remote sensing

MDPI

Article

S^2-PCM: Super-Resolution Structural Point Cloud Matching for High-Accuracy Video-SAR Image Registration

Zhikun Xie [1], Jun Shi [1], Yihang Zhou [1], Xiaqing Yang [2,*], Wenxuan Guo [1] and Xiaoling Zhang [1]

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2] School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: yangxiaqing@uestc.edu.cn

Abstract: In this paper, the super-resolution structural point cloud matching (S^2-PCM) framework is proposed for video synthetic aperture radar (SAR) inter-frame registration, which consists of a feature recurrence super-resolution network (FRSR-Net), structural point cloud extraction network (SPCE-Net) and robust point matching network (RPM-Net). FRSR-Net is implemented by integrating the feature recurrence structure and residual dense block (RDB) for super-resolution enhancement, SPCE-Net is implemented by training a U-Net with data augmentation, and RPM-Net is applied for robust point cloud matching. Experimental results show that compared with the classical SIFT-like algorithms, S^2-PCM achieves higher registration accuracy for video-SAR images under diverse evaluation metrics, such as mutual information (MI), normalized mutual information (NMI), entropy correlation coefficient (ECC), structural similarity (SSIM), etc. The proposed FRSR-Net can significantly improve the quality of video-SAR images and point cloud extraction accuracy. Combining FRSR-Net with S^2-PCM, we can obtain higher inter-frame registration accuracy, which is crucial for moving target detection and shadow tracking.

Keywords: SAR image registration; RPM; super-resolution network; video-SAR

Citation: Xie, Z.; Shi, J.; Zhou, Y.; Yang, X.; Guo, W.; Zhang, X. S^2-PCM: Super-Resolution Structural Point Cloud Matching for High-Accuracy Video-SAR Image Registration. *Remote Sens.* **2022**, *14*, 4302. https://doi.org/10.3390/rs14174302

Academic Editors: Ali Khenchaf and Jean-Christophe Cexus

Received: 17 July 2022
Accepted: 25 August 2022
Published: 1 September 2022

1. Introduction

Video synthetic aperture radar (SAR) is a SAR system capable of high-frame-rate imaging, which enables real-time monitoring of the target area by continuously illuminating the ground target area and processing the received echoes in real-time [1–3]. Due to the unique advantage of all-day all-weather reconnaissance, video-SAR has been widely applied in military and civilian applications [4]. However, due to the existence of IMU measurement errors, there are translation and rotation errors in the video-SAR inter-frame images, which need to be further registered. The registration accuracy directly affects the subsequent processing, such as moving target detection and shadow tracking, etc. [5–7].

Traditional image registration methods can be roughly divided into two categories: intensity-based registration methods and feature-based registration methods [8–10]. The intensity-based registration methods first obtain the grayscale statistics of images, set the similarity criterion, and acquire the correspondence between the images by evaluating the similarity of the corresponding windows of the two images [11]. Widely used similarity criteria include the normalized cross-correlation [12,13] and mutual information [14–16]. However, the required computational cost for these methods is too large in practice.

The feature-based registration methods usually consist of three steps: feature extraction, feature matching and transform parameter estimation [8]. Firstly, the significant features, such as the point features, line features, and area features, are extracted from images. The features are then matched by calculating the similarities between them. Finally, the transformation relationship between images is estimated based on reliable feature pairs. The feature-based registration methods have the advantages of extensive adaptability and

high registration accuracy. Among them, the scale-invariant feature transform (SIFT) [17] descriptor is invariant with respect to scale, rotation and illumination changes, which is one of the most widely used algorithms for optical and radar image registration tasks.

To reduce the dimension of the SIFT descriptor, Ke et al. [18] proposed the PCA-SIFT algorithm by performing a principal component analysis on the feature vector. To suppress the speckle noise in SAR images, Dellinger et al. [19] proposed SAR-SIFT, which uses the exponentially weighted mean ratio to calculate the amplitude and direction of the gradient. The rotation invariance of the SIFT-like algorithms is achieved by assigning a dominant direction [17,20] to the key points. However, the speckle noise in the SAR images greatly affects the calculation of the dominant direction, which severely deteriorates the registration performance [20,21]. In addition, the SIFT-like registration algorithms often suffer from high computational complexity in the face of massive feature points.

