remote sensing

MDPI

Article

SAR and Optical Image Registration Based on Deep Learning with Co-Attention Matching Module

Jiaxing Chen [1,2], Hongtu Xie [1,2,*], Lin Zhang [3], Jun Hu [1,2], Hejun Jiang [2] and Guoqian Wang [4]

[1] School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
[2] Science and Technology on Near-Surface Detection Laboratory, Wuxi 214035, China
[3] Department of Early Warning Technology, Air Force Early Warning Academy, Wuhan 430019, China
[4] The Fifth Affiliated Hospital, Guangzhou Medical University, Guangzhou 510700, China
* Correspondence: xiehongtu@mail.sysu.edu.cn

Abstract: Image registration is the basis for the joint interpretation of synthetic aperture radar (SAR) and optical images. However, the significant nonlinear radiation difference (NRD) and the geometric imaging model difference render the registration quite challenging. To solve this problem, both traditional and deep learning methods are used to extract structural information with dense descriptions of the images, but they ignore that structural information of the image pair is coupled and often process images separately. In this paper, a deep learning-based registration method with a co-attention matching module (CAMM) for SAR and optical images is proposed, which integrates structural feature maps of the image pair to extract keypoints of a single image. First, joint feature detection and description are carried out densely in both images, for which the features are robust to radiation and geometric variation. Then, a CAMM is used to integrate both images' structural features and generate the final keypoint feature maps so that the extracted keypoints are more distinctive and repeatable, which is beneficial to global registration. Finally, considering the difference in the imaging mechanism between SAR and optical images, this paper proposes a new sampling strategy that selects positive samples from the ground-truth position's neighborhood and augments negative samples by randomly sampling distractors in the corresponding image, which makes positive samples more accurate and negative samples more abundant. The experimental results show that the proposed method can significantly improve the accuracy of SAR–optical image registration. Compared to the existing conventional and deep learning methods, the proposed method yields a detector with better repeatability and a descriptor with stronger modality-invariant feature representation.

Keywords: image registration; synthetic aperture radar (SAR) and optical images; deep learning; co-attention matching module (CAMM); keypoint detection; sampling strategy

check for updates

Citation: Chen, J.; Xie, H.; Zhang, L.; Hu, J.; Jiang, H.; Wang, G. SAR and Optical Image Registration Based on Deep Learning with Co-Attention Matching Module. *Remote Sens.* **2023**, *15*, 3879. https://doi.org/10.3390/rs15153879

Academic Editors: Olga Sykioti, Gangyao Kuang, Xin Su and Siqian Zhang

Received: 17 June 2023
Revised: 28 July 2023
Accepted: 1 August 2023
Published: 4 August 2023

1. Introduction

Synthetic aperture radar (SAR) can penetrate the atmosphere actively, and SAR imaging systems are both weather-independent and high-resolution [1], which makes them important tools for macroscopic, long-term, dynamic, and real-time observations of the land and ocean [2–4]. However, SAR images suffer from defects, such as imaging noises and difficult interpretation. In contrast, optical images have rich texture and are easy to interpret, but their imaging quality is highly dependent on the imaging environment. Thus, the joint interpretation of SAR and optical images is beneficial to various earth science fields, such as image fusion [5], image segmentation [6], change detection [7], and three-dimensional reconstruction [8]. SAR and optical image registration is the basis of the joint interpretation of these two types of images. However, the significant nonlinear radiation difference (NRD) and geometric difference between SAR and optical images (shown in Figure 1), as well as the strong speckle noise carried by SAR images [9], make the registration of SAR images and optical images a difficult task.

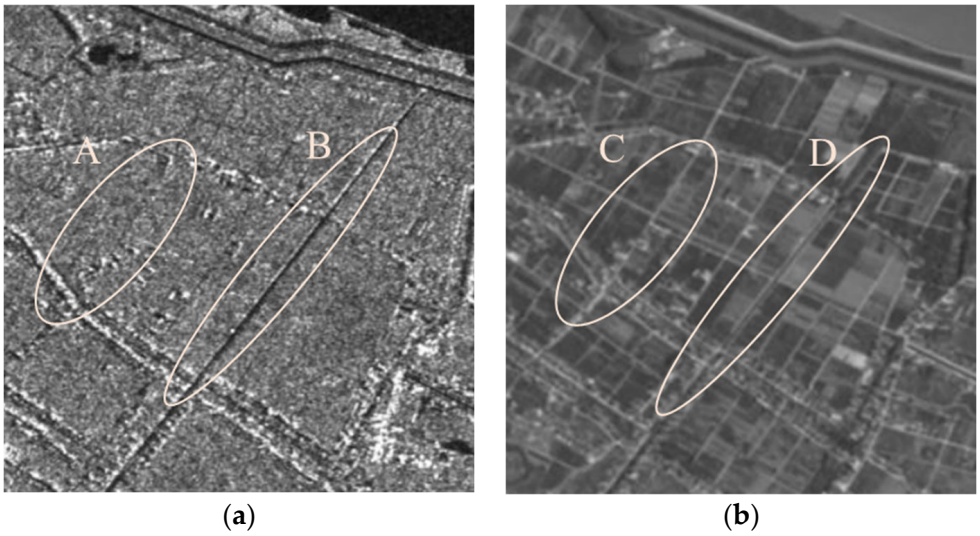


Figure 1. SAR image and optical image of the same area. (**a**) SAR image; (**b**) Optical image. Notice that region A correspond to region C, and region B correspond to region D.

In past decades, a variety of SAR and optical image registration methods have been proposed, which can be mainly classified into area-based methods and feature-based methods. Area-based methods usually cut SAR or optical images into patches, find correspondences between the images by continuously moving the position of the patch on the corresponding images, and finally, search for the maximum similarity based on some criteria [10]. Depending on the different similarity criteria, area-based methods can be classified as the sum of square difference (SSD) [11]-based methods, normalized cross-correlation (NCC) [12]-based methods, etc. Most of these methods rely on the intensity information and gradient information of the images, which are vulnerable to the NRD [13]. To overcome the NRD and adequately acquire salient structures in the images of the different modalities, a dense similarity criterion with a pixelwise description [14] has been proposed. However, area-based methods still tend to be computationally inefficient and are not rotationally invariant [15].

In general, feature-based methods are more robust than area-based methods, which find correspondences by detecting repeatable and salient keypoints, extracting robust descriptors, and matching descriptors. Among them, scale-invariant feature transform (SIFT) [16] has the performance of scale, rotation, and translation invariance and works well in various fields, such as optical image registration. However, in SAR and optical image registration, SIFT cannot achieve a satisfactory result due to NRD. Based on SIFT, researchers have proposed some detectors and descriptors adapted to SAR imaging characteristics, such as the SAR-SIFT [17], position-scale orientation-SIFT (PSO-SIFT) [18], and so on. Still, the feature detection and description of these manual detectors and descriptors are relatively sparse, which can only extract low-level semantic information.

With the booming development of deep learning, neural networks have been increasingly used in the field of target detection, classification, and feature recognition in SAR images [19]. Deep learning-based methods have the advantage of being able to extract high-level semantic features, which fits well with the need of SAR and optical image registration for structural information. Better feature descriptions can be obtained with the help of deep representations of the neural networks. The Siamese fully convolutional network (SFcNet) [20] maximizes the distance between positive and negative samples and learns the deep feature representations of the image patches. The deep dense feature network (DDFN) [21] also uses convolution networks to extract both high-level semantic features and low-level fine-grained information.

The neural networks also facilitate the feature detection stage to acquire structural information. Handcraft detectors usually consider only a few local pixels, for which the

detected keypoints do not represent higher-level structural information. Moreover, after the detection is completed, the feature description only focuses on local areas around the keypoints, and the global spatial information of the images is discarded. To solve this problem, the MAP-Net [22] first performs joint detection and description with a network, which uses an attention module to select the key features from the dense structure feature maps. In addition, a two-stage "detection + matching" approach has been proposed [23], in which the detection part also uses the keypoint detection network to obtain repeatable keypoints.

