remote sensing

MDPI

Article

Optical and SAR Image Registration Based on Pseudo-SAR Image Generation Strategy

Canbin Hu [1], Runze Zhu [1], Xiaokun Sun [1,*], Xinwei Li [1] and Deliang Xiang [1,2]

[1] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; canbinhu@buct.edu.cn (C.H.); 2021200817@buct.edu.cn (R.Z.); 2022210546@buct.edu.cn (X.L.); xiangdeliang@mail.buct.edu.cn (D.X.)

[2] Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China

* Correspondence: sunxk@mail.buct.edu.cn

Abstract: The registration of optical and SAR images has always been a challenging task due to the different imaging mechanisms of the corresponding sensors. To mitigate this difference, this paper proposes a registration algorithm based on a pseudo-SAR image generation strategy and an improved deep learning-based network. The method consists of two stages: a pseudo-SAR image generation strategy and an image registration network. In the pseudo-SAR image generation section, an improved Restormer network is used to convert optical images into pseudo-SAR images. An L2 loss function is adopted in the network, and the loss function fluctuates less at the optimal point, making it easier for the model to reach the fitting state. In the registration part, the ROEWA operator is used to construct the Harris scale space for pseudo-SAR and real SAR images, respectively, and each extreme point in the scale space is extracted and added to the keypoint set. The image patches around the keypoints are selected and fed into the network to obtain the feature descriptor. The pseudo-SAR and real SAR images are matched according to the descriptors, and outliers are removed by the RANSAC algorithm to obtain the final registration result. The proposed method is tested on a public dataset. The experimental analysis shows that the average value of NCM surpasses similar methods over 30%, and the average value of RMSE is lower than similar methods by more than 0.04. The results demonstrate that the proposed strategy is more robust than other state-of-the-art methods.

Keywords: pseudo-SAR; image generation strategy; registration

check for updates

Citation: Hu, C.; Zhu, R.; Sun, X.; Li, X.; Xiang, D. Optical and SAR Image Registration Based on Pseudo-SAR Image Generation Strategy. *Remote Sens.* **2023**, *15*, 3528. https://doi.org/10.3390/rs15143528

Academic Editor: José Fernández

Received: 1 June 2023
Revised: 28 June 2023
Accepted: 11 July 2023
Published: 13 July 2023

1. Introduction

Optical and synthetic aperture radar (SAR) images are two types of products formed by distinct sensors. Optical images result from the passive reception of naturally reflected light, while SAR images are generated by actively transmitting and receiving radar electromagnetic waves. The quality of optical images is heavily affected by cloudy conditions and night variations. In contrast, SAR is capable of observing the earth under various weather conditions and demonstrates strong performance during day and night. Therefore, it is necessary to jointly use the effective information from these two different imaging sensors. Thus, how to match the two kinds of images becomes the top priority. However, SAR images often exhibit unique manifestations, such as shadows, superimposition, and foreshortening, which are caused by special imaging mechanisms. These imaging disparities pose challenges to the registration of optical and SAR images.

In recent years, many researchers have proposed diverse approaches for matching optical and SAR images [1]. Presently, three primary frameworks for the image registration have been established [2]: area-based matching [3], feature-based matching [4], and deep learning-based matching methods [5].

Area-based matching techniques rely on template matching to quantify the similarity of image block templates, employing similarity metrics such as mutual information [6–8], normalized cross-correlation coefficient [9,10], and cross-cumulative residual entropy [11].

Although area-based matching methods tend to achieve higher accuracy than feature-based methods in homologous image registration due to their use of the gray-scale information of image regions, they are sensitive to the nonlinear radiometric differences between optical and SAR images. Additionally, the need to traverse the entire reference image in area-based registration methods results in computationally complexity. Consequently, there are some limitations to the registration of optical and SAR images using area-based methods [12].

Feature-based matching is also an important method framework in image registration. This method is mainly divided into four steps: keypoint extraction, feature descriptor construction, feature matching, and affine transformation. Generally, feature-based matching methods have a higher calculation speed than area-based matching methods. The most famous one is the scale-invariant feature transform (SIFT) [13] method. It has good effects on rotation, scaling, translation, and shall be robust to intensity changes and affine distortion. The different imaging mechanisms of optical and SAR images, such as noise characteristics, may cause distortions in the keypoint detection and matching by just using SIFT. In the research on applying the SIFT algorithm to optical and SAR image registration, some scholars have made attempts. Xiang et al. proposed OS-SIFT [14–16], which improved the method of extracting keypoints from SAR images and built a new descriptor. Sedaghat et al. established an improved descriptor-based framework called UR-SIFT [17]. Ma et al. [18] adopted a new gradient calculation architecture, which introduced an enhanced feature-matching method based on the position, scale, and orientation of each keypoint. However, the traditional feature-matching strategy mentioned above is challenging to eliminate the offset of matched point pairs caused by feature differences.

With the development of artificial intelligence, deep learning technology has gradually been applied to the registration of remote sensing images from different sources. Compared with traditional methods, deep learning-based descriptor construction methods have stronger robustness and are more conducive to keypoint matching [2]. Han et al. [19] proposed a Siamese network, which uses image blocks to construct feature vector descriptors. It has been widely applied in the registration field. Subsequently, numerous deep learning-based registration methods emerged, such as TFeat [20], L2 Net [21], HardNet [22], MatchosNet [23], etc. Xiang et al. [24] proposed a new registration method based on feature and area combinations. The above deep learning-based methods have made significant progress in the feature-based registration field, but there are also some limitations. When dealing with significant feature differences between optical and SAR images, it will lead to an offset of matched point pairs.

To mitigate the feature differences, several scholars have proposed image transformation-based methods. For instance, Maggiolo et al. [25] employed a conditional GAN-based generation strategy to convert optical images into SAR images and then conducted a template matching between the GAN-generated SAR and real SAR images. Huang et al. [26] utilized a CycleGAN network structure to transform SAR images into pseudo-optical images and registered them with real optical images. Many researchers have also explored the potential of the Transformer model [27], which was introduced by Google in 2017, for SAR image processing. Self-attention mechanisms [28–31] included by the Transformer network, which are used to replace the conventional convolution operator, have been employed in SAR image [32–34] interpretation. The attention mechanism allows the network to extend information beyond the convolution kernel's range in the surrounding space, resulting in a more robust feature extraction of spatial relations, particularly for high-resolution images [35]. It can further focus on the imaging differences between optical and SAR images, greatly increasing the success probability of transforming optical images into pseudo-SAR images [27]. In the Transformer architecture, self-attention mechanisms are used instead of CNN structures in the encoder and decoder. This paper employs a

Transformer model for pseudo-SAR image generation; then, it seeks registration between pseudo-SAR and real SAR images.

To address the registration of homologous SAR images, in the field of traditional methods, Schwind et al. [36] combined the Best-Bin-First algorithm with SIFT for SAR image registration. Delinger et al. proposed the SAR-SIFT [37] method, which replaced the DoG pyramid with the Harris pyramid and utilized the ratio gradient [38,39], achieving good results in SAR image registration. These traditional methods based on SAR image gray-scale variations struggle to detect effective keypoints in SAR images with narrow dynamic ranges, but they could lead to unsatisfied results sometimes. In the field of deep learning, Du et al. [40] introduced FM-CycleGAN to achieve feature-matching consistency. Ye et al. [5] achieved remote sensing image registration by fusing SIFT and CNN feature descriptors. However, as the convolutional layers deepen, detail features in SAR images are gradually lost. Effectively utilizing these detail features has become a research direction in SAR image registration. Yun et al. [23] presented an improved Siamese [19] network, called MatchosNet, to avoid the loss of detail features. In proposed framework, a refined scheme including MatchosNet is constructed in the registration stage between pseudo-SAR and real SAR images.

Inspired from the Transformer and the MatchosNet, this paper presents a novel method for the registration of optical and SAR images, adopting a pseudo-SAR generation strategy and an improved registration network between pseudo-SAR and real SAR images. The overall methodology is illustrated in Figure 1. In the first step, an improved Restormer [41] network is utilized to transform optical images to pseudo-SAR images. This network originated from Transformer and comprises several encoding and decoding blocks, which are each equipped with self-attention mechanisms. These mechanisms effectively capture the feature differences present in the local features of optical and SAR images. To enhance performance, we adopt the L2 loss function, which facilitates better estimating similarity between the pseudo-SAR and real SAR images, aiding the convergence of network weights toward an optimal fitting state. In the second step, the registration is conducted between the pseudo-SAR and real SAR images. Initially, the ROEWA [16] operator is applied to construct a multi-scale Harris scale space for both images. Subsequently, extremal points are selected from each scale space and incorporated into the keypoint sets obtained from the pseudo-SAR and real SAR images. The ROEWA operator effectively extracts informative features from SAR images, while the Harris scale space construction facilitates the extraction of keypoints at multiple scales, thereby enhancing the robustness of the extracted keypoints. For each keypoint in the pseudo-SAR and real SAR image keypoint sets, an image patch surrounding it is extracted and fed into the MatchosNet network. The network employs deep feature extraction, resulting in the generation of robust descriptor vectors. MatchosNet utilizes a twin-branch network with shared weights, enabling the maximum utilization of feature information from both pseudo-SAR and real SAR images, thereby producing optimal descriptor matches. Finally, the RANSAC [42] algorithm is employed to eliminate outlier matching pairs. The contributions of this paper are as follows:

- In the pseudo-SAR generation strategy, this paper use the improved Restormer network to eliminate the feature differences between optical and SAR images.
- For the registration part, a refined keypoint extraction method using the ROEWA operator is designed to construct the Harris scale space and used to extract the extreme points in each scale.

The remaining sections of this paper are organized as follows. Section 2 provides a detailed description of the proposed method. The results of registration and the ablation experiment are presented in Section 3. Section 4 shows the experimental results and a discussion of research prospects. Conclusions are given in Section 5.

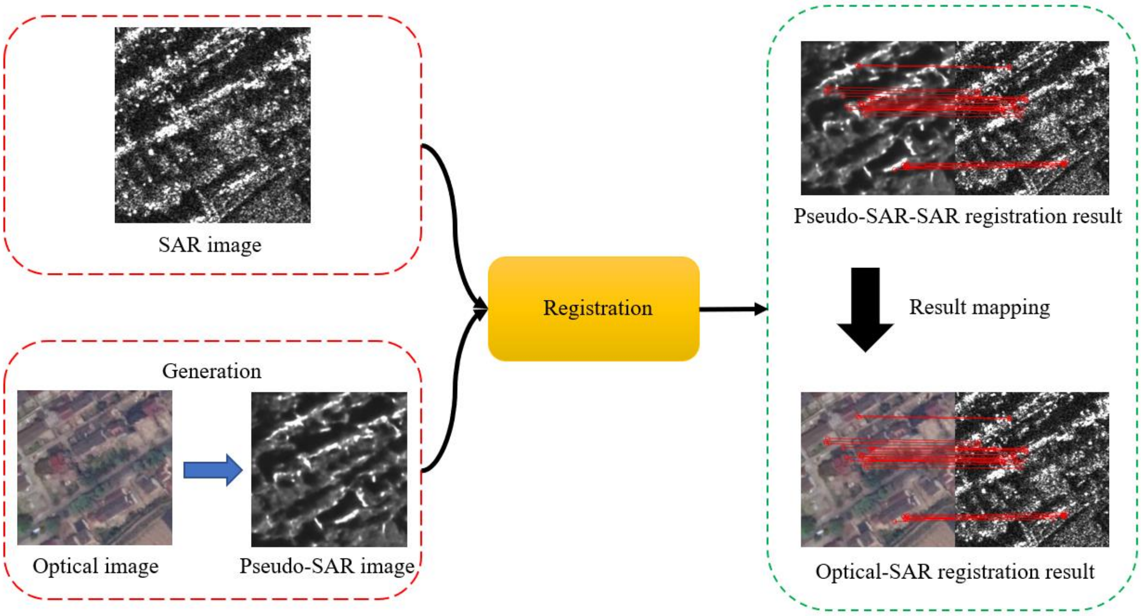


Figure 1. Overall framework of the method.

2. Materials and Methods

As Figure 1 shows, the schematic diagram of our proposed method consists of two main parts. Firstly, this paper adopts an improved Restormer to accomplish our pseudo-SAR generation strategy, transforming optical images into pseudo-SAR images. Secondly, the Harris scale space is constructed for both pseudo-SAR and real SAR images using the ROEWA operator, and we extract extremal points in each scale space. And then, image patches are selected around the keypoints, and we input them into the MatchosNet network to obtain robust descriptors. Based on the feature descriptors, keypoints are matched in pseudo-SAR and real SAR images, and we utilize the RANSAC algorithm to eliminate outliers, resulting in the final matching results.