In recent years, deep-learning-based methods have achieved great success in the field of image processing, and diverse networks have been applied to image registration. Wang et al. [22] proposed a network that directly learns the mapping between image patch pairs and their matched labels. Han et al. [23] extracted features from the two same convolutional networks to estimate the transformation relationship of the image patches. The excellent performance of deep learning is achieved based on a large number of training samples, which means that a large number of paired image blocks need to be given as training samples. However, it is difficult to obtain a large number of labeled training samples in SAR image registration for the reason that manual labeling of paired SAR image blocks is time-consuming and prone to labeling errors [24].

Apart from the above methods, point cloud matching methods, such as iterative closest point (ICP) [25], Go-ICP [26], deep closest point (DCP) [27], PointNetLK [28], PCRNet [29], and robust point matching network (RPM-Net) [30], etc., provide a new approach for image registration. Point cloud is a massive collection of points reflecting the surface characteristics of an object. Unlike the feature-based methods that construct the matching relationship according to the distance between features, point cloud contains the positions of the points only, and the matching relationship is often obtained by solving an optimization problem.

Since it is hard to extract the features from SAR images due to speckle noise, a point-cloud-based video-SAR image registration framework is proposed based on the state-of-the-art RPM-Net due to its robustness in this paper. To improve the registration accuracy, we extract the structural point cloud between the video-SAR frames via a segmentation technique. Then, RPM-Net is used to match the structural point clouds of SAR frames. Considering that the low-resolution SAR images lead to the inaccurate structural point cloud extraction process, a super-resolution network is designed for better point cloud extraction in this paper. The main contributions of the methodology of this paper can be summarized as follows:

1. A super-resolution structural point cloud matching framework for inter-frame registration of video-SAR is first proposed by integrating FRSR-Net, SPCE-Net with RPM-Net, which can significantly improve the registration accuracy and stableness of SAR images.
2. A feature recurrent super-resolution network for video-SAR image super-resolution is proposed by refining the low-level features with the high-level features through feature recurrence, which is able to achieve elaborate image reconstruction.

The organization of the remaining sections of this paper is as follows. Section 2 reviews the principles of RPM-Net. The proposed S^2-PCM framework is introduced in Section 3. Section 4 introduces the proposed FRSR-Net. Experimental results and analysis are given in Section 5. Finally, the conclusion is given in Section 6.

2. Review on Point Cloud Matching

As one of the most popular point cloud matching methods, the classical point cloud matching algorithms, including RPM and RPM-Net, are first reviewed in this section.

RPM-Net [30] was proposed based on the RPM algorithm [31], which consists of two main modules: a feature extraction network and a parameter prediction network. It uses a differentiable Sinkhorn [32] layer and annealing [33] to obtain soft assignment values corresponding to points from hybrid features extracted from spatial coordinates and local geometry. To further improve the registration accuracy, a secondary network is added to predict the optimal annealing parameters.

Given the source point cloud $\mathbf{X}$ and the reference point cloud $\mathbf{Y}$:

$$\begin{aligned} \mathbf{X} &= \left\{ \mathbf{x}_j \in \mathbb{R}^3 \middle| j = 1, \ldots, J \right\} \\ \mathbf{Y} &= \left\{ \mathbf{y}_k \in \mathbb{R}^3 \middle| k = 1, \ldots, K \right\} \end{aligned} \tag{1}$$

where J and K are the numbers of points of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and J is not necessarily equal to K.

Point cloud registration is to find out the transformation relationship $\{\mathbf{R_p}, \mathbf{t_p}\}$ between $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{R_p} \in SO(3)$ is a rotation matrix and $\mathbf{t_p} \in \mathbb{R}^3$ is a translation vector. To this end, a matching matrix $\mathbf{M} \in \mathbb{R}^{J \times K}$ is defined to represent the correspondence between points, where each element m_{jk} can be determined by the following equation:

$$m_{jk} = \begin{cases} 1 & \text{if point } \mathbf{x}_j \text{ corresponds to } \mathbf{y}_k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The registration problem can be described as finding the $\{\mathbf{R_p}, \mathbf{t_p}\}$ and the $\mathbf{M}$ that optimally maps points in $\mathbf{X}$ to $\mathbf{Y}$, i.e.,