However, these deep learning-based detectors utilize only the structural information of a single image, which means that the keypoints are extracted without the knowledge of another image, and then the structure and shape features of the image pair are underutilized. It is noted that in realistic complex registration tasks, a human vision system would look forward and backward at both images of an image pair to determine the appropriate keypoints. As shown in Figure 1, when a human vision system is searching for keypoints in Figure 1b, it observes that region B in Figure 1a is very salient, while region A in Figure 1a is almost obscured by noise, and thus, tends to increase the attention to region D in Figure 1b and decrease the attention to region C in Figure 1b accordingly, so as to obtain more repeatable and distinctive keypoints. This means that while determining the reliable keypoints, not only the structure features in one image should be considered, but also the features in the corresponding location of the other image.

Inspired by the human vision system, we tried to determine more reliable keypoints for a single image based on the corresponding features of both images of the image pair. However, modeling this dependency between the features of the corresponding regions is a challenge since, in practice, the coordinates of the image pairs to be aligned may not correspond to each other. In this paper, a co-attention matching module (CAMM) is proposed, which models long-range dependencies of features in the image pairs based on the co-attention, and then uses the corresponding structure features to correct the keypoint feature maps. Traditionally, this was not feasible, but the flourishing of neural networks has made this possible [24]. On the other hand, the georeferenced information could be used to align images coarsely, which reduces the relative coordinate differences of the images and makes modeling easier.

The main contributions of this paper are summarized as follows:

1. A pair-based deep learning registration framework is proposed. The previous methods are image-based or patch-based, which extract the keypoints by focusing only on a single image or a patch, while a pair-based approach is proposed in this paper. Firstly, deep and dense structure features are extracted by DCNN from individual images in a joint detect-and-describe way, and then the detected keypoint candidates are corrected based on the structure features of the corresponding areas.
2. A CAMM for integrating the structure feature maps of both images is proposed. A CAMM is proposed for exploiting the structure features of both images to generate reliable keypoint feature maps. First, the dependencies among all the corresponding features of an image pair are computed based on the co-attention mechanism. Then, the keypoint features in the other image are incorporated into a complementary feature map, according to the attention scores for the keypoint feature map to be extracted.
3. A sampling strategy for SAR and optical image descriptor learning is proposed. To better train the Siamese network, new sampling strategies for the descriptors to learn the modality-invariant representations are proposed. Considering the imaging mechanism differences between the SAR and optical images, the pixels that are semantically corresponding in the image pair may not be corresponding in the spatial position of the images, so the positive samples are selected from the neighborhoods of the corresponding pixels, instead of directly adopting the patches centered at the corresponding pixels as the positive samples. Moreover, patches randomly sampled from the whole image are added to the negative samples as distractors in order to augment the distinguishing ability of the descriptor.

2. Related Works

2.1. Joint Image Keypoint Detection and Feature Description

Handcraft feature-based methods follow the detect-then-describe pattern, while the neural networks-based methods make joint detect-and-describe pattern possible. The learned invariant feature transform (LIFT) [25] is the first end-to-end feature detection and description method that learns descriptors, detectors, and orientation estimators using contrastive loss and hard mining strategies. Superpoint [26] uses a multi-stage self-supervised method for learning the detectors and descriptors, which utilizes both the ability of dense description and the practicality of sparse output in a learnable framework. The authors of the detect-and-describe network (D2-Net) [27] point out that conventional detectors under large intensity differences only consider small areas, which are not conducive to embedding the higher-level structural features. They propose a keypoint detector depending on the high-level features encoded in the descriptors, which shares all parameters among the detection and description networks and uses a joint formulation that simultaneously optimizes for both tasks. As a result, better robustness of image matching under extreme appearance changes is achieved. The authors of the repeatable and reliable detector and descriptor (R2D2) [28] claim that repeatability and reliability of the feature points are not consistent and need to be considered separately, and then add a separate reliability mask to obtain repeatable and simultaneously reliable keypoints.

The application scenarios of the above methods are mainly optical images. In the field of the registration of SAR images and optical images, MAP-Net [22] proposed an image-based deep matching method for the first time to learn detectors and descriptors jointly, which achieved outstanding results. Overall, the joint detection-and-description pattern creates dense high-level feature maps for both detection and description, and is worth following.

2.2. Attention Mechanism

The attention mechanism is an important technique in modern deep learning, which performs reasoning and decision making by selectively focusing on the key parts given in the contextual information. In the attention mechanism, the attention score is derived using the similarity of the key and query features, which is used to determine the weight of the value features. An attention mechanism has been applied to fields such as natural language processing, computer vision, and recommender systems with great success. Among them, the most famous work is the Transformer [29], which utilizes a self-attention mechanism.

In the field of SAR and optical image registration, the use of an attention mechanism has been a hot topic in recent years. The MAP-Net [22] proposed an attention block to select key features, and a matching network based on self-attention was proposed [23]. However, these methods still only consider a single image for detection, which is not enough for extracting reliable keypoints for the challenging registration. Unlike all of them, this paper proposes a CAMM, which achieves the structure information exchange between the pair of the images to be aligned through the attention mechanism early in the stage of detection.

3. Methodology

3.1. Overall Framework

To solve the problem of SAR and optical image registration, a dense feature description and sparse feature extraction network architecture is proposed in this paper, as shown in Figure 2. The proposed co-attention matching network (called the CAM-Net) extracts a series of reliable sparse keypoints and robust descriptors from the input image pairs.

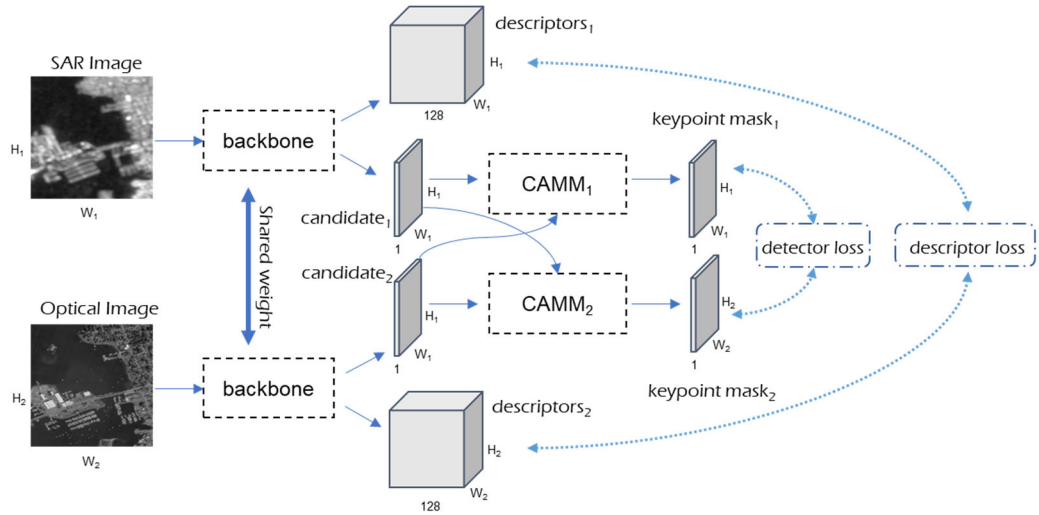


Figure 2. The overall architecture of the proposed CAM-Net.

First, to fully extract the structure features, the images are fed into a fully convolutional backbone network to obtain high-resolution dense feature maps. Specifically, a Siamese network is used, through which the input image is embedded into the dense descriptor feature map and the candidate feature map. Given the hierarchy property of the neural network, each of the descriptors at (i, j) of the feature maps correspond to a small patch centered at (i, j) of the original input image. In the training process, the corresponding patches in the two images should be as close as possible in the feature space, while the non-corresponding patches in the feature space as far away as possible.

