Geometric Orientation of Optical Satellite Images. *Remote Sensing*, 13, 3535. doi.org/10.3390/rs13173535.

Hou, Q., Zhou, D., and Feng, J., 2021. Coordinate Attention for Efficient Mobile Network Design, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20-25 June 2021, 13708-13717. doi.org/10.1109/CVPR46437.2021.01350.

Huang, M., Xu, Y., Qian, L., Shi, W., Zhang, Y., Bao, W., Wang, N., Liu, X., and Xiang, X., 2021. The QXS-SAROPT Dataset for Deep Learning in SAR-Optical Data Fusion.

Lan, C., Lu, W., Yu, J., and Xu, Q., 2021: Deep learning algorithm for feature matching of cross modality remote sensing images. *Acta Geodaetica et Cartographica Sinica*, 50, 189-202.

Li, J., Hu, Q., and Ai, M., 2020a: RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Transactions on Image Processing*, 29, 3296-3310. doi.org/10.1109/TIP.2019.2959244.

Li, X., Yang, Y., Yang, B., and Yin, F., 2020b: A Multi-source Remote Sensing Image Matching Method Using Directional Phase Feature. *Geomatics and Information Science of Wuhan University*, 45, 488-494. doi.org/10.13203/j.whugis20180445.

Li, X., Wang, T., Cui, H., Zhang, G., Cheng, Q., Dong, T., and Jiang, B., 2022: SARPointNet: An Automated Feature Learning Framework for Spaceborne SAR Image Registration. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 6371-6381. doi.org/10.1109/JSTARS.2022.3196383.

Liao, Y., Di, Y., Zhou, H., Li, A., Liu, J., Lu, M., and Duan, Q., 2022: Feature Matching and Position Matching Between Optical and SAR With Local Deep Feature Descriptor. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 448-462. doi.org/10.1109/JSTARS.2021.3134676.

Liu, F., Jiao, L., Tang, X., Yang, S., Ma, W., and Hou, B., 2019: Local Restricted Convolutional Neural Network for Change Detection in Polarimetric SAR Images. *IEEE Transactions on Neural Networks and Learning Systems*, 30, 818-833. doi.org/10.1109/TNNLS.2018.2847309.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., and Liu, L., 2017: Remote Sensing Image Registration With Modified SIFT and Enhanced Feature Matching. *IEEE Geoscience and Remote Sensing Letters*, 14, 3-7. doi.org/10.1109/LGRS.2016.2600858.

Merkle, N., Luo, W., Auer, S., Müller, R., and Urtasun, R., 2017: Exploiting Deep Matching and SAR Data for the Geo-Localization Accuracy Improvement of Optical Satellite Images. *Remote Sensing*, 9, 586. doi.org/10.3390/rs9060586, 2017.

Mishchuk, A., Mishkin, D., Radenović, F., and Matas, J., 2017: Working hard to know your neighbor's margins: Local descriptor learning loss.

Quan, D., Wei, H., Wang, S., Lei, R., Duan, B., Li, Y., Hou, B., and Jiao, L., 2022: Self-Distillation Feature Learning Network for Optical and SAR Image Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-18. doi.org/10.1109/TGRS.2022.3173476.

Quan, Y., Zhang, D., Zhang, L., and Tang, J., 2023: Centralized Feature Pyramid for Object Detection. *IEEE Transactions on Image Processing*, 32, 4341-4354. doi.org/10.1109/TIP.2023.3297408.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C., 2018: MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-23 June 2018, 4510-4520. doi.org/10.1109/CVPR.2018.00474.

Schmitt, M., Hughes, L., and Zhu, X., 2018: The SEN1-2 dataset for deep learning in SAR-optical data fusion. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1, 141-146. dio.org/10.5194/isprs-annals-IV-1-141-2018.

Xiang, Y., Wang, F., and You, H., 2018: OS-SIFT: A Robust SIFT-Like Algorithm for High-Resolution Optical-to-SAR Image Registration in Suburban Areas. *IEEE Transactions on Geoscience and Remote Sensing*, 56, 3078-3090. doi.org/10.1109/TGRS.2018.2790483.

Xufeng, H., Leung, T., Jia, Y., Sukthankar, R., and Berg, A. C., 2015. MatchNet: Unifying feature and metric learning for patch-based matching, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7-12 June 2015, 3279-3286. doi.org/10.1109/CVPR.2015.7298948.

Yan, B. and Kong, Y., 2020. A Fusion Method of SAR Image and Optical Image Based on NSCT and Gram-Schmidt Transform. IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 26 Sept.-2 Oct. 2020, 2332-2335. doi.org/10.1109/IGARSS39084.2020.9323158.

Ye, Y., Shan, J., Bruzzone, L., and Shen, L., 2017: Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55, 2941-2958. doi.org/10.1109/TGRS.2017.2656380.

Yoo, J.-C. and Han, T. H., 2009: Fast Normalized Cross-Correlation. *Circuits, Systems and Signal Processing*, 28, 819-843. doi.org/10.1007/s00034-009-9130-7.

Zhang, H., Lin, Y., Teng, F., and Hong, W., 2022a: A Probabilistic Approach for Stereo 3D Point Cloud Reconstruction from Airborne Single-Channel Multi-Aspect SAR Image Sequences. *Remote Sensing*, 14, 5715. doi.org/10.3390/rs14225715.

Zhang, H., Lei, L., Ni, W., Tang, T., Wu, J., Xiang, D., and Kuang, G., 2022b: Explore Better Network Framework for High-Resolution Optical and SAR Image Matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-18. doi.org/10.1109/TGRS.2021.3126939.

Zhao, Y., Celik, T., Liu, N., and Li, H. C., 2022: A Comparative Analysis of GAN-Based Methods for SAR-to-Optical Image Translation. *IEEE Geoscience and Remote Sensing Letters*, 19, 1-5. doi.org/10.1109/LGRS.2022.3177001.