$$\underset{\mathbf{M}, \mathbf{R_p}, \mathbf{t_p}}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{k=1}^{K} m_{jk} \left(\left\| \mathbf{R_p} \mathbf{x}_j + \mathbf{t_p} - \mathbf{y}_k \right\|_2^2 - \alpha \right) \tag{3}$$

where $\sum_{k=1}^{K} m_{jk} = 1, \forall j$, $\sum_{j=1}^{J} m_{jk} = 1, \forall k$, and $m_{jk} \in \{0, 1\}$. These three constraints require that $\mathbf{M}$ must be a permutation matrix. α is a parameter that controls the number of correspondences rejected as outliers.

In RPM, Equation (3) is minimized by two steps: soft assignment and transformation relationship estimation. Firstly, the constraint of the permutation matrix is relaxed to a double random matrix, i.e., $m_{jk} \in [0, 1]$. Therefore, m_{jk} is initialized as:

$$m_{jk} \leftarrow e^{-\beta(\left\| \mathbf{R_p} \mathbf{x}_j + \mathbf{t_p} - \mathbf{y}_k \right\|_2^2 - \alpha)} \tag{4}$$

where β is an annealing parameter in each iteration. Then, the matching matrix $\mathbf{M}$ is estimated in the soft assignment. Once the optimal matching matrix is obtained, the transformation relationship can be computed.

In RPM-Net, the spatial distance in Equation (4) is replaced by the learned hybrid feature distance. In addition, α and β are obtained by network prediction. Specifically, RPM-Net solves the optimal transformation relationship by multiple iterations, as shown in Figure 1. In the ith iteration, the point cloud $\mathbf{X}$ is first transformed into the initial transformed point cloud $\widetilde{\mathbf{X}}^i$ by the transformation $\left\{ \mathbf{R_p}^{i-1}, \mathbf{t_p}^{i-1} \right\}$ estimated in the previous iteration. Then the feature extraction module extracts the hybrid features of $\widetilde{\mathbf{X}}^i$ and $\mathbf{Y}$. Meanwhile, the optimal annealing parameters α and β are predicted by the secondary parameter prediction network. The initial matching matrix $\mathbf{M}$ is calculated using the hybrid features, and parameters α and β, i.e., m_{jk} is initialized as:

$$m_{jk} \leftarrow e^{-\beta(\left\| F_{\widetilde{\mathbf{x}}_j} - F_{\mathbf{y}_k} \right\|_2^2 - \alpha)} \tag{5}$$

where $F_{\widetilde{\mathbf{x}}_j}$ and $F_{\mathbf{y}_k}$ denote the hybrid features of points $\widetilde{\mathbf{x}}_j \in \widetilde{\mathbf{X}}^i$ and $\mathbf{y}_k \in \mathbf{Y}$ learned by the feature extraction network, respectively. Then the final matching matrix $\mathbf{M}^i$ can be obtained

by Sinkhorn normalization with enhanced double random constraints. Next, for each point $\mathbf{x}_j$ in $\mathbf{X}$, its corresponding coordinate in $\mathbf{Y}$ is calculated as:

$$\hat{\mathbf{y}}_j = \frac{1}{\sum_k^K m_{jk}} \sum_k^K m_{jk} \bullet \mathbf{y}_k \tag{6}$$

Finally, the transformation relationship is solved by SVD [34].

Since RPM-Net uses the learned feature distances instead of spatial distances to initialize the matching matrix, it avoids the initialization sensitivity and local minimum problems and is able to handle partial-to-partial point cloud registration. In addition, it improves the robustness to outliers by progressively reinforcing the soft assignment of point correspondences through Sinkhorn [32] and annealing [33], which is applied in this paper.