Secondly, directly utilizing the dense features for registration is computationally expensive, so the backbone network simultaneously outputs the candidate masks for the subsequent feature selection. Moreover, it is noticed that the structural information in the corresponding image can also assist in extracting the distinctive structure features, so we propose a pair-based image keypoint extraction network module (i.e., the CAMM). The keypoint masks are generated by the CAMM, and then would be applied to the dense descriptors at the inference stage to obtain a series of reliable sparse keypoints and descriptors for efficient and accurate feature matching.

The structure of the proposed backbone network is shown in Figure 3, where 3×3 Conv means a 3×3 convolution of 1-pixel padding, followed by a batch normalization layer and a ReLU layer, which is used to keep the feature map the same size as the original input image. The number after Conv represents the amount of feature map output by the convolution layer. To increase the perceptual field of the convolution kernel, the dilated convolution and the successive convolution are used. Dilation 2 means using 2-pixel-sized dilated convolution to increase the perceptual field while maintaining the resolution. Successive 3 means repeating the convolution three times without using the ReLU in the middle convolution layers, which is used to imitate a larger convolution kernel.

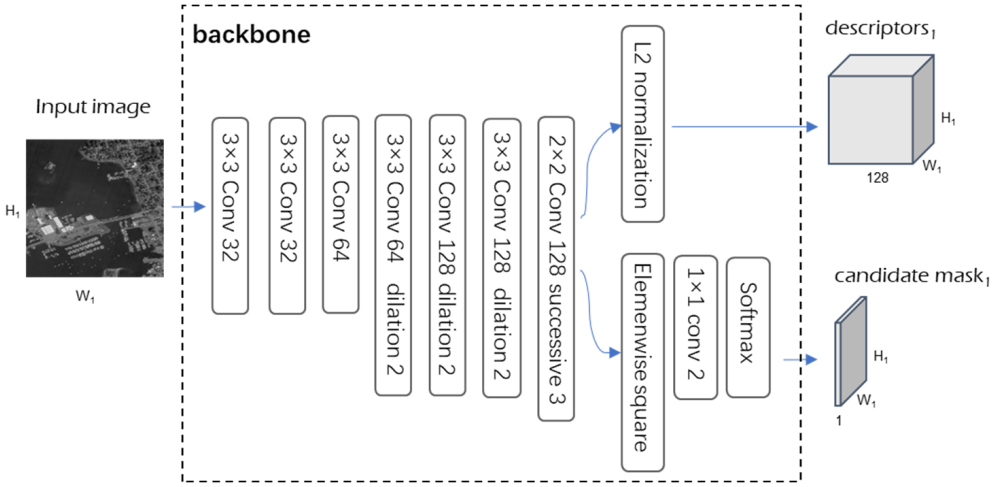


Figure 3. The structure of the proposed backbone network. Conv: convolution of 1-pixel padding, followed by a batch normalization layer and a ReLU layer; dilation 2: the convolution layer has a dilated rate of 2; successive 3: the convolution layer repeats convolution 3 times.

3.2. CAMM for Pairwise Structural Information

It is beneficial for the two input images to integrate each other's salient features to find their own valid keypoints under a strong NRD. However, the location of the relevant or corresponding features in the other image is unknown. For this reason, a CAMM is proposed to model the dependencies of the features at different locations in the different images with an attention mechanism.

As shown in Figure 4, the masks candidate$_1$ and candidate$_2$ generated in the feature extraction of the backbone include the salient feature responses of each pixel point in two images, and the CAMM$_1$ utilizes the structural information in mask candidate$_2$ to assist mask candidate$_1$ in generating a more reliable keypoint mask$_1$. Specifically, in order to generate multi-scale information and improve the computation efficiency, the two candidate masks obtained from the backbone are first down-sampled by half to obtain the feature map G and feature map H. The structural information in feature map H needs to be corrected by feature map G, and then the dependencies A_{ij} for each feature g_i in feature map G on features h_j in feature map H are represented by the attention score, which is given by

$$A_{ij} = \frac{\exp(g_i^T h_j)}{\sum_k \exp(g_i^T h_k)} \tag{1}$$

Then, the correction effect of each feature h_j on the location i can be characterized as

$$\hat{g}_i = \sum_j A_{ij} h_j \tag{2}$$

Let the complementary feature map for the feature map G be $\hat{G}$, which comprises $\hat{g}_i (i \in H_1' \times W_1')$. To fuse the structural information of different modalities, $\hat{G}$ and G are cascaded and convolved with a 1×1 kernel, and then up-sampled to recover the original image size. In order to reserve more detail information about the images and improve the accuracy of the registration, the resultant feature map is cascaded and fused with the original candidate$_1$ by a skip connection, which imitates the network structure of the U-Net [30]. The Softmax step is to obtain the normalized detection keypoint mask$_1$.

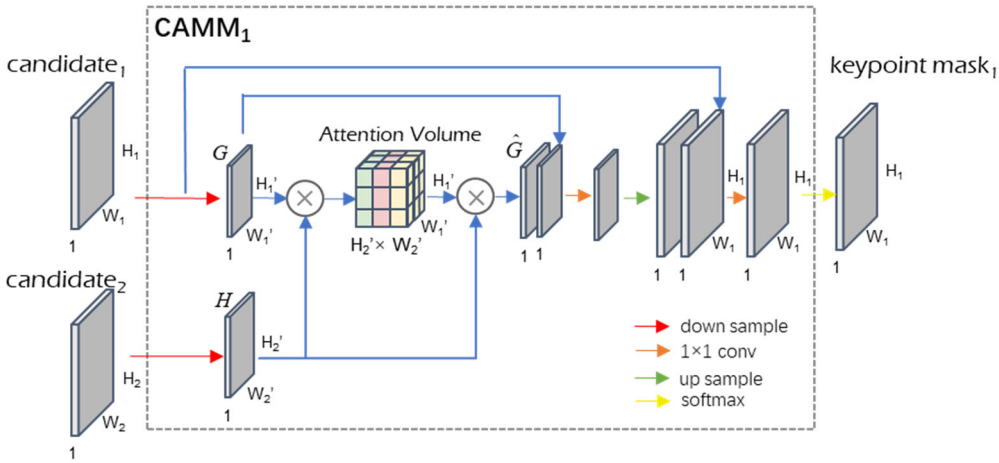


Figure 4. The structure of the proposed CAMM_1.

3.3. Loss Function

3.3.1. Detector Loss

Let I and I' be a SAR image and an optical image in a pair of pre-registered images, whose keypoint masks are S and S', respectively, with the coordinate $x = (i, j)^T$ of the pixel point S_{ij} in keypoint mask S corresponding to the coordinate $x' = (i', j')^T$ of the pixel point S'_{ij} in keypoint mask S'. To generate geometrically diverse data and obtain dense matching labels at a low cost, the self-supervised training is used in practice, i.e., the pre-aligned image pairs are subjected to a known homography transformation matrix H, so the dense correspondence can be characterized as

$$\begin{bmatrix} x \\ 1 \end{bmatrix} = H \begin{bmatrix} x' \\ 1 \end{bmatrix} \tag{3}$$

Since the local maxima of the two keypoint masks are selected as keypoints, to make the keypoints of the two masks correspond to each other, we should make the positions of the local maxima of the two keypoint masks correspond to each other. For this purpose, the keypoint masks are divided into a number of overlapping $N \times N$ patches P, and the similarity of them is measured by the cosine similarity $\cos im(\cdot)$. Let S'_H be the keypoint mask S' after a homography transformation H, the following loss is defined to maximize the cosine similarity between the corresponding patches:

$$L(I, I', H) = 1 - \frac{1}{|P|} \sum_{p \in P} \cos im(S[p], S'_H[p]) \tag{4}$$