2.1. Pseudo-SAR Image Generation Strategy

2.1.1. Network Architecture

Inspired by the Restormer network, a deep learning-based transformation method is proposed. Specifically, an encoder network is employed to extract feature information from the optical image and a decoder to decode the feature information. Within the encoder–decoder network, the transposed self-attention mechanism is used to enhance the model's robustness of the feature differences in optical and SAR images. Finally, a pseudo-SAR image consistent with the feature imaging of real SAR images is obtained. The specific network structure is illustrated in Figure 2.

As shown in Figure 2, the input to the network is an optical image to be transformed, and the output is a pseudo-SAR image. The input image is first processed by convolution to expand the number of channels to obtain a high-dimensional feature matrix. These features are then transformed into deeper feature maps through a symmetric encoder–decoder at each level. There are a total of four levels of corresponding to the encoder and decoder blocks. Each level of encoder and decoder has multiple Transformer blocks. The number of Transformer blocks gradually increases from top to bottom, mainly for deep feature extraction. The attention mechanism establishes a connection between local and global features. The encoder uses downsampling to continuously reduce the spatial size of the input image and increase feature dimension. The decoder employs upsampling to progressively enlarge the image while compressing the feature dimensions. In order to transfer the features extracted from each downsampling layer, skip connections are added after each downsampling layer, and the feature matrix obtained by upsampling are concatenated in the channel dimension and compressed by convolution. Then, the feature matrix is further refined through several Transformer blocks and a convolution layer. To better learn the differences between the optical and SAR image features, the network introduces element-wise operation between the original optical image feature and the high-level feature matrix to help restore the lost texture and semantic details in the image. Finally, the network outputs the generated pseudo-SAR image.

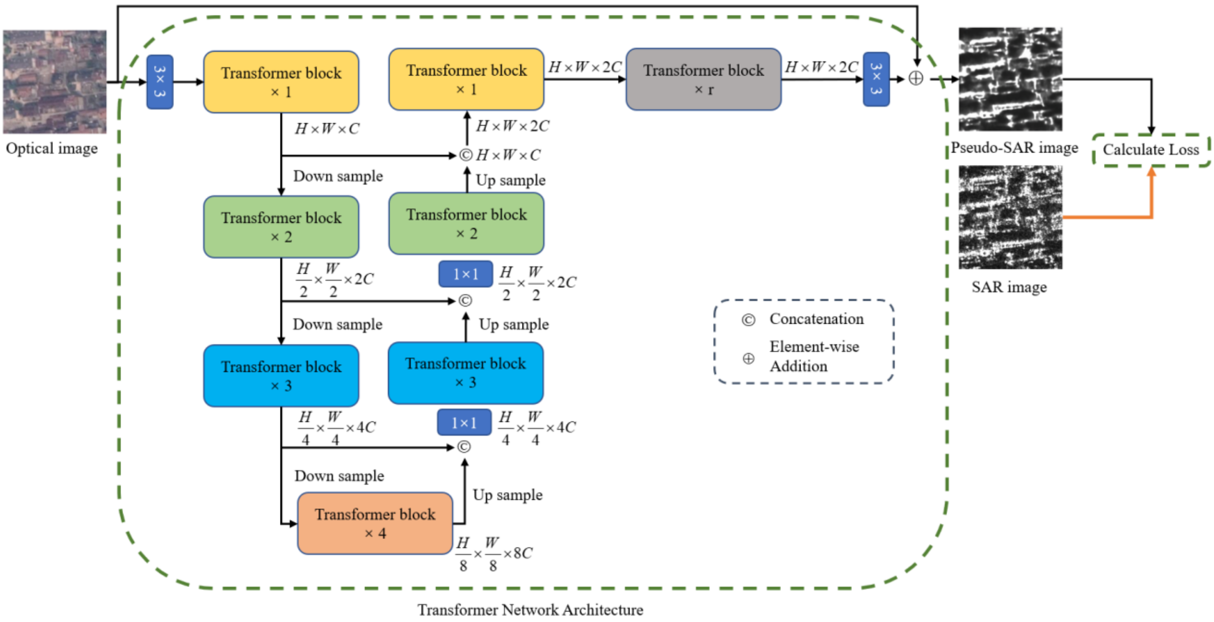


Figure 2. Transformation network architecture.

The structure of the Transformer block is shown in Figure 3, which includes two modules, MDTA [41] and GDFN. The MDTA module is Multi-Dconv Head Transposed Attention. The main network structure of this module is a self-attention mechanism. Firstly, normalize the input feature matrix, and then generate **Q**, **K**, **V** projections through 1 × 1 point-wise convolution and 3 × 3 depth-wise convolution, respectively. Here, **Q** projection is query projection, **K** projection is keyword projection, and **V** is value projection. Then, multiply the **Q** and **K** matrices to obtain the Transposed Attention Map. The Transposed Attention Map is element-wise multiplied with the **V**, and the resulting matrix is obtained through 1 × 1 point-wise convolution to generate the output matrix of MDTA. The principle formula is as follows:

$$\begin{gathered}\hat{\mathbf{X}} = W_p \text{ Attention } (\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) + \mathbf{X} \\ \text{Attention}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = \hat{\mathbf{V}} \cdot \text{Softmax}(\hat{\mathbf{K}} \cdot \hat{\mathbf{Q}})\end{gathered} \tag{1}$$

where **X** and $\hat{\mathbf{X}}$ are the input and output vector feature maps, respectively, and the $\hat{\mathbf{Q}} \in \mathbb{R}^{\hat{H}\hat{W}\times\hat{C}}$, $\hat{\mathbf{K}} \in \mathbb{R}^{\hat{C}\times\hat{H}\hat{W}}$, and $\hat{\mathbf{V}} \in \mathbb{R}^{\hat{H}\hat{W}\times\hat{C}}$ matrices are obtained by transformation from the original matrix $\mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$.

The GDFN module is Gated Dconv Feedforward Network [41]. As shown in Figure 3, the GDFN structure is divided into two parallel paths. Both paths undergo 1 × 1 convolution and 3 × 3 depth convolution. One of the paths is nonlinear activated by the GELU. Finally, the outputs of the two paths are multiplied element-wise and passed through a 1 × 1 convolution layer. The resulting matrix is then added element-wise to the input matrix to obtain the output matrix. The corresponding formula for this method is as follows:

$$\begin{gathered}\hat{\mathbf{X}} = W_p^0 \text{Gating}(\mathbf{X}) + \mathbf{X} \\ \text{Gating}(\mathbf{X}) = \phi\Big(W_d^1 W_p^1(\text{LN}(\mathbf{X}))\Big) \odot W_d^2 W_p^2(\text{LN}(\mathbf{X}))\end{gathered} \tag{2}$$

where ⊙ represents the multiplication of each element in the vector, and ϕ represents the nonlinear GELU function. LN is a layer normalization operation. $W_p^{(\cdot)}$ is 1×1 pixel by pixel convolution, and $W_d^{(\cdot)}$ is 3×3 convolution.

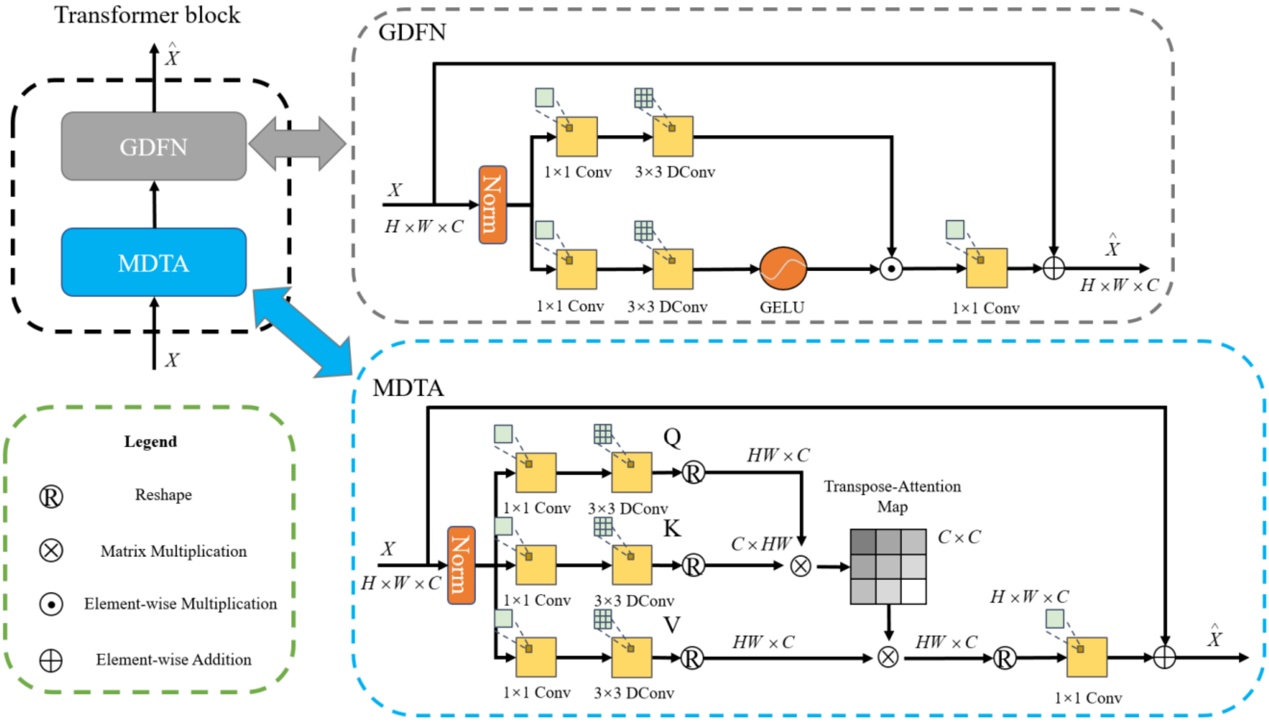


Figure 3. Transformer block detail.

2.1.2. Pseudo-SAR Generation Network Loss Function

In the application of Restormer to optical image, an L1 loss function is usually employed. Compared to the L1 loss function, the L2 loss function emphasizes the penalization of erroneous pixels, resulting in smoother fluctuations around the best fit in the model [43,44]. The later experiments will further demonstrate its effectiveness. Therefore, the L2 loss function is used here to calculate the difference between the pseudo-SAR and real SAR image. The formula for the L2 loss function is given below:

$$\mathrm{Loss}(x,y) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij} - f(x)_{ij}\right)^2 \tag{3}$$

where y_{ij} is the pixel values of real SAR images at coordinates (i,j), and $f(x)_{ij}$ is the pixel values of pseudo-SAR image at coordinates (i,j). n and m represent the size parameters of images.

2.1.3. Pseudo-SAR Generation Performance Evaluation

The performance evaluation of the pseudo-SAR generation strategy is conducted using a combination of subjective visual assessment and objective evaluation metrics. The objective quantitative evaluation indicators used for testing are the Average Gradient (AG), Structural Similarity (SSIM), Peak Signal-to-noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [45], and Mean Absolute Error (MAE) index. The pseudo-SAR image and results of these indicators will also be mentioned in the ablation experiments.

(1) **AG**

The Average Gradient index is calculated using the following formula:

$$AG = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \sqrt{\frac{\Delta I_x^2 + \Delta I_y^2}{2}} \tag{4}$$

where m and n are the size parameters of the image, and ΔI_x and ΔI_y are the differences on the horizontal and vertical coordinates, respectively. The AG reflects the difference in gray-scale near the edge of an image and is used to measure the clarity of the image, and a larger value represents a clearer image.

(2) **SSIM**

The formula for $SSIM$ is given as follows:

$$\begin{gathered} SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \\ c_1 = (k_1 L)^2 \\ c_2 = (k_2 L)^2 \end{gathered} \tag{5}$$

where μ_x is the mean of x, μ_y is the mean of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y, and c_1 and c_2 are used to maintain stability in the dynamic range of pixel values. Among c_1 and c_2, L is the dynamic value range of the pixel value, k_1 and k_2 represent the hyperparameters, which are generally 0.01 and 0.03. In this evaluation indicator, a value closer to 1 indicates a higher similarity between the pseudo-SAR and the real SAR image.