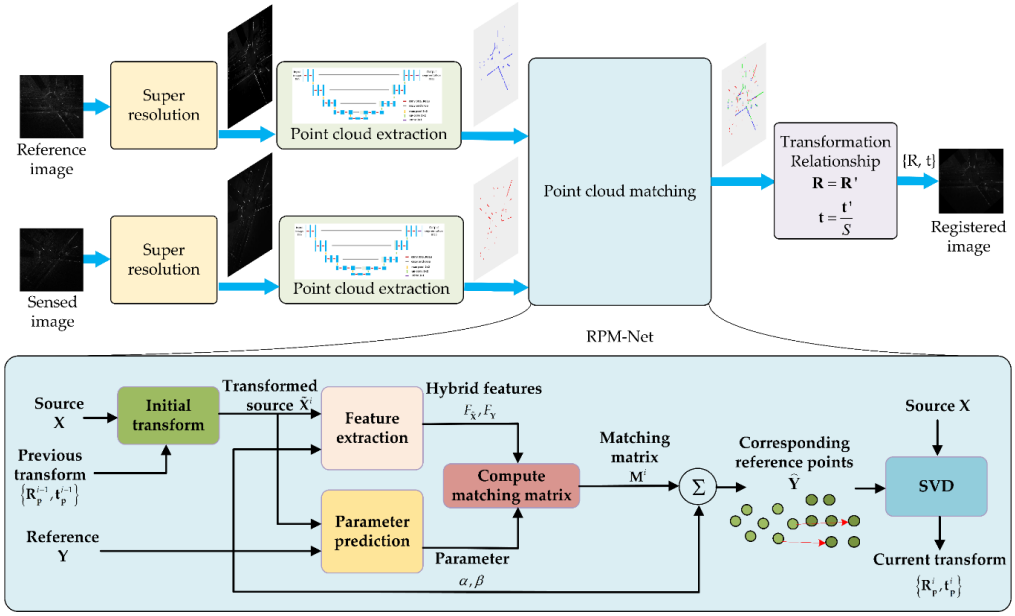

Figure 1. Super-resolution structural point cloud matching (S^2-PCM) framework.

3. Super-Resolution Structural Point Cloud Matching Framework

Based on RPM-net, we present the framework of the super-resolution structural point cloud matching algorithm in detail in this section. In order to use the point clouds to register the SAR images, the point clouds first need to be extracted, which can be considered a segmentation problem in deep learning. Furthermore, to ensure the performance of the registration, the positions of the extracted point clouds should be accurate. However, for the actual SAR system, the images suffer from low resolution, speckle noise, blur and defocusing, which greatly reduce the image quality and lead to the difficulty and inaccuracy of the extraction of point clouds. Therefore, a novel super-resolution structural point cloud matching framework is presented in this section for better performance.

3.1. Network Structure

The proposed super-resolution structural point cloud matching (S^2-PCM) framework for inter-frame registration of video-SAR consists of a super-resolution network, structural point cloud extraction network (SPCE-Net) and RPM-Net, as shown in Figure 1.

Firstly, the reference image I_{ref} and the sensed image I_{sen} are fed into a super-resolution network to improve the resolution and suppress the speckle noise. Most of the current super-resolution networks [35–40] are designed for optical images, but there are great differences between the textures of optical images and SAR images. Through experiments, it is found that the super-resolution network for optical images directly applying to SAR

images results in poor performance. For this reason, we propose a super-resolution network, the details of which will be given in Section 4.

After image enhancement, feature point cloud extraction is required for SAR images. Extracting the point cloud can be achieved by fixing a threshold and selecting the points with higher intensity. It can also be achieved by segmentation. However, due to the different observation angles of the circular SAR system and the effect of speckle noise, the intensity of scattered points varies greatly among different images. The point clouds extracted by intensity-based methods vary greatly and reduce the registration accuracy significantly. Thus, we use the segmentation technique to extract the structural point clouds, i.e., point clouds reflecting the geometric structure of video-SAR frames. Firstly, we manually label the typical region edges in the video-SAR image to train a structural point cloud extraction network (SPCE-Net implemented by U-Net [41]). Secondly, the images are fed into the trained SPCE-Net to obtain the segmented masks. Then, the coordinates are extracted from the masks to generate the initial point clouds. To reduce the computational cost of the point cloud matching process, we down-sample the initial point clouds to generate the final structural point clouds via random sampling. Each structural point cloud contains approximately 1000 points. Too few points do not reflect the geometric structure of the image, which reduces the registration accuracy.