On the other hand, to prevent the detection mask from overfitting into a constant during training, the regular term is added:

$$\text{Reg}(I) = 1 - \frac{1}{|P|} \sum_{p \in P} \left(\max_{(i,j) \in p} S_{ij} - \text{mean}_{(i,j) \in p} S_{ij} \right) \tag{5}$$

The total loss function for the detector is characterized as

$$L_{\text{det}}(I, I', H) = L_{\cos im}(I, I', H) + \frac{1}{2}(\text{Reg}(I) + \text{Reg}(I')) \tag{6}$$

3.3.2. Descriptor Loss

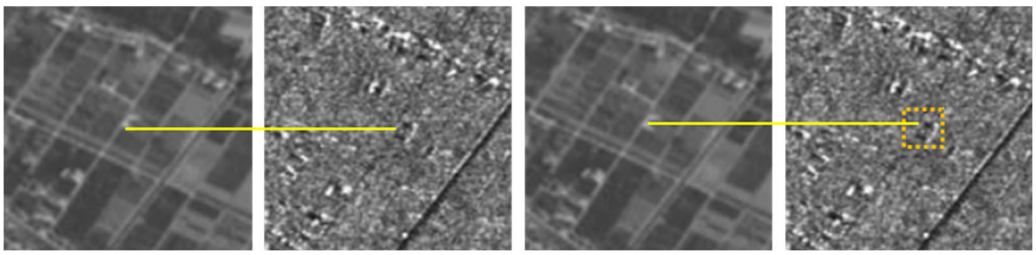
It is unfeasible to acquire a ground truth for the supervised learning of descriptors. Moreover, if a hand-crafted descriptor is used as the ground truth of the learning, the learned network is only an imitation of that manual descriptor. As a result, the descriptor should be trained in a self-supervised manner on a pre-registered image pair dataset. And the only restriction of the descriptor is that the corresponding patches in an image pair should share similar descriptors, while the non-corresponding patches obtain dissimilar descriptors. This is well suited to be implemented with the average precision (*AP*) loss [31], which is reformulated from the *AP* as a soft assignment version, and thus, differentiable. The formula for the descriptor loss is

$$L_{desc} = \frac{1}{B}\sum_{ij} 1 - AP(p_{ij}) \tag{7}$$

where B is the number of the query descriptors of patches $\{p_{ij}\}$ at location (i, j).

While in training, the dense descriptor feature maps should be sampled to improve the computational efficiency. Usually, $H \times W/(8 \times 8)$ points are randomly sampled from the query image as the query descriptors, while the positive samples generally are sampled from the exactly spatially corresponding location at the other image, and the negative samples are sampled from the neighborhood of that location. However, this sampling strategy could be improved.

For the positive samples, it was noticed that the spatially corresponding pixel points on the two images did not necessarily semantically correspond to each other because of the difference between the optical and SAR imaging systems. Therefore, a sampling strategy for the positive samples is proposed. We computed similarities between the query descriptor and each descriptor of the neighborhood centered within 3 pixels from the spatially corresponding position in the second image, and the descriptor with the greatest similarity to the query descriptor was selected as the positive sample. This method is called the positive sample pooling (PSP), which is shown in Figure 5.



(**a**) (**b**)

Figure 5. Sampling strategies for the positive samples. (**a**) Take the patch centered in the spatially corresponding pixel as the positive sample; (**b**) select the positive sample from the center patches in the neighborhood of the spatially corresponding pixel.

As for the negative samples, usually, they are sampled at a distance of 5 to 8 pixels away from the spatially corresponding position. However, we found this was not enough for training the proposed network. In this paper, a negative sample augmentation (NSA) strategy is proposed to strengthen the distinguishing ability of the descriptor. In addition to the negative samples sampled at a distance of 5 to 8 pixels away from the corresponding position, M patches are randomly sampled from the second image again as distractors to augment the negative samples, with the patches that have been already sampled as the positive samples excluded. In the experiments, it was found that M taking the number of the positive samples was both effective and computationally efficient. To the best of our knowledge, this sample mechanism has not been previously applied in the research of SAR and optical image registration.

3.4. Model Inference

The inference of the proposed CAM-Net is shown in Figure 6. The keypoint $mask_1$ is first processed using the NMS, and then the pixels whose responses are above a certain threshold are considered the keypoints. In the experiment, it was found that a keypoint threshold of 0.4 achieved good results.

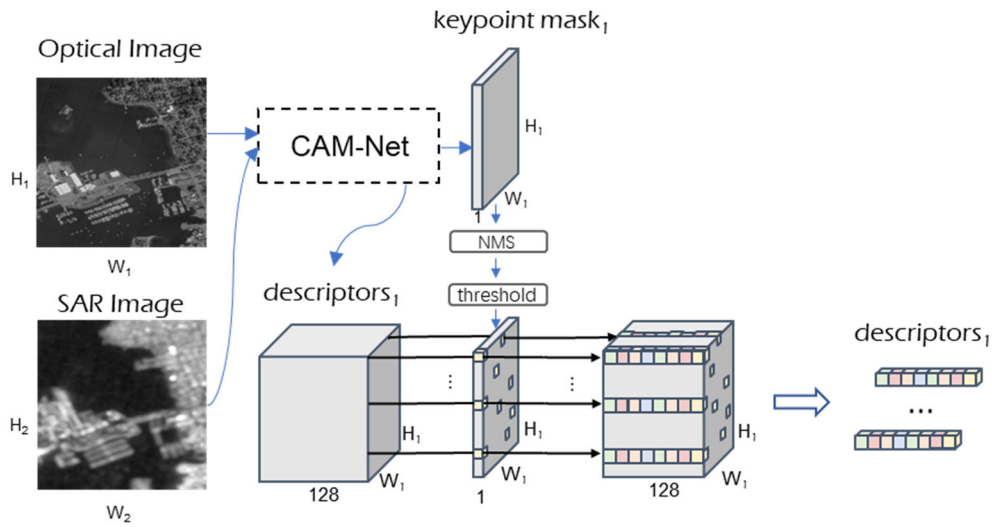


Figure 6. Illustration of the inference of the proposed CAM-Net.

Furthermore, to improve the detection accuracy and generate multi-scale features, input images are down-sampled during the inference, and the keypoint mask generated from the inference of the several down-sampled multi-scale images is integrated to obtain the final keypoints.

To complete the registration, after the network extracts the keypoints and descriptors of both images, the descriptors are matched using mutual nearest neighbor matching. Then, random sample consensus (RANSAC) is used to filter the correspondences and estimate the homography transformation matrix.

4. Experiment and Analysis

4.1. Experimental Datasets

The training dataset included 5000 pre-registered image pairs from the SEN1-2 dataset [32]. To reduce the difficulty of the model convergence, a phased training method was used. The model was first trained with the original dataset, and then trained with the dataset with data augmentation to acquire the homography transformation invariance.

The test dataset comprised dozens of the pre-registered image pairs from the SEN1-2 datasets (with the training data excluded), as well as the dataset provided in [23], and it included urban areas, farmlands, mountains, rivers and snowfields, which can reflect the complex NRD between the SAR and optical images. While testing, the pre-registered dataset was transformed to test our model, and then the ground truth coordinates could be obtained by recording the transformation matrix.

For this experiment, an Intel i5-11400F CPU and a RTX3060 GPU was used. In the training process, the Adam optimizer was used with an initial learning rate of 0.0001. It should be noted that the padding mode of all the convolution operations needs to be set to circular or mirror (symmetric or reflect) in order to overcome the "ghost shadow" [33] in the training process.