(3) **PSNR**

Regarding $PSNR$, the formula is given as follows:

$$\begin{gathered} MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \\ PSNR = 20 \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \end{gathered} \tag{6}$$

where MSE refers to the mean square error, m and n represent the dimensions of the image, and i and j represent the positions of the pixels. MAX_I represents the maximum pixel value. Generally, a higher PSNR value represents the better similarity of pseudo-SAR and real SAR images.

(4) **LPIPS**

In this paper, LPIPS [45] represents the distance between pseudo-SAR and real SAR image features. LPIPS utilizes a CNN network to extract features from images and calculates distances using these features. A smaller value of LPIPS indicates a higher similarity between the two images. The formula is shown as follows:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}_{hw}^l - \hat{y}_{0hw}^l \right) \right\|_2^2 \tag{7}$$

where x and x_0 represent the pseudo-SAR and real SAR image, respectively. $\hat{y}_{hw}^l$ and $\hat{y}_{0hw}^l$ denote the features extracted from the CNN network at the L-th layer. w_l represents the weights of the L-th layer of the CNN network. H and W represent the size parameters of the image.

(5) **MAE**

MAE represents the difference value between pseudo-SAR and real SAR images and can be expressed using the following formula:

$$MAE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)] \quad (8)$$

where i and j represent the coordinates of a pixel, and m and n are the dimension parameters of the image. A smaller *MAE* value indicates a higher similarity between the pseudo-SAR and real SAR image.

2.1.4. Parameter Analysis

During the model training process, it is necessary to determine the training iteration value, learning rate, and optimizer. The training iteration is set to an empirical value. As for the learning rate and optimizer settings, this paper follows the strategies mentioned in the literature [41]. The specific setting approach is provided in Section 3.1.2.

2.2. Image Registration

A point-matching-based registration framework is adopted in this paper. The framework includes keypoint extraction, feature descriptor construction, feature descriptor matching, and RANSAC to remove outliers. The flow chart of this method is shown in Figure 4.

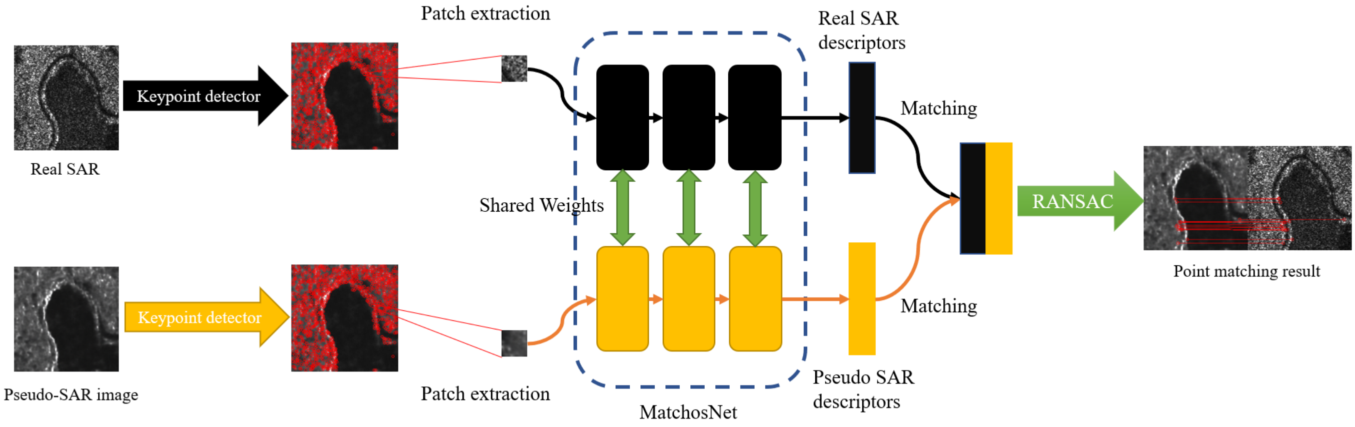


Figure 4. Procedure of registration.

The overall process of the registration scheme is described below. First of all, extract keypoints from real SAR and pseudo-SAR images. Secondly, image patches are extracted centered around the keypoints and fed into the MatchosNet to extract feature descriptors. Then, the keypoints are matched according to the descriptor. Finally, remove the outliers by RANSAC [42] and obtain the final registration result.

2.2.1. Keypoint Detection

The registration framework proposed by Yun et al. [23], based on MatchosNet, utilizes the Difference of Gaussians (DoG) operator to extract keypoints from optical and SAR images. However, directly applying the DoG operator in SAR images with significant speckle noise will lead to the inability to extract repeatable keypoints, affecting the orientation assignment and descriptor construction, and resulting in registration errors [14]. The ROEWA operator proposed by Fjortoft et al. [16] is commonly used for the edge detection of SAR images with good results. It uses gradient by ratio (GR) instead of a differential gradient, which consists of two orthogonal one-dimensional filters to form a two-dimensional separable filter. In this paper, the ROEWA [16] operator is used to extract images at different scales to help establish the Harris scale space. The comparison between our proposed keypoint extraction method and the DoG operator is given in the following ablation experiment chapter.

The calculation process of constructing the Harris scale space using the ROEWA operator is described by Equations (9)–(13) as shown. A series of scale parameters are constructed, which are denoted as α.

$$\alpha = \alpha_0 \times c^i \tag{9}$$

where i represents the spatial layer of the scale, c is a constant, α denotes the scale space parameter, and α_0 represents the initial value of the scale space parameter.

Hereafter, the ROEWA operators oriented in the horizontal and vertical directions are defined as follows:

$$R_{h,\alpha} = \frac{\sum\limits_{m=-M/2}^{M/2} \sum\limits_{n=1}^{N/2} I(x+m, y+n) e^{-\frac{|m|+|n|}{\alpha}}}{\sum\limits_{m=-M/2}^{M/2} \sum\limits_{n=-N/2}^{-1} I(x+m, y+n) e^{-\frac{|m|+|n|}{\alpha}}} \tag{10}$$

$$R_{v,\alpha} = \frac{\sum\limits_{m=1}^{M/2} \sum\limits_{n=-N/2}^{N/2} I(x+m, y+n) e^{-\frac{|m|+|n|}{\alpha}}}{\sum\limits_{m=-M/2}^{-1} \sum\limits_{j=-N/2}^{N/2} I(x+m, y+n) e^{-\frac{|m|+|n|}{\alpha}}} \tag{11}$$

where M and N are the size of the sliding processing window, $I(x, y)$ represents the pixel intensity of the image, x and y represent the coordinates of the center point. $R_{h,\alpha}$ and $R_{v,\alpha}$ denote the horizontal and vertical ROEWA operator, separately.

For the next step, the horizontal and vertical gradients are calculated using $R_{h,\alpha}$ and $R_{v,\alpha}$, respectively.

$$\begin{aligned} G_{h,\alpha} &= \log(R_{h,\alpha}) \\ G_{v,\alpha} &= \log(R_{v,\alpha}) \end{aligned} \tag{12}$$

where $G_{h,\alpha}$ and $G_{v,\alpha}$ represent the horizontal and vertical gradients, respectively.

Finally, the Harris scale space is constructed by the following formula:

$$\begin{aligned} C_{\mathrm{SH}}(h, v, \alpha) &= g_{\sqrt{2}\alpha} * \begin{bmatrix} (G_{h,\alpha})^2 & (G_{h,\alpha}) \cdot (G_{v,\alpha}) \\ (G_{h,\alpha}) \cdot (G_{v,\alpha}) & (G_{v,\alpha})^2 \end{bmatrix} \\ R_{\mathrm{SH}}(h, v, \alpha) &= \det(C_{\mathrm{SH}}(h, v, \alpha)) - d \cdot \mathrm{tr}(C_{\mathrm{SH}}(h, v, \alpha))^2 \end{aligned} \tag{13}$$

where $g_{\sqrt{2}\alpha}$ represents the Gaussian convolution kernel with scale α, $*$ denotes the convolution operation, d is a hyperparameter, and R_{SH} represents the Harris scale space.

Then, in each scale level of the Harris scale space, a local extrema value is found as a candidate keypoint and added to the keypoint set. The flow chart for extracting keypoints is illustrated as shown in Figure 5.

2.2.2. Feature Descriptor Construction

For the obtained keypoints, the image patches centered on the keypoint coordinates are selected. Then, we input the extracted image patches to the MatchosNet. The network adopts a twin structure and shares weights. The backbone of MatchosNet is CSP-DenseNet [46]. The network reduces the computation and enhances the learning ability. The network structure is shown in Figure 6.

There are three DenseBlocks with the same structure in the network. The structure of each DenseBlock is shown in Figure 7.

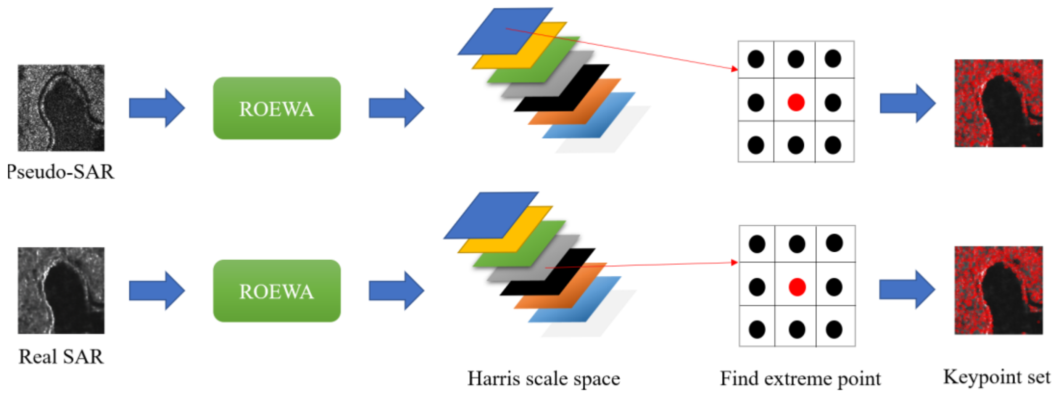


Figure 5. Keypoint detection flowchart.

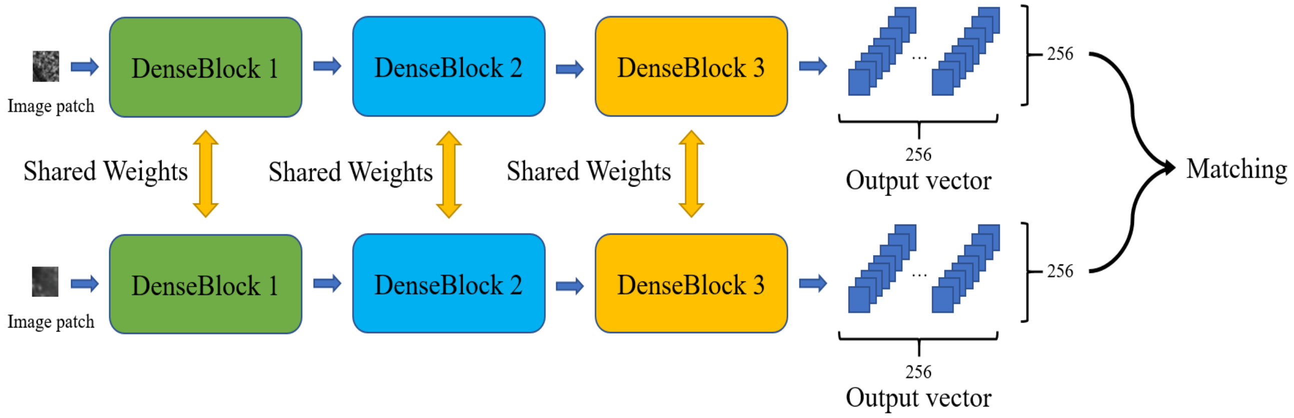


Figure 6. Descriptor construction network architecture.