We use the trained SPCE-Net to segment the video-SAR frames to generate a sequence of point clouds, which is applied as the training dataset for RPM-Net. Then, we use the trained RPM-Net to match the structural point clouds. In order to achieve higher matching accuracy, RPM-Net requires at least five iterations. It is worth noting that since RPM-Net is a 3D point cloud registration network, we adapt the 2D point clouds extracted from the SAR images to RPM-Net by expanding a new dimension and setting it to 0. By feeding the structural point clouds into RPM-Net, we can obtain the transformation relationship $\{\mathbf{R}', \mathbf{t}'\}$ between images, where $\mathbf{R}' \in SO(3)$ is a rotation matrix and $\mathbf{t}' \in \mathbb{R}^3$ is a translation vector. Due to the scaling of super-resolution processing, the transformation relationship estimated by RPM-Net is not equal to the original one, which should be adjusted before use.

3.2. Transformation Relationship

Assume that the transformation relationship between I_{sen} and I_{ref} is $\{\mathbf{R}, \mathbf{t}\}$, where $\mathbf{R} \in SO(3)$ is a rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is a translation vector. According to [42], $\mathbf{R}$, $\mathbf{t}$ can be expressed as:

$$\begin{aligned} \mathbf{R} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \cos\psi & 0 & -\sin\psi \\ 0 & 1 & 0 \\ \sin\psi & 0 & \cos\psi \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ \mathbf{t} &= \begin{pmatrix} t_x, t_y, t_z \end{pmatrix}^T \end{aligned} \tag{7}$$

where ϕ, ψ, and θ denote the rotation angles around the x-axis, y-axis, and z-axis, respectively, and t_x, t_y, and t_z denote the translations in the x-, y-, and z-directions, respectively. Since the sensed image I_{sen} and the reference image I_{ref} are located only in the x–y plane, it can be obtained that]ϕ and ψ are 0 and t_z is 0. Equation (7) can be simplified as:

$$\begin{aligned} \mathbf{R} &= \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ \mathbf{t} &= \begin{pmatrix} t_x, t_y, 0 \end{pmatrix}^T \end{aligned} \tag{8}$$

Take any point $A(x, y, 0)$ from I_{sen} and assume that its corresponding point in I_{ref} is $B(x', y', 0)$, and then the relationship between A and B can be expressed as:

$$\begin{pmatrix} x' \\ y' \\ 0 \end{pmatrix} = \mathbf{R} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} + \mathbf{t} = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ 0 \end{pmatrix} \tag{9}$$

$$\cos\theta = \frac{\vec{OA} \bullet \vec{OB}}{\left|\vec{OA}\right|\left|\vec{OB}\right|} \tag{10}$$

where O is the coordinate origin.

Given the scaling factor S of super-resolution, the points $A(x, y, 0)$ and $B(x', y', 0)$ after scale transformation are $A'(Sx, Sy, 0)$ and $B'(Sx', Sy', 0)$, respectively. Thus, the relationship between A' and B' can be expressed as:

$$\begin{pmatrix} Sx' \\ Sy' \\ 0 \end{pmatrix} = \mathbf{R}' \begin{pmatrix} Sx \\ Sy \\ 0 \end{pmatrix} + \mathbf{t}' \tag{11}$$

Similar to Equation (10), we can obtain that the rotation angle around the z-axis of I_{ref} and I_{sen} after super-resolution is still θ. Thus, combining Equations (9) and (11), it can be obtained that:

$$\begin{aligned} \mathbf{R} &= \mathbf{R}' \\ \mathbf{t} &= \frac{\mathbf{t}'}{S} \end{aligned} \tag{12}$$

In a word, the transformation relationship $\{\mathbf{R}', \mathbf{t}'\}$ with a scaling operator between I_{ref} and I_{sen} is acquired by registration, and the transformation relationship {R, t} without a scaling operator between the original I_{ref} and I_{sen} can then be obtained according to Equation (12).