4.2. Evaluation Metrics

(a) Repeatability. This metric is used to determine the repeatability of the keypoints detected in the image pair. Suppose that image 1 is detected with N_1 keypoints and the set of the coordinates is $\{x\}$, while image 2 is detected with N_2 keypoints and the set of the coordinates is $\{\hat{x}\}$. The points in image 1 are called the inner points if they satisfy the following equation:

$$\min_j \|x - \hat{x}_j\| \leq \varepsilon \tag{8}$$

where ε is the predefined threshold, which is 3 pixels in this paper. Let the number of the inner points in image 1 be n_1 and the number of the inner points in image 2 be n_2. The formulation for repeatability is

$$\text{Rep} = \frac{n_1 + n_2}{N_1 + N_2} \tag{9}$$

(b) Localization error (*LE*). *LE* is used to calculate the error between the coordinates of the inner point in (a) and the ground-truth coordinates in order to further evaluate the accuracy of the keypoints extracted by the detector.

$$LE = \frac{1}{N} \sum_{i \in n_1} \min_j \|x_i - \hat{x}_j\| \tag{10}$$

(c) Mean matching accuracy (MMA) [34]. MMA is also the percentage of the inner points, but unlike (a), the inner points here are defined as those selected by the homography transformation model estimated by the RANSAC. This metric measures the accuracy of the correspondences generated by mutual brute-force matching of the descriptors, reflecting the performance of both the detectors and descriptors.

(d) Nearest neighbor mean average precision (NN mAP) [26]. This criterion shows the detection ability of detectors and the distinguishing ability of descriptors. It is computed by measuring the area under the curve (AUC) of the precision–recall (PR) curve, using the nearest neighbor matching strategy.

(e) Average corner error (ACE) [35]. This metric integrates the performance of the entire keypoint detection, feature description, and feature-matching pipeline. It is defined as the distance between real coordinates and estimated coordinates of the four corners, which are computed separately according to the ground-truth homography matrix and the estimated one by the RANSAC.

4.3. Experimental Results and Analysis

The performance of the proposed CAM-Net was evaluated with state-of-the-art methods (i.e., SIFT, SAR-SIFT, RIFT, Superpoint, and R2D2) on SAR and optical image pairs, with both geometric difference and radiometric difference. To thoroughly analyze the performance of the network, its detection capability, description ability, and its overall registration capability were verified respectively. Specifically, in the implementation of the comparison algorithms, for SIFT, the default OpenCV implementation was used. For RIFT and SAR-SIFT, its online public code was used. For the two deep learning algorithms, Superpoint and R2D2, the open-source implementations were also used. For the test dataset, dozens of the images with a variety of terrain types with complex geometric distortion and radiometric differences were used to verify the registration performance.

4.3.1. Detection Capability

To validate the detection capability of the proposed CAM-Net, the repeatability and LE were measured, where the distance threshold in the repeatability was set to 3 pixels.

In Table 1, it is seen that the LE of the keypoints detected by different algorithms are similar, so the repeatability can be a fair and valid indication of the effectiveness of

the detected keypoints. The handcrafted methods have better LE metrics than the deep learning-based methods because the learning-based methods still use a loose loss function for keypoint detection to reduce the difficulty of convergence. Among all the handcrafted methods, RIFT is the best designed one, with NRD resistance and rotation invariance, and thus, it has the best LE metric. The experimental results show that the proposed CAM-Net has the highest repeatability because it considers the feature information of the corresponding structures in both images of the image pairs. For SIFT and SAR-SIFT, the two traditional methods can extract a large number of keypoints, but the repeatabilities of them are very low, since they are hand-designed. RIFT solves NRD well and detects keypoints with high repeatability; however, it is still less repeatable than the proposed CAM-Net, probably because RIFT is designed manually and has limited generalization capabilities for the multimodality, and the scenario in this test is especially complex. In contrast, the deep learning-based methods can extract high-level structural information, and all performed well, with the proposed CAM-Net achieving the best results.

Table 1. The comparison of the detection capability of different algorithms.

	SIFT	SAR-SIFT	RIFT	Superpoint	R2D2	CAM-Net
LE	1.9149	1.8319	**1.7311**	1.9429	1.9018	1.9619
Repeatability	0.1309	0.1402	0.4238	0.3914	0.3626	**0.4338**

4.3.2. Description Capability

To verify the description capability of the proposed CAM-Net, the MMA and NN mAP were measured. Based on the results in Table 2 and Figure 7, it is seen that the proposed CAM-Net achieved large leads in the MMA and mAP. On one hand, this shows that the proposed CAM-Net descriptor is effective in its ability to describe the corresponding regions consistently in both SAR and optical modalities. On the other hand, the feature description is also affected by the performance of the detectors. The more distinctive the keypoints are, the more the descriptors can extract structural information from the patches centered at the keypoints, so the results also show that the proposed CAM-Net is more effective in extracting reliable and repeatable keypoints in the detection phase.

Table 2. The NN mAP of the different algorithms.

	SIFT	SAR-SIFT	RIFT	Superpoint	R2D2	CAM-Net
NN mAP	0.0011	0.1692	0.0595	0.1796	0.1794	**0.3090**

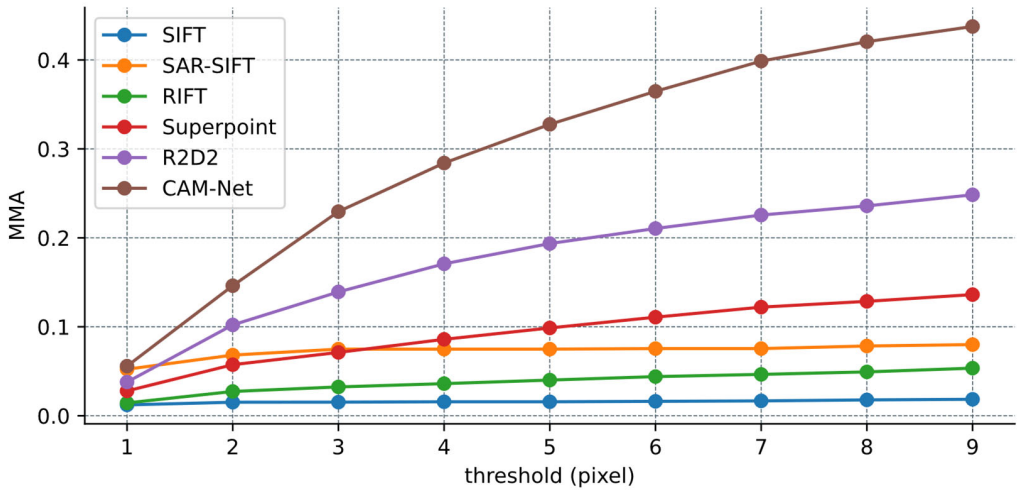


Figure 7. The MMA under a series of thresholds of the different algorithms.

It should be noticed that RIFT achieves the similar results to the proposed CAM-Net in the repeatability, but it lost to the CAM-Net in the mAP and MMA by a large margin, which implies that solely optimizing repeatability does not lead to a better registration result. Although RIFT detects repeatable keypoints in the image pairs to be aligned, these points are not necessarily suitable for subsequent registration. If the detected keypoints are not located in the regions containing salient structural features, they can instead hinder the mutual nearest neighborhood matching.

As a non-learning method, SIFT can hardly find correspondences accurately under a strong NRD. SAR-SIFT is a variant of SIFT that considers the imaging characteristics of SAR images, and its descriptor performance is improved. The learning-based methods are overall better than the manual methods due to their strong nonlinear learning capability. Superpoint uses simulated data as well as homography adaptation to train the network, and R2D2 uses optical datasets with complex scenes and large illumination differences for training. Both of them have strong generalization capabilities. However, they do not consider the modality differences between the two images, and thus, their description performances are lower than the proposed methods under a complex NRD in SAR and optical image registration.

4.3.3. Overall Registration Performance

The comprehensive capability of different algorithms was compared by computing the metric ACE and visualizing the registration results. To obtain the ACE, we obtained the keypoints and their descriptors of the two images by different algorithms and performed a mutual brute-force match on the descriptors to obtain preliminary correspondences, and then we used OpenCV implementation (findHomography() with RANSAC) to compute the final homography matrix of all correspondences.