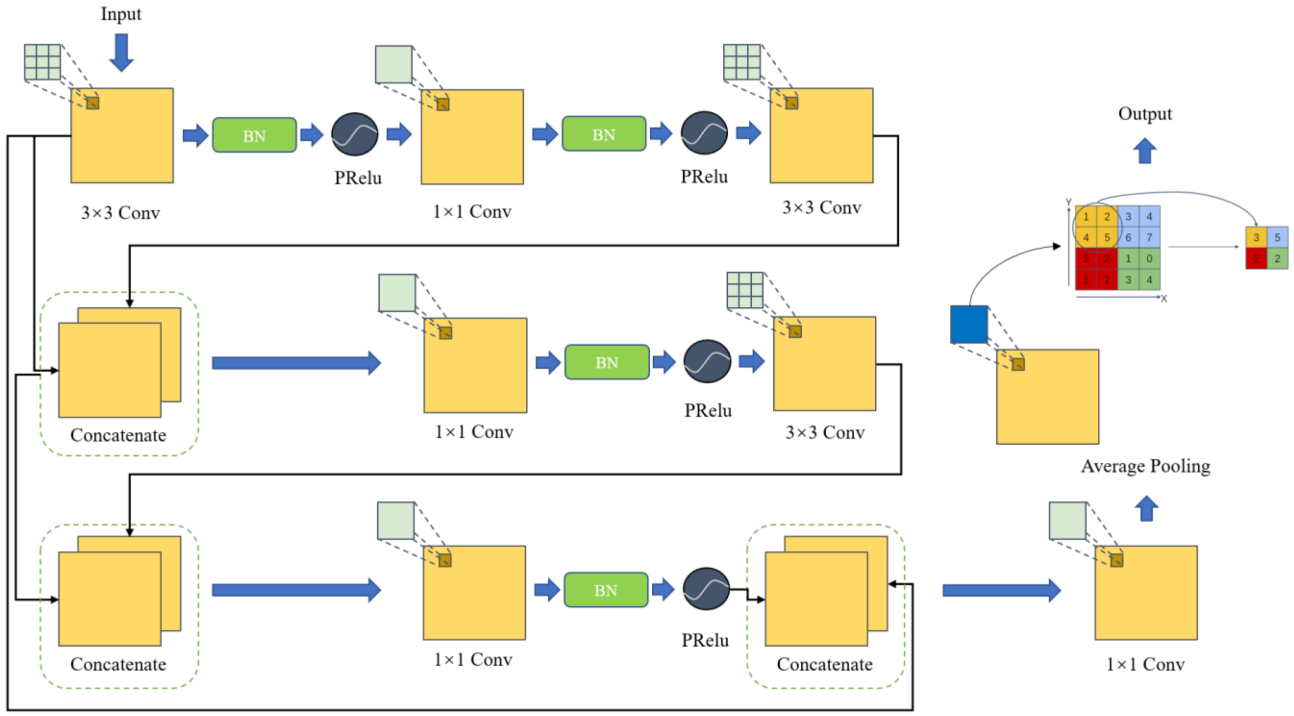


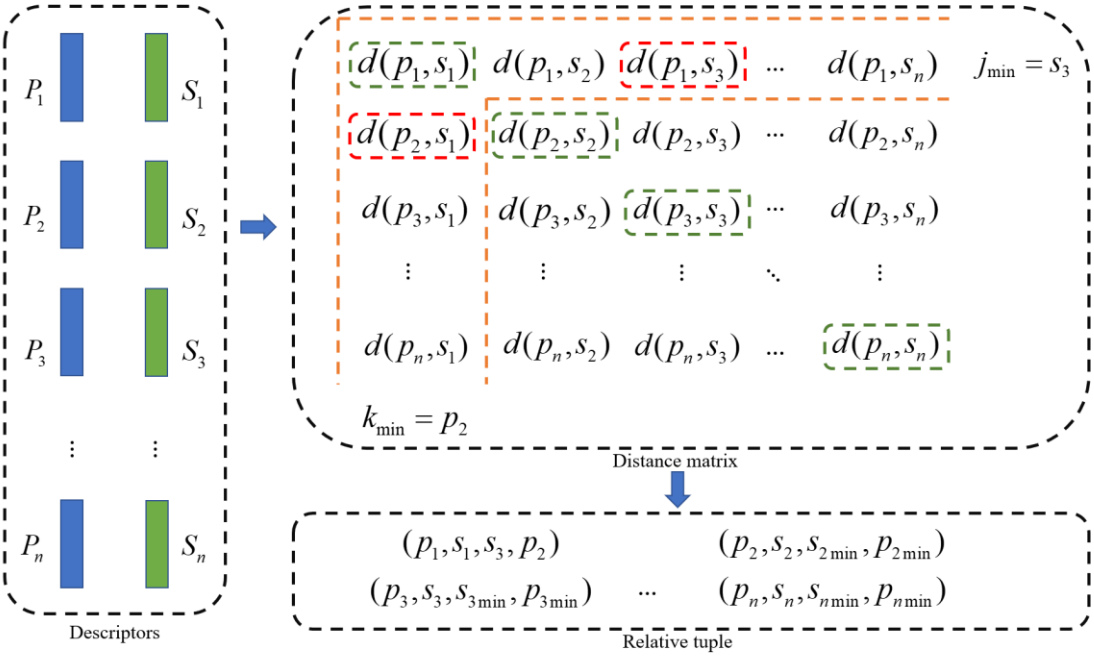
Figure 7. DenseBlock architecture.

Each DenseBlock consists of 11 layers, including 3 layers of 3 × 3 depth-wise convolutions, 4 layers of 1 × 1 pixel-wise convolutions, 3 layers of channel-wise feature concatenation, and 1 layer of average pooling. The DenseBlock removes redundant layer connections and retains some important layer connections, resulting in the improvement of the operation efficiency.

2.2.3. Descriptor Matching Loss Function

Figure 8 illustrated the computation process of the loss function. The figure consists of three components: namely, descriptors, distance matrix, and relative tuple. In the descriptors section, there are total of n matched descriptor pairs. The distance matrix section represents the distance matrix formed by calculating the L2 distances between all descriptors [21]. The relative tuple section consists of the four-tuple collection computed from the distance matrix. This collection is used for the final calculation of the loss function, which is shown below:

$$L_{HardL2}(p_i, s_j, s_{jmin}, p_{imin}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\max(0, (1 + d(p_i, s_j) - \min(d(p_i, s_{jmin}), d(p_{imin}, s_j))) \tag{14}$$



P_1
S_1
P_2
S_2
P_3
S_3
P_n
S_n
Descriptors
$d(p_1,s_1)$ $d(p_1,s_2)$ $d(p_1,s_3)$ ··· $d(p_1,s_n)$ $j_{\min} = s_3$
$d(p_2,s_1)$ $d(p_2,s_2)$ $d(p_2,s_3)$ ··· $d(p_2,s_n)$
$d(p_3,s_1)$ $d(p_3,s_2)$ $d(p_3,s_3)$ ··· $d(p_3,s_n)$
$d(p_n,s_1)$ $d(p_n,s_2)$ $d(p_n,s_3)$ ··· $d(p_n,s_n)$
$k_{\min} = p_2$
Distance matrix
(p_1,s_1,s_3,p_2) $(p_2,s_2,s_{2\min},p_{2\min})$
$(p_3,s_3,s_{3\min},p_{3\min})$ ··· $(p_n,s_n,s_{n\min},p_{n\min})$
Relative tuple

Figure 8. Computation process of the hard L2 loss function.

$d(p_i, s_j)$ represents the distance of matching descriptor pairs, where $i = 1\ldots n, j = 1\ldots n$. $d(p_{imin}, s_j)$ and $d(p_i, s_{jmin})$ are the distances of non-matching descriptor pairs closest to $d(p_i, s_j)$, where $imin = 1\ldots n$ and $jmin = 1\ldots n$. The descriptors involved in this computation are referred to as a quadruple $(p_i, s_j, s_{jmin}, p_{imin})$.

After extracting feature descriptors using the network, the RANSAC algorithm is used to remove outliers and obtain the point matching result. Finally, the registration results of the pseudo-SAR and SAR images are mapped back to the optical and SAR images.

2.2.4. Parameter Analysis

For the registration part, it is necessary to determine the Harris scale space constant c, d, and α_0. In the registration network, the image patch size, number of training iterations, learning rate, optimizer, and RANSAC threshold need to be determined. The specific parameter values for the c, d, α_0, image patch size, learning rate, and optimizer are set according to the literature [23,37]. As for the training iteration value and the RANSAC threshold, these rely on empirical values. These values are provided in Section 3.1.2.

3. Results

3.1. Experiment Preparation

3.1.1. Dataset Preparation

The QXS-SAROPT [47] and OSDataset [48] datasets are adopted in this experiment. For the QXS-SAROPT dataset, the SAR image part of the dataset is from the Gaofen-3 satellite and the resolution is 1 m × 1 m, and the optical image part is from Google Earth images. These images cover three port cities: Santiago, Shanghai, and Qingdao, and they contain 20,000 pairs of optical and SAR image pairs. The dataset is split into three parts for training, validation, and testing, with a ratio of 8:1:1. In the training process of the pseudo-SAR generation network, the training and validation sets of this dataset are adopted. The test set is used to test the pseudo-SAR generation network and the registration network. The OSDatase contains 10,692 pairs of optical and SAR images, each with a size of 256 × 256 pixels. These SAR images have the same sensor source and resolution as QXS-SAROPT. The dataset collects scenes from several cities around the world, including Beijing, Shanghai, Suzhou, Wuhan, Sanhe, Yuncheng, Dengfeng, Zhongshan, and Zhuhai in China, Rennes in France, Tucson, Omaha, Guam, and Jacksonville in the United States, and Dehradun and Agra in India. The SAR images from this dataset are used to train the registration network.

3.1.2. Parameter Setting

For the pseudo-SAR generation strategy, the total number of iterations for training the original and improved Restormer network is set to 600,000 empirically. Regarding the literature [41], the initial learning rate is set to 3×10^{-4} and gradually reduced to 1×10^{-6} using the cosine annealing algorithm. The optimizer is set to Adam.

For constructing the Harris scale space, according to the literature [37], the parameters c, d, and α_0 are set as $2^{1/3}$, 0.04, and 2 respectively. On the basis of the literature [23], the training optimizer is Adam, the learning rate is set to 1×10^{-4}, and the size of the image patches is specified to be 64 × 64. Based on the empirical values, the training epoch is set to 100, and the RANSAC threshold is set to 1.

In the ablation experiment, to ensure fairness, the CycleGAN model is trained for 600,000 iterations. The optimizer used is Adam with a learning rate of 3×10^{-4}. Additionally, the learning rate gradually reduce to 1×10^{-6} using the cosine annealing algorithm.

3.1.3. Registration Comparison Method

The comparative experiments used in this paper are as follows.

(1) PSO-SIFT [49]: According to the existing SIFT method, PSO-SIFT adopts a new gradient definition to eliminate the nonlinear radiation differences between optical and SAR images.

(2) MatchosNet [23]: MatchosNet proposes a deep convolution Siamese network based on CSPDenseNet to obtain powerful matching descriptors to improve the matching effect.

(3) CycleGAN + MatchosNet [26]: This method uses CycleGAN [50] to generate pseudo-optical images from SAR images, and it uses SIFT to match the pseudo-optical and optical images to obtain the final registration results. In the ablation experiment, the CycleGAN network and the improved Restormer network are compared in the pseudo-SAR generation strategy. In the registration experiment, we make improvements to this method by converting the optical image into a pseudo-SAR image and replacing the SIFT with MatchosNet to better evaluate the registration method proposed in this paper.

3.1.4. Experimental Platform

The platform and environment used in the experiment are shown in the Table 1.

Table 1. Experiment environment.

Environment	Version
Platform	Windows 11, Linux
Torch	V 1.9.0
Matlab	2021a
CPU	Inter Core i7-10700
Memory	16 G
Video memory	6 G

3.2. Experiment Result

3.2.1. Comparison of Registration Results

(1) **Keypoint matching analysis**

To evaluate the proposed registration method, this paper compares its performance with the methods discussed in Section 3.1.3. The results of point matching are analyzed in this section. Figure 9 show the visual results of point matching.