3.3. Training Strategy

Since it is hard to acquire noise-free high-resolution SAR images, we use optical images to construct the dataset for training a super-resolution network by data augmentation. Firstly, optical images are converted to grayscale images as high-resolution images for training. The high-resolution images are Bicubic [43] down-sampled and then blurred with a point spread function (PSF). Then, speckle noise of level L is added to construct low-resolution images. As a result, multiple pairs of training data are generated. We choose L1 loss for training the network, which is defined as:

$$L_{SR} = \sum_{y=1}^{W} \sum_{x=1}^{H} |I_{SR}(x, y) - I_{HR}(x, y)| \tag{13}$$

where I_{HR} and I_{SR} denote the high-resolution image and the super-resolution image, respectively. W and H are the width and height of the image, respectively.

To train SPCE-Net, we label the images manually. Then the same affine transformation is applied to the images and the corresponding labels, thus achieving data augmentation and reducing the workload of image labeling. The dice coefficient [44] is used as the loss function for training SPCE-Net. It is an ensemble similarity measure function and calculated as follows:

$$Dice = \frac{2|T \cap P|}{|T| + |P|} \tag{14}$$

where T denotes the set of segmented true labels and P denotes the set of output predicted values. To calculate $|T \cap P|$, it is approximated as the sum of the dot product of T and P. $|T| + |P|$ denotes a direct summation over all elements of T and P. Thus, *Diceloss* can be obtained:

$$Diceloss = 1 - Dice \tag{15}$$

The overall loss L_{seg} uses the sum of *Diceloss* and the cross-entropy loss E:

$$L_{seg} = Diceloss + E \tag{16}$$

For training RPM-Net, we use the original point cloud as the source point cloud **X**. Then, we generate a random rotation angle around the z-axis in the range of $[-90°, 90°]$ and a random translation vector in the range of $[-300, 300]$ pixel on the x and y axes, which results in a transformation relationship $\{\mathbf{R_{gt}}, \mathbf{t_{gt}}\}$. This transformation relationship is applied to the source point cloud **X** to obtain the reference point cloud **Y**. For each point cloud, a hyperplane is randomly generated to sample a half-space, and it is continuously shifted so that 70% of the points are retained. Finally, Gaussian noise is added. We choose the L1 distance as the loss for training RPM-Net, which is defined as:

$$L_{reg} = \frac{1}{J}\sum_{j}^{J}\left|\left(\mathbf{R_{gt}}\mathbf{x}_j + \mathbf{t_{gt}}\right) - \left(\mathbf{R_{pred}}\mathbf{x}_j + \mathbf{t_{pred}}\right)\right| \tag{17}$$

where $\{\mathbf{R_{pred}}, \mathbf{t_{pred}}\}$ is the prediction transformation.

4. Feature Recurrence Super-Resolution Network

In this section, we present our proposed feature recurrence super-resolution network in detail.

Most current super-resolution networks [35–40] for optical images reconstruct high-resolution images with only a single prediction, ignoring the connection between higher-level information and lower-level information. Since the resolution degradation of SAR images is affected by several factors, a single prediction may not be able to accurately recover the detailed information of the image. To this end, we design a feature recurrence super-resolution network (FRSR-Net) using the residual dense block (RDB) [40]. FRSR-Net utilizes a recurrence structure to refine low-level features with high-level features for better reconstruction of high-resolution SAR images, and the recurrence structure is able to reduce network parameters.

4.1. Super-Resolution Sub-Network via Feature Recurrence

The proposed FRSR-Net contains three modules: feature extraction, feature enhancement, and image reconstruction, as shown in Figure 2. The feature extraction module extracts the low-level features of the image. The feature enhancement module is mainly constructed by RDB, where RDB makes full use of the features of all convolutional layers within the block and establishes feature-to-feature associations with dense connections. Then, the local features generated by different RDBs are fused to generate high-level features. The image reconstruction module recovers high-resolution images by sub-pixel convolution [37].