The ACE metric obtained from the experiments is shown in Table 3, where the ACE above 100 pixels represents a complete failure of the registration. According to the results, the proposed CAM-Net outperformed all the other algorithms in the test, which indicates that the proposed CAM-Net is very effective in SAR and optical image registration. SIFT and SAR-SIFT, as two hand-designed methods, fail in the registration test. RIFT has the modality-invariant feature representation ability between the multimodal images; however, as a non-learning method, it can hardly extract robust features for SAR and optical images. In the learning-based methods, the proposed CAM-Net exchanges structural information between pairs of images and outperforms Superpoint and R2D2, which suggests that the more distinctive and repeatable keypoints can improve the overall performance of the registration.

Table 3. The ACE of the different algorithms.

	SIFT	**SAR-SIFT**	**RIFT**	**Superpoint**	**R2D2**	**CAM-Net**
ACE	-	-	49.7911	37.0389	27.0098	**7.1459**

To further analyze the performance of these algorithms, a comparison of the matching results for several images is shown in Figure 8. The distance threshold for the accurate matches of keypoints was set to 3 pixels, and the accurate matches are indicated by the blue line, while the incorrect matches are indicated by the red line. It is seen that the proposed CAM-Net generated a large number of correct matches which densely covered the whole images, given the dense feature description. In contrast, SIFT and SAR-SIFT extracted sparse features, which rarely corresponded. Superpoint and R2D2 are also deep learning methods that depict images densely; however, they were infeasible in extracting structural information in the context of a strong NRD. In contrast, the proposed CAM-Net integrated the information among image pairs, considered both the salience and repeatability of the keypoints, and then obtained more effective correspondences.

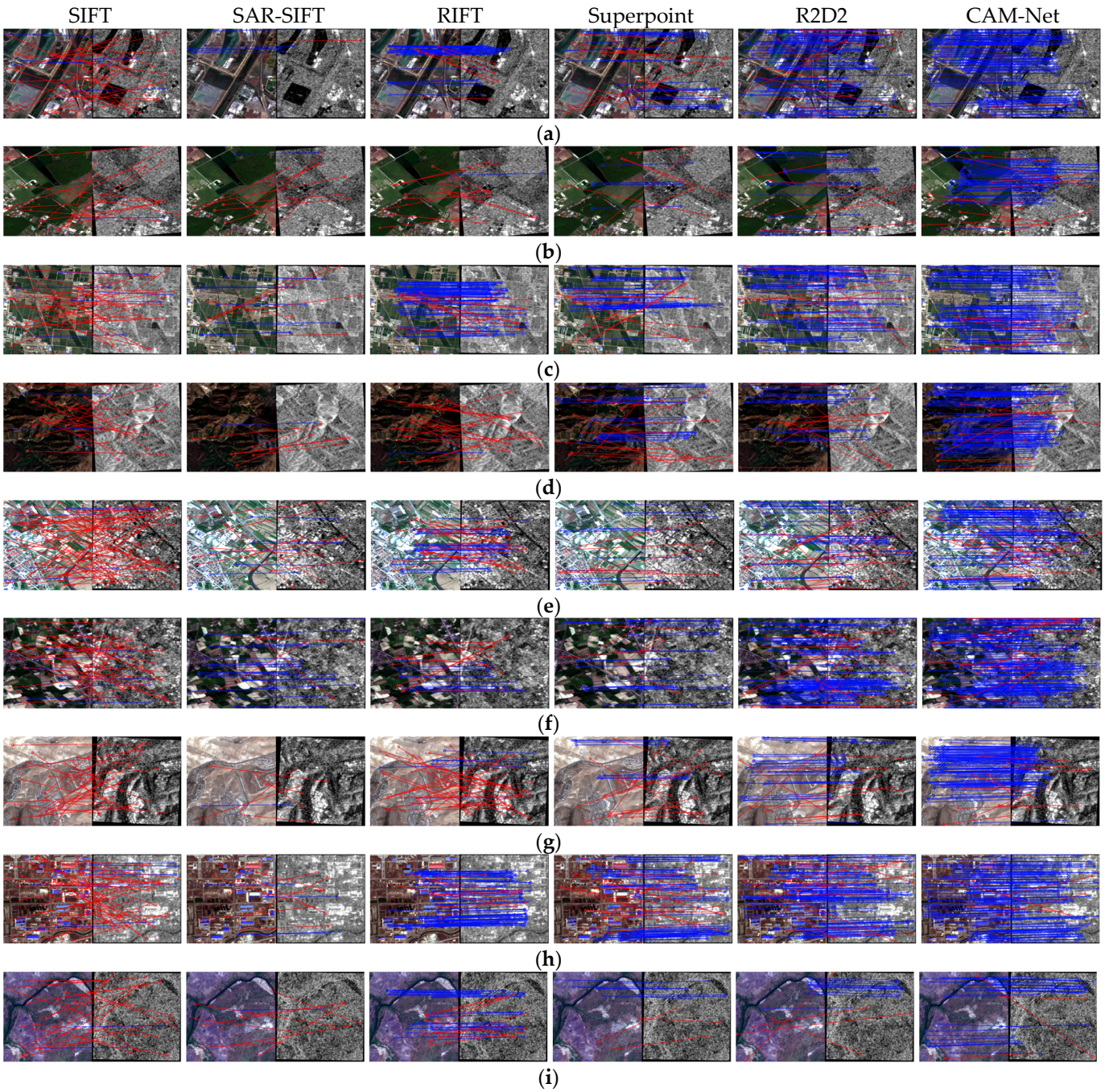


Figure 8. Visualization of the matching results. (**a**–**i**) Registration results of the different scenes. The blue lines indicate the correct matches, while the red lines indicate the wrong matches. CAM-Net generates a large number of the correct matches, which densely cover the whole images, given the dense feature description. SIFT and SAR-SIFT extract the sparse features, which rarely correspond. Well-designed RIFT generates dense matches because of its NRD resistance and rotation invariance. Superpoint and R2D2 are also deep learning methods that depict images densely; however, they are infeasible in extracting structural information in the context of the strong NRD.

4.3.4. Component Analysis

In this section, the roles of specific components of the network are evaluated. For the ablation experiment of the proposed CAMM, the results of the model tests with the CAMM and without the CAMM in inference were compared. For the ablation experiment

of the sampling strategies, the performance of the networks trained with different sampling strategies was compared.

(a) CAMM

Since it is difficult to visualize the CAMM directly, a registration case is used to illustrate the effect of the CAMM. As shown in Figure 9, Figure 9a,d shows the optical image and SAR image of the same snowy scene, respectively. In the optical image, the white snow area reflects almost all the sunlight and has a low feature response, for which the structural information is mainly distributed in the bottom half. In contrast, the snowy terrain in the SAR image is even more prominent than the bottom part of the image, so the keypoints in the SAR image are concentrated more in the snowy part, while in the optical image. few keypoints are detected there. Consequently, the widely different detected keypoints lead to unsatisfactory matching results, as shown in Figure 9c.