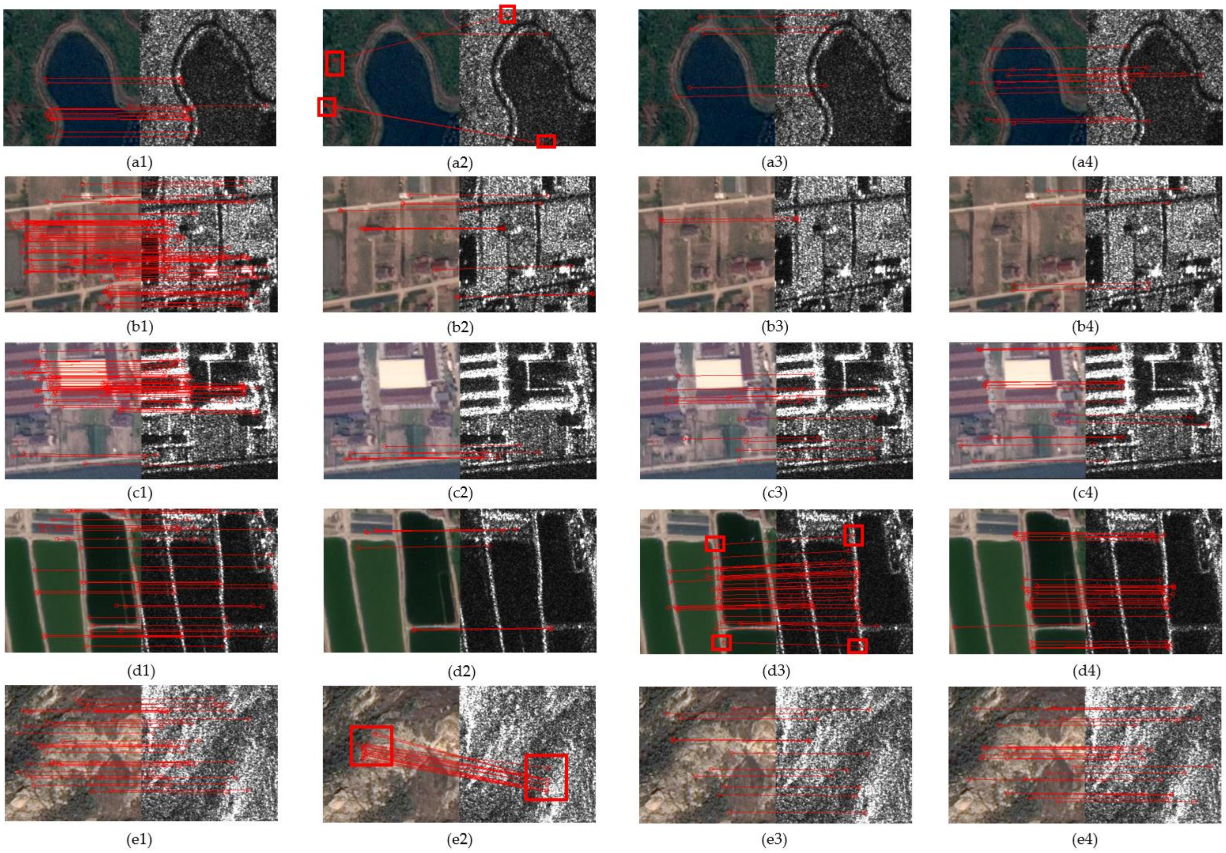


Figure 9. Experimental results of registration for different scenes. (**a1–a4**,**b1–b4**,**c1–c4**,**d1–d4**,**e1–e4**) represent the forest and lake, rural and road, urban, farmland, and mountain scenes, respectively. (**a1–e1**,**a2–e2**,**a3–e3**), and (**a4–e4**) are the results of the proposed method, PSO-SIFT, CycleGAN + MatchosNet, and MatchosNet, respectively.

Figure 9(a1–a4) show the visual registration results of four methods in forest and lake scenes. Our method achieves a higher number of matched points compared to PSO-SIFT and CycleGAN + MatchosNet, and there are no noticeable positional errors in the matched points. As shown in Figure 9(a2), the red box highlights the mismatches produced by PSO-SIFT in those areas.

The visual registration results of the rural and road scenes are shown in Figure 9(b1–b4). From the overall comparison of the four methods, it can be observed that our method has a significant advantage in terms of the number of matched points. PSO-SIFT, CycleGAN + MatchosNet, and MatchosNet exhibit fewer matched points.

Figure 9(c1–c4) show the registration results of the four methods in an urban scene. It can be noticed that our proposed method still has a significant number of matched points in areas with strong textures. Both CycleGAN + MatchosNet and MatchosNet show more matched points than PSO-SIFT in the urban scene.

Referring to Figure 9(d1–d4), the registration results of the four methods in a farmland scene are presented. Our proposed method obtains a greater number of matched point pairs in the farmland scene. PSO-SIFT shows fewer matched point pairs compared to the proposed method, CycleGAN + MatchosNet, and MatchosNet methods. The CycleGAN + MatchosNet method exhibits slight errors in matched points, and Figure 9(c3) highlights the incorrectly matched point pair within the red box.

Figure 9(e1–e4) illustrate the registration results of the four methods in the mountain scene. In the PSO-SIFT method, all keypoints are completely mismatched, as indicated by the red box. Among the remaining three methods, the proposed method outperforms CycleGAN + MatchosNet and MatchosNet in terms of the number of matched points, and the keypoints are evenly distributed.

From the visual effect of the above five scenes, deep learning-based registration methods demonstrate superior point-matching performance compared to traditional methods, especially in areas with strong textures. This is because deep learning networks can effectively learn the similarity between features of heterogeneous images. The combination of the Transformer-based pseudo-SAR generation strategy and deep learning registration mitigates the majority of feature differences in the pseudo-SAR generation stage; this strategy significantly enhance the robustness of the registration network.

Table 2 presents the quantitative evaluation of the registration results. Two quantitative metrics are used here—namely, the number of correctly matched points (NCM) and Root Mean Squared Error (RMSE)—to assess the effectiveness of our registration. Among them, a larger NCM and smaller RMSE value indicates a better matching effect. The data in Table 2 indicate that the proposed method outperforms the other three methods in terms of the NCM metric across all scenes. Regarding the RMSE metric, our method is slightly higher than the MatchosNet method only in the rural and highway scenes, but it performs lower than the other methods in the remaining four scenes. However, in the rural and road scenes, the MatchosNet registration method only achieves 4 matched point pairs, while our method achieves 84 matched point pairs.

Table 2. Comparison of evaluation indicators for registration result.

Method	Scene	NCM	RMSE (pix)
PSO-SIFT	Forest and lake	4	1.15
	Rural and road	12	1.57
	Urban	6	0.98
	Farmland	7	0.90
	Mountain	15	0.99
CycleGAN + MatchosNet	Forest and lake	5	0.89
	Rural and road	3	1.30
	Urban	12	0.96
	Farmland	33	0.88
	Mountain	11	0.90

Table 2. *Cont.*

Method	Scene	NCM	RMSE (pix)
MatchosNet	Forest and lake	13	0.87
	Rural and road	4	**0.96**
	Urban	15	0.79
	Farmland	29	0.89
	Mountain	26	0.94
Proposed Method	Forest and lake	**15**	**0.83**
	Rural and road	**84**	0.99
	Urban	**57**	**0.76**
	Farmland	**34**	**0.82**
	Mountain	**36**	**0.84**

(2) **Checkerboard image experiment analysis**

The visual appearance of the checkerboard pattern is also an essential evaluation criterion for registration results. In order to further prove the effectiveness of the proposed method, the checkerboard image experiments are added based on the point matching results, and the experimental results are shown in Figure 10. In Figure 10(a1–a4), the PSO-SIFT method completely fails to register, while MatchosNet exhibits a matching error in the red-boxed region. As shown in Figure 10(b1–b4), the PSO-SIFT method generates incorrect matches in the red-boxed region. In the urban scene, only our method successfully matched the images; the other three methods exhibit mismatches in the red-boxed region, as depicted in Figure 10(c1–c4). From Figure 10(d1–d4), it can be observed that in the farmland scene, the other three methods also exhibit matching errors in the red-boxed region, while the proposed method still achieves successful registration. According to Figure 10(e1–e4), in the mountain scene, the PSO-SIFT method fails to achieve a complete registration. The proposed method accomplishes the registration successfully. These results indicate that deep learning-based methods have advantages in registration, and the proposed Transformer-based pseudo-SAR generation strategy further improves the registration performance between optical and SAR images.

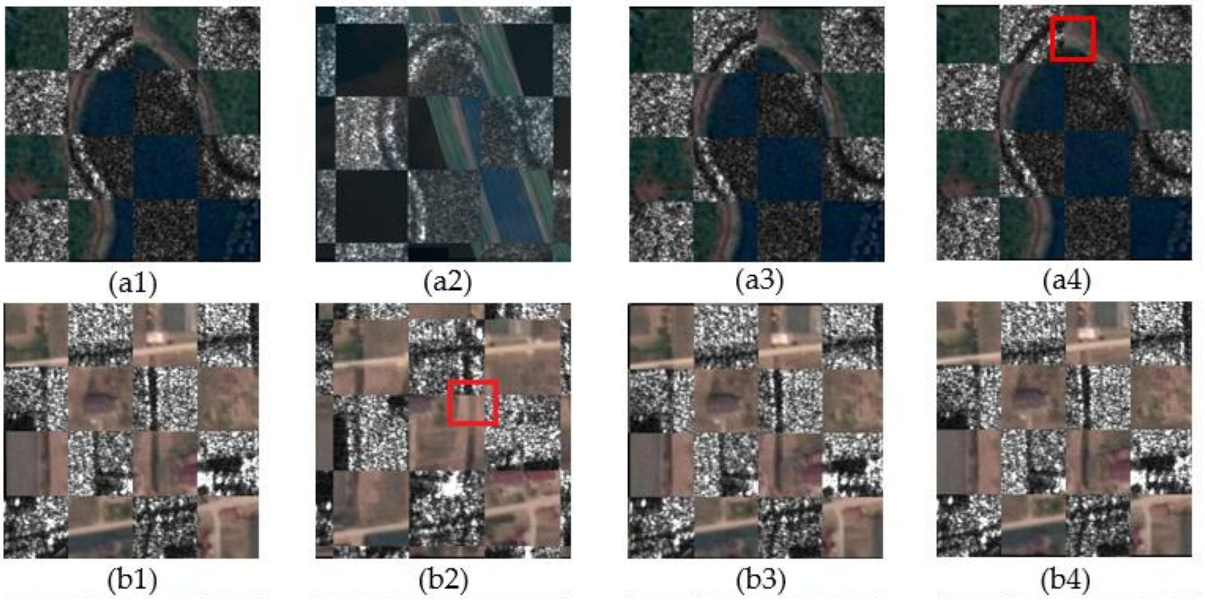


Figure 10. *Cont.*

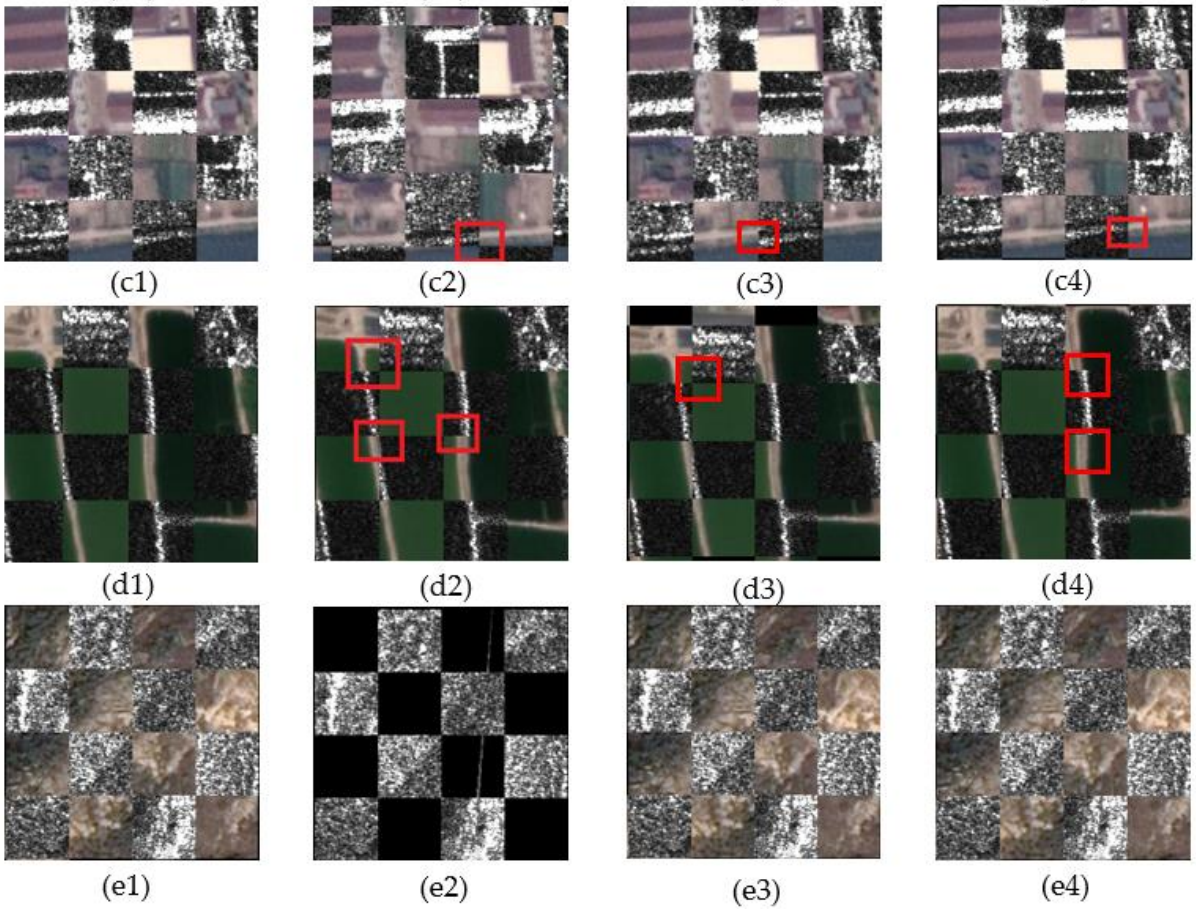


Figure 10. Registration checkerboard results for five scenarios. (**a1**–**a4**) represent forest and lake scenes, (**b1**–**b4**) represent rural and road scenes, (**c1**–**c4**) represent urban scenes, (**d1**–**d4**) represent farmland scenes, and (**e1**–**e4**) represent mountain scenes. (**a1**–**e1**) are the experimental results of the proposed method, (**a2**–**e2**) are the experimental results of the PSO-SIFT method, (**a3**–**e3**) are the experimental results of the CycleGAN + MatchosNet method, and (**a4**–**e4**) are the experimental results of the MatchosNet method.

3.2.2. Ablation Experiment

This section presents a series of ablation experiments, including pseudo-SAR generation strategy validity analysis, the validity analysis of the pseudo-SAR generation strategy for registration, and keypoints extraction strategy validity analysis.

(1) **Pseudo-SAR generation strategy validity analysis**

In this experiment, AG is used to evaluate the pseudo-SAR image. SSIM, PSNR, LPIPS, and MAE are calculated between the pseudo-SAR and real SAR images. The objective evaluation indicator values for the generation results are shown in the Table 3. Bold font indicates optimal values.

The experiment results show that the improved Restormer outperforms the original Transformer and CycleGAN in terms of AG, SSIM, and PSNR metrics. The improved Restormer achieves the best performance in most scenes based on the LPIPS and MAE metrics. Therefore, the experimental result indicates that the improved Restormer outperforms the CycleGAN and the original Restormer, and L2 loss function is superior to L1 loss function in the pseudo-SAR generation strategy.

The subjective evaluation indicators for this experiment were based on visual evaluation. Figure 11 presents a comparison of five different scenes: forest and lake, rural and road, urban, farmland, and mountain. Each scene comparison comprises five images, arranged from left to right: a real optical image, a pseudo-SAR image generated by CycleGAN, a pseudo-SAR image generated by original Restormer, a pseudo-SAR image generated by improved Restormer, and a real SAR image.

Table 3. Quantitative evaluation index of pseudo-SAR generation strategy.

Method	**Evaluation Metrics**	**Scenes**				
		Forest and Lake	**Rural and Road**	**Urban**	**Farmland**	**Mountain**
CycleGAN	AG↑	11.39	16.19	17.56	12.00	17.05
	SSIM↑	0.64	0.80	0.78	0.79	0.72
	PSNR↑	12.70	10.02	8.64	12.14	10.02
	LPIPS↓	0.62	0.61	0.62	0.57	0.56
	MAE↓	173.28	127.16	121.04	**89.15**	136.67
Original Restormer	AG↑	15.86	25.04	23.05	19.96	24.86
	SSIM↑	0.89	0.91	0.90	0.81	0.97
	PSNR↑	15.30	12.33	11.78	15.76	14.30
	LPIPS↓	0.56	0.60	0.58	0.54	**0.50**
	MAE↓	143.58	**113.39**	115.57	95.24	123.57
Improved Restormer	AG↑	**16.91**	**27.03**	**26.73**	**21.19**	**25.34**
	SSIM↑	**0.93**	**0.92**	**0.92**	**0.88**	**0.99**
	PSNR↑	**16.17**	**12.80**	**12.69**	**16.00**	**14.54**
	LPIPS↓	**0.50**	**0.53**	**0.53**	**0.51**	0.54
	MAE↓	**141.02**	123.16	**114.45**	92.89	**123.12**

Based on the comparison of pseudo-SAR generation strategy, in Figure 11(a2–e2), as indicated by the red-box marked, the CycleGAN method only focuses on rendering the style of the SAR image onto the optical image in the rural and rode, urban, farmland, and mountain scenes without fully eliminating the feature difference of the target. The original Restormer and improved Restormer exhibit better visual effects compared to CycleGAN. As shown in Figure 11(a3–e3,a4–e4), Restormer produces pseudo-SAR images that resemble real SAR images more closely. Specifically, improved Restormer provides clearer textures in the generated images compared to the original Restormer, as depicted in Figure 11(b4,d4,e4). The original Restormer generates images with relatively blurred texture details in some areas of the pseudo-SAR images, as shown in the red box marked in Figure 11(b3,d3,e3). In conclusion, in the pseudo-SAR generation strategy, the original Restormer outperforms similar methods, and the improved Restormer further improves the generating effect. The above conclusions also prove that the L2 loss function has advantages in the field of pseudo-SAR image generation.

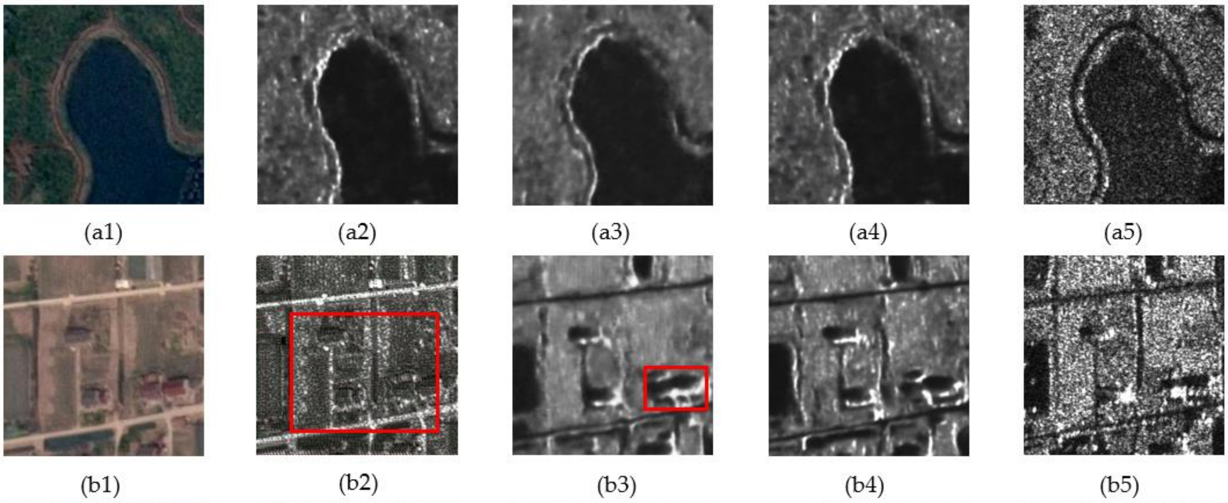


Figure 11. *Cont.*

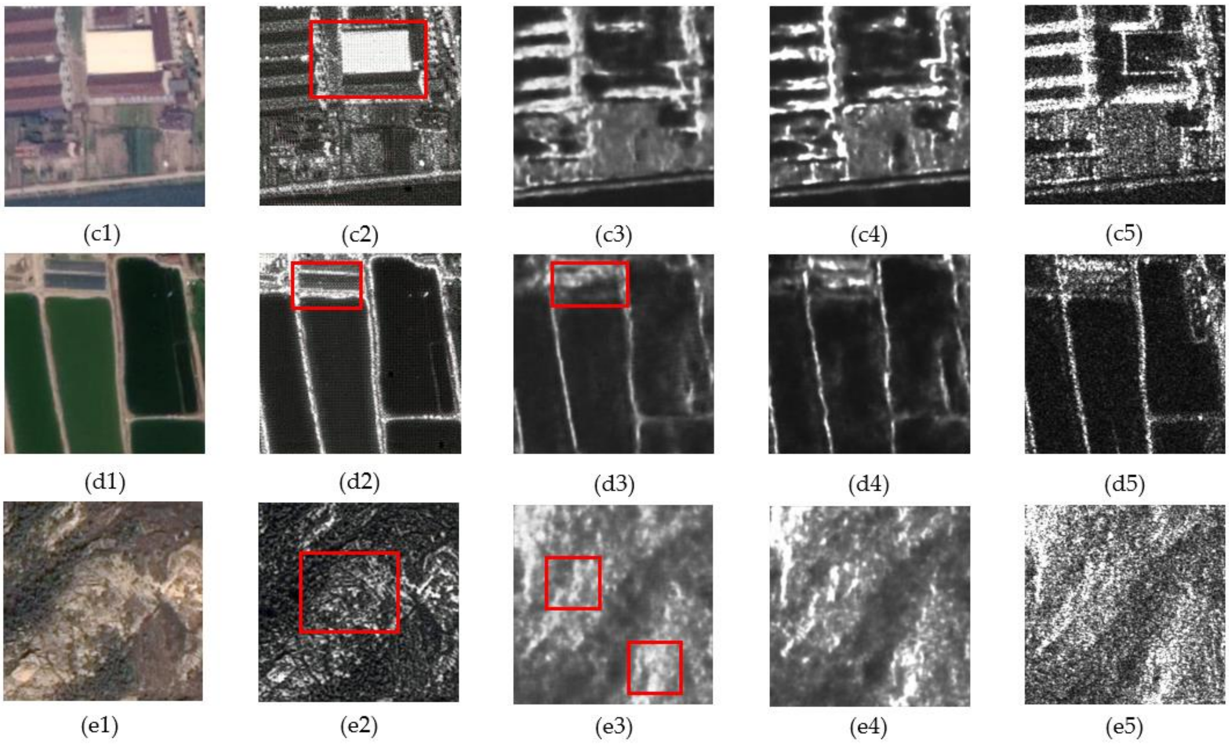


Figure 11. Comparison of pseudo-SAR generation strategy. (**a1**–**a5**) represent the forest and lake scenes, (**b1**–**b5**) represent the rural and road scenes, (**c1**–**c5**) represent the urban scenes, (**d1**–**d5**) represent the farmland scenes, and (**e1**–**e5**) represent the mountain scenes. (**a1**–**e1**) are the real optical images, (**a2**–**e2**) are pseudo-SAR images generated by CycleGAN, (**a3**–**e3**) are pseudo-SAR images generated by original Transformer, (**a4**–**e4**) are pseudo-SAR images generated by improved Restormer, and (**a5**–**e5**) are the real SAR images.

(2) The validity analysis of pseudo-SAR generation strategy for registration

Figure 12 illustrates a comparison of the registration results between the proposed method, the original Transformer, CycleGAN + MatchosNet, and MatchosNet. Among them, MatchosNet directly registers optical and SAR images. The proposed method, the original Transformer and CycleGAN + MatchosNet are two-stage registration methods that involve pseudo-SAR generation and registration. From the results in Figure 12, it can be observed that the proposed method exhibits a more even distribution and a higher number of matching points. The registration results based on the Transformer pseudo-SAR generation strategy outperform the direct registration of optical and SAR images. In the registration methods based on CycleGAN, one scene shows a matching error, as indicated by the red box in Figure 12(d3).