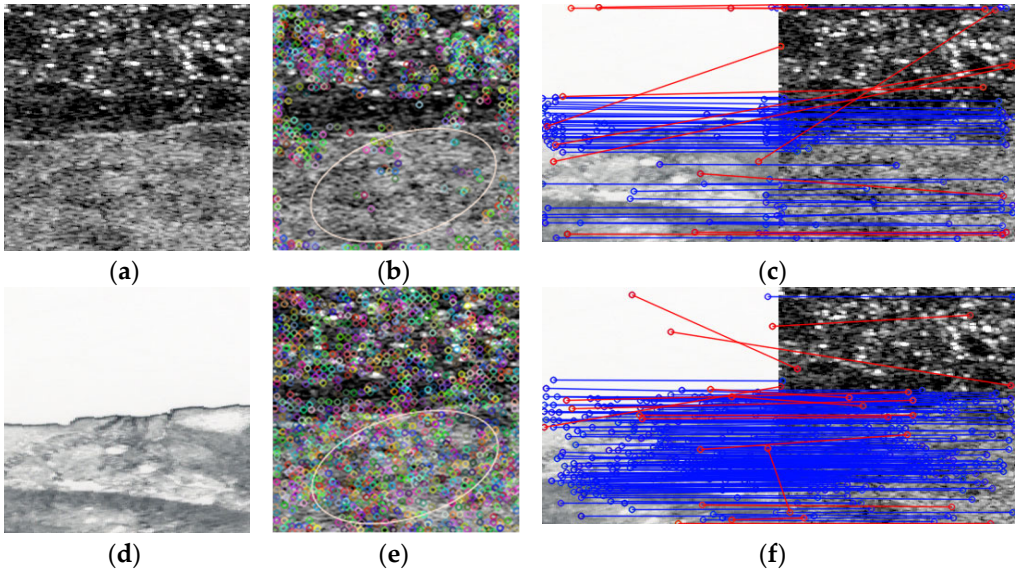


Figure 9. The visualization of the effect of the proposed CAMM. The blue lines indicate the correct matches, while the red lines indicate the wrong matches. (**a**) SAR image of the snowy area; (**d**) Optical image of the same area; (**b**) keypoints extracted from the SAR image by the networks without the CAMM; (**e**) keypoints extracted from the optical image by the networks with the CAMM; (**c**) matching result of the SAR image and optical image without the CAMM; (**f**) matching result of the SAR image and optical image with the CAMM.

However, with the CAMM, the optical image and SAR image can correct the keypoint feature maps by the co-attention to each other and obtain more reliable keypoints. As shown in Figure 9e, more keypoints are distributed in the mountainous area in the SAR image with the CAMM, and finally, more successive matching results were achieved, as shown in Figure 9f. It should be noted that the numbers of the keypoints extracted in Figure 9c,f were artificially controlled to be the same.

To further illustrate the effectiveness of the CAMM, an ablation experiment was conducted. The results are shown in Table 4 and Figure 10, indicating that the network using the CAMM was significantly superior to the one without the CAMM in all metrics, which quantitatively proves the effectiveness of the CAMM.

Table 4. The results of the ablation experiment for the proposed CAMM.

	Repeatability	**LE**	**NN mAP**	**ACE**
CAM-Net	**0.4338**	1.9619	**0.3090**	**7.1459**
CAM-Net w/o CAMM	0.3839	**1.9293**	0.2476	20.5072

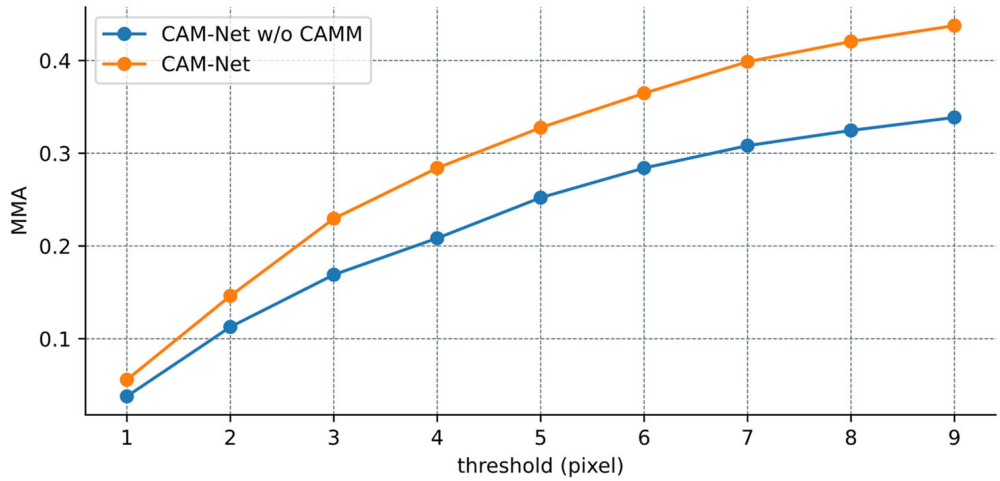


Figure 10. The MMA result of the ablation experiment of the proposed CAMM.

(b) Sampling strategies

To illustrate effectiveness of the proposed sampling strategies, an ablation experiment of the sampling strategies was conducted. As shown in Table 5 and Figure 11, PSP improved the description ability of the descriptor and further reduced the ACE. NSA provided sufficient and typical negative samples and PSP provided more accurate positive samples for the learning of the descriptors, both of which were beneficial for learning the modality-invariant feature representation. Overall, the contribution of NSA was more significant than PSP because PSP improved the accuracy of the selected positive samples to a certain extent, while NSA provided the necessary quantity of the negative samples.

Table 5. The results of the ablation experiment for the sampling strategies.

	PSP	NSA	Repeatability	LE	NN mAP	ACE
CAM-Net w/o PSP and NSA			0.2062	1.9891	0.0307	-
CAM-Net w/o NSA	√		0.2030	1.9886	0.0417	-
CAM-Net w/o PSP		√	0.3238	**1.9474**	0.2950	8.0344
CAM-Net	√	√	**0.4338**	1.9619	**0.3090**	**7.1459**

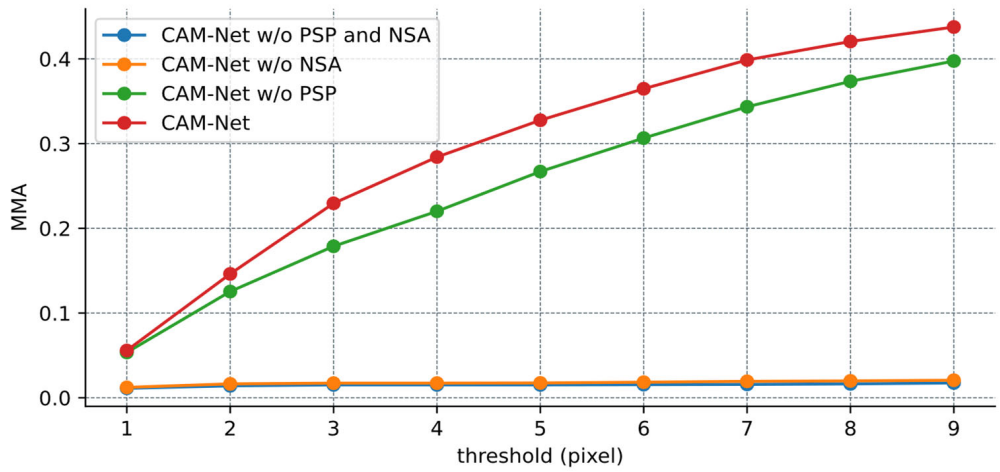


Figure 11. The MMA results of the ablation experiment of the different sampling strategies.

5. Conclusions

In this paper, a SAR and optical image registration method based on a CAMM is proposed, aiming to solve the problems of NRDs and geometric model differences between SAR and optical images. Traditional methods and deep learning methods often neglect the utilization of the information between image pairs when acquiring distinctive keypoints, and this paper extracted keypoints by integrating the structural information of the image pairs through a co-attention mechanism.

First, joint feature description and keypoint extraction were performed for two images to be aligned. Then, the CAMM was proposed to model the dependencies of the features. Finally, the structure feature maps of the two images were integrated to extract the keypoints, which made the extracted keypoints more reliable and repeatable, and facilitated their global registration. The experimental results verify the correctness and effectiveness of the proposed method. Compared to the traditional and deep learning methods, the proposed method had better repeatability, ACE, NN mAP and MMA in SAR and optical image registration. Thus, the proposed method provides an effective solution to the problem of SAR and optical image registration.