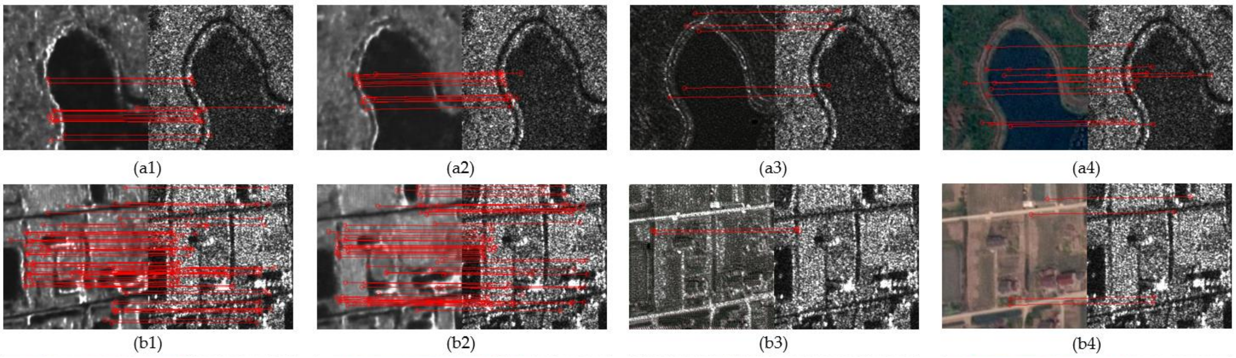


Figure 12. *Cont.*

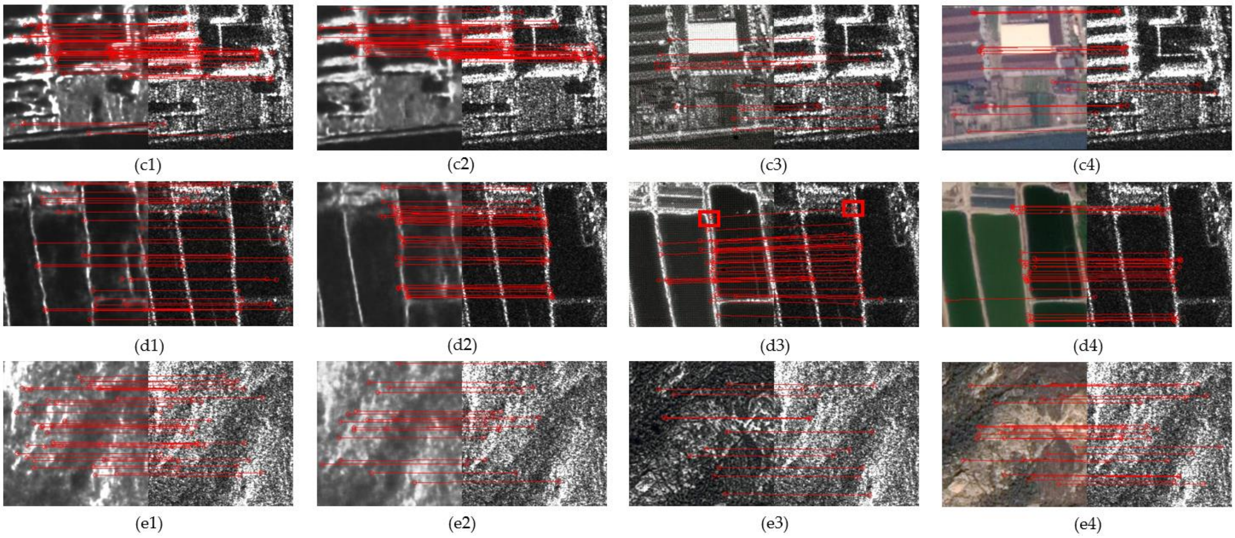


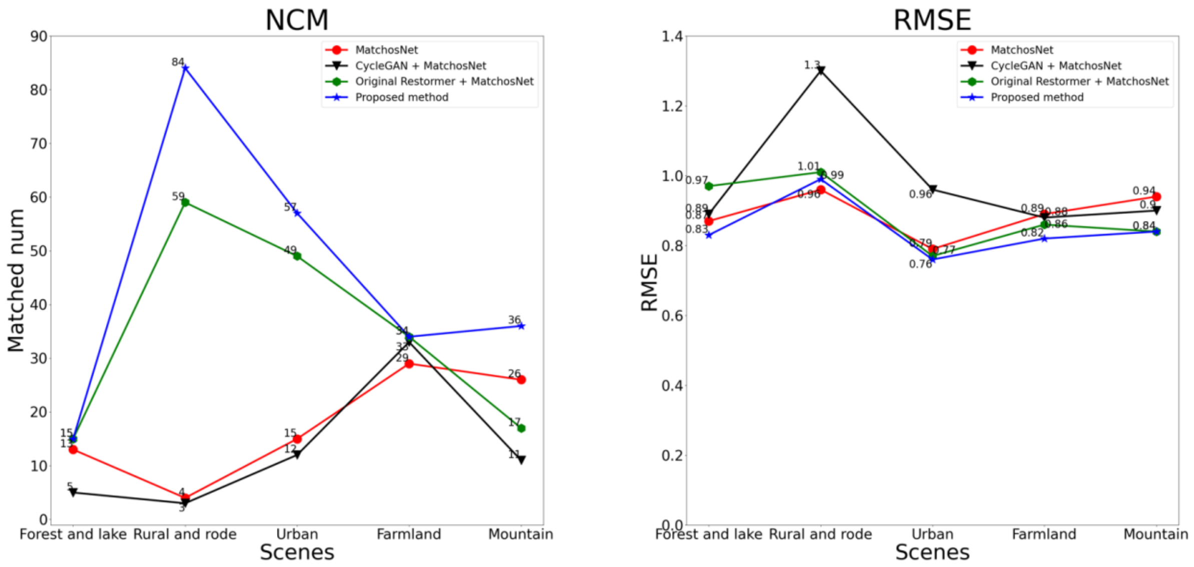
Figure 12. The results of pseudo-SAR generation strategies for registration and direct registration. (**a1**–**a4**) represent the forest and lake scenes, (**b1**–**b4**) represent the rural and road scenes, (**c1**–**c4**) represent the urban scenes, (**d1**–**d4**) represent the farmland scenes, and (**e1**–**e4**) represent the mountain scenes. (**a1**–**e1**) are the results of the proposed method, (**a2**–**e2**) are the results of the original Transformer + MatchosNet method, (**a3**–**e3**) are the results of CycleGAN + MatchosNet, and (**a4**–**e4**) are the results of direct MatchosNet. (**a1**–**e3**) depict the results of the two-stage registration method, showcasing the registration of the pseudo-SAR and real SAR images. (**a4**–**e4**) illustrate the results of the direct registration method, demonstrating the registration results between the optical and SAR images.

Figure 13 displays a quantitative evaluation line chart of the registration results shown in Figure 12. Figure 13a represents the NCM for the four methods; it can be observed that the proposed method achieves the highest NCM value. The original two-stage registration method using Transformer performs well and obtains the second-highest NCM values in most scenes. Figure 13b shows the RMSE for the four methods; it is evident that the proposed method achieves the lowest RMSE in most scenes. These experimental results provide substantial evidence of the proposed pseudo-SAR generation strategy in terms of image registration effectiveness.

(3) **Keypoint extraction strategy validity analysis**

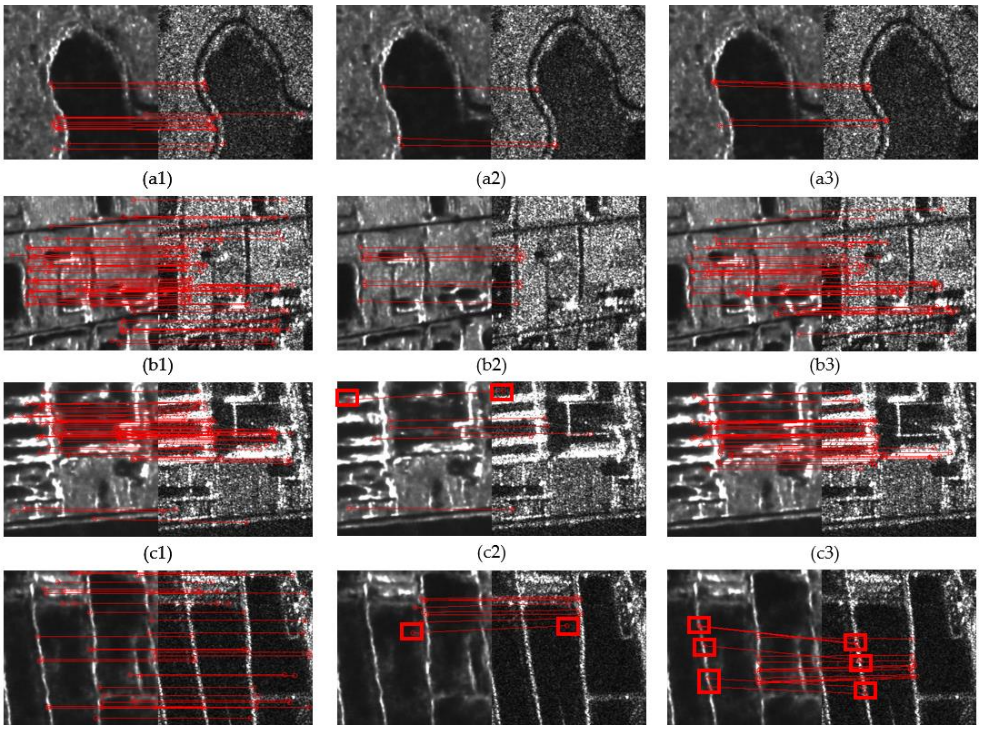
Figure 14 presents the comparison of keypoint extraction strategies for pseudo-SAR and real SAR images. It can be observed that the proposed strategy yields a higher number of keypoint matches without any matching errors. However, the DoG and FAST extraction strategies exhibit fewer matching points. Especially in some scenarios there are some erroneous keypoint matches, as indicated by the red boxes in Figure 14(c2,d2,d3).

In Figure 15, a comparison of three keypoint extraction methods is presented. In Figure 15a, the number of keypoints extracted by the three methods is compared. Figure 15b illustrates the NCM for the three methods. It can be observed that proposed keypoint extraction strategy outperforms the other methods in terms of both the number of keypoints selected and the final NCM. FAST operator extracts fewer keypoints than DoG in most cases but more keypoints than DoG in urban areas, which fully shows that there are more corners in urban areas and FAST is easier to extract keypoints than DoG. However, in weak texture areas, the corner feature is not significant, so FAST extracts fewer keypoints. The proposed keypoint extraction strategy gives consideration to both shallow and deep features, it can extract large keypoints in both weak and strong texture regions, and the final matching keypoints are more than FAST and DoG. This provides substantial evidence for the robustness of our keypoint extraction strategy.



(a) (b)

Figure 13. Comparison of different pseudo-SAR generation strategies for registration and direct registration. (**a**) represents NCM statistics. Among them, there are 4 curves, representing direct registration MatchosNet, CycleGAN + MatchosNet, original Restormer + MatchosNet, and the proposed method. (**b**) represents the RMSE statistics. Among them, there are 4 curves, representing direct registration MatchosNet, CycleGAN + MatchosNet, original Restormer + MatchosNet, and proposed method.



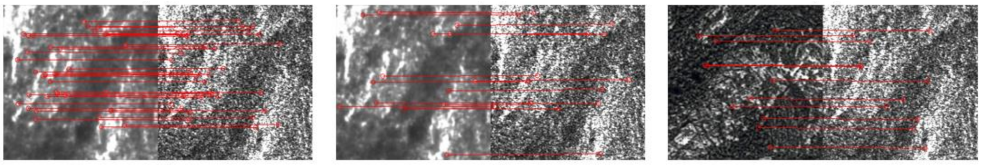
(a1) (a2) (a3)

(b1) (b2) (b3)

(c1) (c2) (c3)

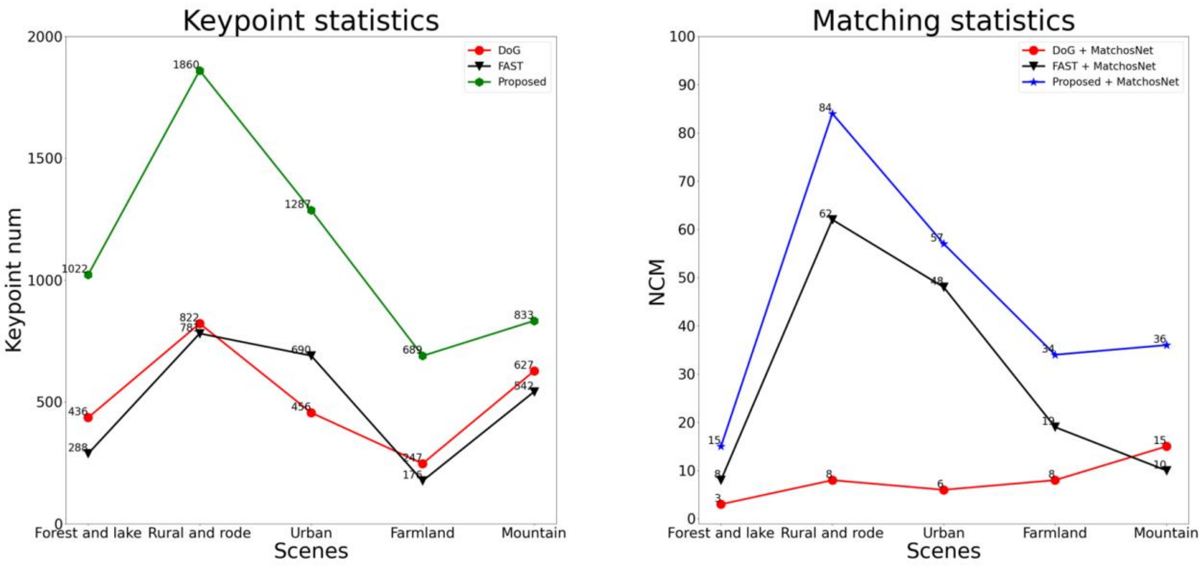
(d1) (d2) (d3)

Figure 14. *Cont.*



(e1) (e2) (e3)

Figure 14. Results of registration methods with different keypoint extraction strategies. (**a1**–**a3**) represent the forest and lake scenes, (**b1**–**b3**) represent the rural and road scenes, (**c1**–**c3**) represent the urban scenes, (**d1**–**d3**) represent the farmland scenes, and (**e1**–**e3**) represent the mountain scenes. (**a1**–**e1**) are the results of the proposed strategy, (**a2**–**e2**) are the results of the DoG strategy, and (**a3**–**e3**) are the results of the FAST strategy.



(a) (b)

Figure 15. Comparison of different keypoint extraction strategies. (**a**) represents keypoints statistics. Among them, there are 3 curves representing the number of keypoints for pseudo-SAR by DoG, FAST, and the proposed strategy. (**b**) represents matching statistics. Among them, there are 3 curves representing the NCM by DoG + MatchosNet, FAST + MatchosNet, and the proposed keypoint extraction strategy + MatchosNet.

4. Discussion

The registration experimental results from Section 3.2 indicate that deep learning-based methods are more robust compared to traditional point matching methods. Comparing the results of the proposed method with CycleGAN + MatchosNet and MatchosNet methods further demonstrates that the registration method based on the Restormer's pseudo-SAR generation strategy improves the accuracy of deep learning models in the registration process. The registration scheme based on the pseudo-SAR generation strategy can avoid the feature differences between heterogeneous images, making the registration network easier to train.

In the conducted ablation experiments, this paper investigated the effectiveness of Restormer's pseudo-SAR generation and a Harris scale space keypoint extraction strategy. The experimental results demonstrate that both of these strategies outperform similar methods. Specifically, compared to similar methods, the proposed Restormer pseudo-SAR generation strategy exhibits smaller RMSE and larger NCM. Contrasting with the original method, L2 loss function is used to instead of L1 loss function, and this improvement has achieved better results in experiments. The proposed keypoints extraction strategy shows a higher number of extracted and matched keypoints. Therefore, relative to other deep learning-based methods, the proposed method has more advantages.

However, in some weak texture scenes, generating pseudo-SAR images may be challenging, which could be a direction for future research in the field of optical and SAR image registration based on pseudo-SAR generation strategy. In other fields of research, such as underwater acoustic or sonar [51,52], Transformer-based simulation may be explored for pseudo-SAS (Synthetic Aperture Sonar) imagery generation.

5. Conclusions

This paper proposes a registration method based on a pseudo-SAR generation strategy. In this approach, Restormer is used to transform an optical image to a pseudo-SAR image. During the training of Restormer, the original loss function is replaced with L2 so that the model fluctuates less at the best fit. In the registration process, the DoG operator is replaced with the ROEWA operator, which is used to construct the Harris scale space for pseudo-SAR and real SAR images, this strategy increases both the extracted and matched keypoints. The extreme points are extracted in each layer of the Harris scale space and added to the keypoint set. The image patches around the keypoints are extracted and fed to MatchosNet to obtain feature descriptors for initial matching, and the RANSAC algorithm is used to remove outliers to obtain the final matching results. The feasibility and robustness of this method have been demonstrated by experiments compared to similar methods.

Author Contributions: Conceptualization, X.S. and C.H.; methodology, R.Z.; software, R.Z. and X.L.; resources, D.X.; writing, R.Z. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62171015, in part by the Fundamental Research Funds for the Central Universities under Grant buctrc202121 and buctrc202218, and in part by the funding provided by the Alexander von Humboldt Foundation.

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

References

1. Le Moigne, J.; Netanyahu, N.S.; Eastman, R.D. *Image Registration for Remote Sensing*; Cambridge University Press: Cambridge, UK, 2011.
2. Zhang, X.; Leng, C.; Hong, Y.; Pei, Z.; Cheng, I.; Basu, A. Multimodal remote sensing image registration methods and advancements: A survey. *Remote Sens.* **2021**, *13*, 5128. [CrossRef]
3. Sotiras, A.; Davatzikos, C.; Paragios, N. Deformable medical image registration: A survey. *IEEE Trans. Med. Imaging* **2013**, *32*, 1153–1190. [CrossRef] [PubMed]
4. Yu, L.; Zhang, D.; Holden, E.-J. A fast and fully automatic registration approach based on point features for multi-source remote-sensing images. *Comput. Geosci.* **2008**, *34*, 838–848. [CrossRef]
5. Ye, F.; Su, Y.; Xiao, H.; Zhao, X.; Min, W. Remote sensing image registration using convolutional neural network features. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 232–236. [CrossRef]
6. Lehureau, G.; Tupin, F.; Tison, C.; Oller, G.; Petit, D. Registration of metric resolution SAR and optical images in urban areas. In Proceedings of the 7th European Conference on Synthetic Aperture Radar, Friedrichshafen, Germany, 2–5 June 2008; pp. 1–4.
7. Yang, W.; Han, C.; Sun, H.; Cao, Y. Registration of high resolution SAR and optical images based on multiple features. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005, IGARSS'05, Seoul, Republic of Korea, 29 July 2005; pp. 3542–3544.
8. Chen, H.-M.; Varshney, P.K.; Arora, M.K. Performance of mutual information similarity measure for registration of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2445–2454. [CrossRef]
9. Luo, J.; Konofagou, E.E. A fast normalized cross-correlation calculation method for motion estimation. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control.* **2010**, *57*, 1347–1357.
10. Zhao, F.; Huang, Q.; Gao, W. Image matching by normalized cross-correlation. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; p. II.
11. Wang, F.; Vemuri, B.C. Non-rigid multi-modal image registration using cross-cumulative residual entropy. *Int. J. Comput. Vis.* **2007**, *74*, 201. [CrossRef]

12. Heinrich, M.P.; Jenkinson, M.; Bhushan, M.; Matin, T.; Gleeson, F.V.; Brady, M.; Schnabel, J.A. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Med. Image Anal.* **2012**, *16*, 1423–1435. [CrossRef]
13. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
14. Xiang, Y.; Wang, F.; You, H. OS-SIFT: A Robust SIFT-Like Algorithm for High-Resolution Optical-to-SAR Image Registration in Suburban Areas. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3078–3090. [CrossRef]
15. Mikolajczyk, K.; Schmid, C. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1615–1630. [CrossRef] [PubMed]
16. Fjortoft, R.; Lopes, A.; Marthon, P. An optimal multiedge detector for SAR image segmentation. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 793–802. [CrossRef]
17. Sedaghat, A.; Mokhtarzade, M.; Ebadi, H. Uniform Robust Scale-Invariant Feature Matching for Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4516–4527. [CrossRef]
18. Ma, T.; Ma, J.; Yu, K.; Zhang, J.; Fu, W. Multispectral Remote Sensing Image Matching via Image Transfer by Regularized Conditional Generative Adversarial Networks and Local Feature. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 351–355. [CrossRef]
19. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3279–3286.
20. Balntas, V.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In Proceedings of the 27th British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; p. 3.
21. Tian, Y.; Fan, B.; Wu, F. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6128–6136.
22. Mishchuk, A.; Mishkin, D.; Radenovic, F.; Matas, J. Working hard to know your neighbor′s margins: Local descriptor learning loss. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
23. Liao, Y.; Di, Y.; Zhou, H.; Li, A.; Liu, J.; Lu, M.; Duan, Q. Feature Matching and Position Matching Between Optical and SAR With Local Deep Feature Descriptor. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 448–462. [CrossRef]
24. 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. [CrossRef]
25. Maggiolo, L.; Solarna, D.; Moser, G.; Serpico, S.B. Registration of Multisensor Images through a Conditional Generative Adversarial Network and a Correlation-Type Similarity Measure. *Remote Sens.* **2022**, *14*, 2811. [CrossRef]
26. Huang, X.; Wen, L.; Ding, J. SAR and optical image registration method based on improved CycleGAN. In Proceedings of the 2019 6th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Xiamen, China, 26–29 November 2019; pp. 1–6.
27. Fan, Y.; Wang, F.; Wang, H. A Transformer-Based Coarse-to-Fine Wide-Swath SAR Image Registration Method under Weak Texture Conditions. *Remote Sens.* **2022**, *14*, 1175. [CrossRef]
28. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach Convention & Entertainment Center, CA, USA, 9–15 June 2019; pp. 7354–7363.
29. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1–8.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Ma, J.; Li, M.; Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Homo–heterogenous transformer learning framework for RS scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2223–2239. [CrossRef]
33. Hao, S.; Wu, B.; Zhao, K.; Ye, Y.; Wang, W. Two-stream swin transformer with differentiable sobel operator for remote sensing image classification. *Remote Sens.* **2022**, *14*, 1507. [CrossRef]
34. Bazi, Y.; Bashmal, L.; Rahhal, M.; Dayil, R.; Ajlan, N. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
35. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
36. Schwind, P.; Suri, S.; Reinartz, P.; Siebert, A. Applicability of the SIFT operator to geometric SAR image registration. *Int. J. Remote Sens.* **2010**, *31*, 1959–1980. [CrossRef]
37. 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. [CrossRef]
38. Bovik, A.C. On detecting edges in speckle imagery. *IEEE Trans. Acoust. Speech Signal Process.* **1988**, *36*, 1618–1627. [CrossRef]
39. Touzi, R.; Lopes, A.; Bousquet, P. A statistical and geometrical edge detector for SAR images. *IEEE Trans. Geosci. Remote Sens.* **1988**, *26*, 764–773. [CrossRef]

40. Du, W.-L.; Zhou, Y.; Zhao, J.; Tian, X.; Yang, Z.; Bian, F. Exploring the potential of unsupervised image synthesis for SAR-optical image matching. *IEEE Access* **2021**, *9*, 71022–71033. [CrossRef]
41. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
42. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
43. An, T.; Zhang, X.; Huo, C.; Xue, B.; Wang, L.; Pan, C. TR-MISR: Multiimage super-resolution based on feature fusion with transformers. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1373–1388. [CrossRef]
44. Ye, C.; Yan, L.; Zhang, Y.; Zhan, J.; Yang, J.; Wang, J. A Super-resolution Method of Remote Sensing Image Using Transformers. In Proceedings of the 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Cracow, Poland, 22–25 September 2021; pp. 905–910.
45. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
46. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.H.; Ieee Comp, S.O.C. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Seattle, WA, USA, 14–19 June 2020; pp. 1571–1580.
47. Huang, M.; Xu, Y.; Qian, L.; Shi, W.; Zhang, Y.; Bao, W.; Wang, N.; Liu, X.; Xiang, X. The QXS-SAROPT Dataset for Deep Learning in SAR-Optical Data Fusion. *arXiv* **2021**, arXiv:2103.08259.
48. Xiang, Y.; Tao, R.; Wang, F.; You, H.; Han, B. Automatic registration of optical and SAR images via improved phase congruency model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5847–5861. [CrossRef]
49. 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. [CrossRef]
50. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
51. Choi, H.-M.; Yang, H.-S.; Seong, W.-J. Compressive underwater sonar imaging with synthetic aperture processing. *Remote Sens.* **2021**, *13*, 1924. [CrossRef]
52. Zhang, X.; Wu, H.; Sun, H.; Ying, W. Multireceiver SAS imagery based on monostatic conversion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10835–10853. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.