In further work, more prior information between the image pairs can be considered to further improve the accuracy and robustness of SAR and optical image registration. In addition, the proposed method has currently been performed on several SAR and optical image datasets, and it can be further transferred to other datasets and extended to more modalities.

Author Contributions: Conceptualization, J.C., H.X. and J.H.; methodology, J.C., H.X. and J.H.; software, J.C.; validation, J.C., H.X. and L.Z.; formal analysis, J.C., H.X. and G.W.; investigation, J.C., H.X. and L.Z.; resources, H.J. and J.H.; data curation, J.C. and H.X.; writing—original draft preparation, J.C. and H.X.; writing—review and editing, J.C., H.X. and G.W.; visualization, J.C., J.H. and H.J.; supervision, H.X. and G.W.; project administration, H.X. and G.W.; funding acquisition, H.X., J.H., H.J. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Guangdong Basic and Applied Basic Research Foundation (Grants No. 2021A1515010768 and No. 2023A1515011588), the Shenzhen Science and Technology Program (Grant No. 202206193000001, 20220815171723002), the Science and Technology on Near-Surface Detection Laboratory Pre-Research Foundation (Grant No. 6142414200607), the National Natural Science Foundation of China (Grant No. 62001523, No. 62203465, No. 62201614, and No. 6210593), and by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant No. 23lgpy45). Hongtu Xie is the corresponding author.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the editors and reviewers for their very competent comments and helpful suggestions to improve this paper. Moreover, the authors would like to thank the authors of [23] for the experimental dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xie, H.; Shi, S.; An, D.; Wang, G.; Wang, G.; Xiao, H.; Huang, X.; Zhou, Z.; Xie, C.; Wang, F.; et al. Fast Factorized Backprojection Algorithm for One-Stationary Bistatic Spotlight Circular SAR Image Formation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1494–1510.
2. Hu, X.; Xie, K.; Xie, H.; Zhang, L.; Hu, J.; He, J.; Yi, S.; Jiang, H. Fast Factorized Backprojection Algorithm in Orthogonal Elliptical Coordinate System for Ocean Scenes Imaging Using Geosynchronous Spaceborne-Airborne VHF UWB Bistatic SAR. *Remote Sens.* **2023**, *15*, 2215.
3. Jiang, X.; Xie, H.; Chen, J.; Zhang, J.; Wang, G.; Xie, K. Arbitrary-Oriented Ship Detection Method Based on Long-Edge Decomposition Rotated Bounding Box Encoding in SAR Images. *Remote Sens.* **2023**, *14*, 3599.
4. Xie, H.; Hu, J.; Duan, K.; Wang, G. High-Efficiency and High-Precision Reconstruction Strategy for P-Band Ultra-Wideband Bistatic Synthetic Aperture Radar Raw Data Including Motion Errors. *IEEE Access* **2020**, *8*, 31143–31158.
5. Kulkarni, S.; Rege, P. Pixel Level Fusion Techniques for SAR and Optical Images: A Review. *Inf. Fusion* **2020**, *59*, 13–29.

6. Wurm, M.; Stark, T.; Zhu, X.; Weigand, M.; Taubenboeck, H. Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69.
7. Sun, Y.; Lei, L.; Guan, D.; Kuang, G. Iterative Robust Graph for Unsupervised Change Detection of Heterogeneous Remote Sensing Images. *IEEE Trans. Image Process.* **2021**, *30*, 6277–6291.
8. Hartmann, W.; Havlena, M.; Schindler, K. Recent Developments in Large-Scale Tie-Point Matching. *ISPRS J. Photogramm. Remote Sens.* **2016**, *115*, 47–62.
9. Xiang, D.; Xie, Y.; Cheng, J.; Xu, Y.; Zhang, H.; Zheng, Y. Optical and SAR Image Registration Based on Feature Decoupling Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5235913.
10. Ye, Y.; Shan, J.; Bruzzone, L.; Shen, L. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2941–2958.
11. Inglada, J.; Giros, A. On the Possibility of Automatic Multisensor Image Registration. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 2104–2120.
12. Hel-Or, Y.; Hel-Or, H.; David, E. Fast Template Matching in Non-Linear Tone-Mapped Images. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 1355–1362.
13. Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; Guo, X. A Review of Multimodal Image Matching: Methods and Applications. *Inf. Fusion* **2021**, *73*, 22–71.
14. Ye, Y.; Bruzzone, L.; Shan, J.; Bovolo, F.; Zhu, Q. Fast and Robust Matching for Multimodal Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9059–9070.
15. Li, J.; Hu, Q.; Ai, M. RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Trans. Image Process.* **2020**, *29*, 3296–3310.
16. Lowe, D. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
17. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. SAR-SIFT: A SIFT-Like Algorithm for SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 453–466.
18. Ma, W.; Wen, Z.; Wu, Y.; Jiao, L.; Gong, M.; Zheng, Y.; Liu, L. Remote Sensing Image Registration with Modified SIFT and Enhanced Feature Matching. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 3–7.
19. Zhu, X.; Montazeri, S.; Ali, M.; Hua, Y.; Wang, Y.; Mou, L.; Shi, Y.; Xu, F.; Bamler, R. Deep Learning Meets SAR: Concepts, Models, Pitfalls, and Perspectives. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 143–172.
20. Zhang, H.; Ni, W.; Yan, W.; Xiang, D.; Wu, J.; Yang, X.; Bian, H. Registration of Multimodal Remote Sensing Image Based on Deep Fully Convolutional Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3028–3042.
21. Zhang, H.; Lei, L.; Ni, W.; Tang, T.; Wu, J.; Xiang, D.; Kuang, G. Optical and SAR Image Matching Using Pixelwise Deep Dense Features. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6000705.
22. Cui, S.; Ma, A.; Zhang, L.; Xu, M.; Zhong, Y. MAP-Net: SAR and Optical Image Matching via Image-Based Convolutional Network with Attention Mechanism and Spatial Pyramid Aggregated Pooling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1000513. [CrossRef]
23. Li, L.; Han, L.; Ye, Y. Self-Supervised Keypoint Detection and Cross-Fusion Matching Networks for Multimodal Remote Sensing Image Registration. *Remote Sens.* **2022**, *14*, 3599.
24. Wiles, O.; Ehrhardt, S.; Zisserman, A. Co-Attention for Conditioned Image Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15915–15924.
25. Yi, K.; Trulls, E.; Lepetit, V.; Fua, P. LIFT: Learned Invariant Feature Transform. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 457–483.
26. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-Supervised Interest Point Detection and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 337–349.
27. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8084–8093.
28. Revaud, J.; Weinzaepfel, P.; De Souza, C.; Pion, N.; Csurka, G.; Cabon, Y.; Humenberger, M. R2D2: Repeatable and reliable detector and descriptor. *arXiv* **2019**, arXiv:1906.06195.
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advanced Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
31. He, K.; Lu, Y.; Sclaroff, S. Local Descriptors Optimized for Average Precision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 596–605.
32. Schmitt, M.; Hughes, L.; Zhu, X. The Sen1-2 Dataset for Deep Learning in Sar-Optical Data Fusion. *arXiv* **2018**, arXiv:1807.01569.

33. Alsallakh, B.; Kokhlikyan, N.; Miglani, V.; Yuan, J.; Reblitz-Richardson, O. Mind the Pad--CNNs Can Develop Blind Spots. *arXiv* **2020**, arXiv:2010.02178.
34. Mikolajczyk, K.; Schmid, C. Scale & Affine Invariant Interest Point Detectors. *Int. J. Comput. Vis.* **2004**, *60*, 63–86.
35. Ye, Y.; Tang, T.; Zhu, B.; Yang, C.; Li, B.; Hao, S. A Multiscale Framework with Unsupervised Learning for Remote Sensing Image Registration. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622215.