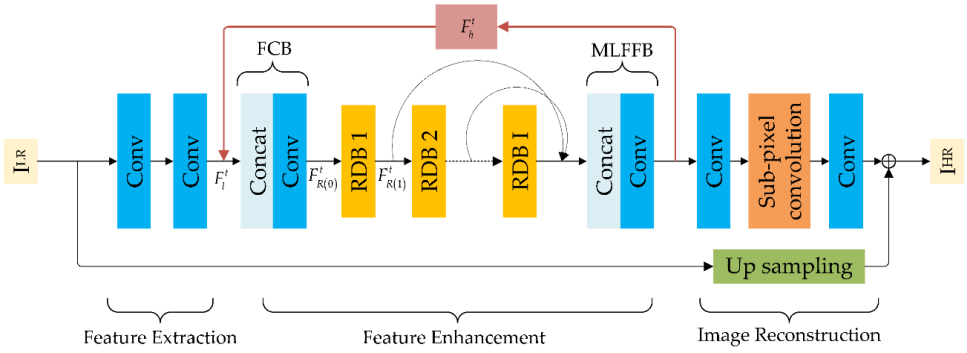

Figure 2. The architecture of the feature recurrence super-resolution network (FRSR-Net).

The feature extraction module consists of two cascaded convolutional layers, and a ReLU activation function is connected after each convolutional layer. At the tth iteration, the low-level feature F_l^t, output by the feature extraction network, can be expressed as:

$$F_l^t = f_{FE}(I_{LR}) \tag{18}$$

where $f_{FE}(\bullet)$ denotes the feature extraction module, and I_{LR} is the input of a low-resolution image.

The feature enhancement module consists of a feature compression block (FCB) cascaded with I RDBs, and a multi-level feature fusion block (MLFFB) is added last. At the tth iteration, the high-level feature F_h^t is obtained by feeding the low-level feature F_l^t and the hidden feature F_h^{t-1} output from the feature enhancement module in the previous iteration into the feature enhancement module, which can be expressed by the following equation:

$$F_h^t = f_{EN}\left(F_l^t, F_h^{t-1}\right) \tag{19}$$

where $f_{EN}(\bullet)$ denotes the feature enhancement module. At the initial iteration, F_h^{t-1} is initialized to F_l^t. Specifically, F_l^t is concatenated with F_h^{t-1} as the input to FCB to obtain the feature $F_{R(0)}^t$ as the input to RDB 1. Then, the output of the previous RDB is used as the input to the next RDB, i.e.,

$$\begin{aligned} F_{R(0)}^t &= f_{FCB}\left(F_l^t, F_h^{t-1}\right) \\ F_{R(i)}^t &= f_{RDB,i}\left(F_{R(i-1)}^t\right) = f_{RDB,i}\left(f_{RDB,i-1}\left(\cdots\left(f_{RDB,1}\left(F_{R(0)}^t\right)\right)\cdots\right)\right) \end{aligned} \tag{20}$$

where $f_{FCB}(\bullet)$ denotes FCB, $f_{RDB,i}(\bullet)$ denotes the ith RDB, and $F_{R(i)}^t$ denotes the local features of the output of the ith RDB. Finally, the MLFFB connects all the local features and applies a convolution operation to obtain the high-level feature F_h^t, i.e.,

$$F_h^t = f_{MLFFB}\left(F_{R(1)}^t, F_{R(2)}^t, \ldots, F_{R(I)}^t\right) \tag{21}$$

where $f_{MLFFB}(\bullet)$ denotes MLFFB.

The image reconstruction module consists of two convolutional layers and a sub-pixel convolution block [37], and more details of sub-pixel convolution will be introduced in Section 4.3. F_h^t is used as the input of the first convolution layer, and then the feature map is up-sampled by a subpixel convolution block. Finally, I_{LR} is bilinearly interpolated and added to the output of the image reconstruction module to obtain the reconstruction output I_{SR}, i.e.,

$$I_{SR}^t = f_{IR}(F_h^t) + f_{up}(I_{LR}) \tag{22}$$

where $f_{IR}(\bullet)$ denotes the image reconstruction module and $f_{up}(\bullet)$ denotes the bilinear up-sampling.

After T_a rounds of iterations, the output of the T_ath iteration is taken as the final reconstructed image.

4.2. Residual Dense Block

Since the details of SAR images are obscured by the strong speckle noise, the standard convolutional networks that are composed of simple chain stacking may ignore some information about each convolutional layer. On the contrary, thee RDB proposed by Zhang et al. [40] uses dense connection and residual connection, which fully utilizes the information of the convolutional layer within the block and has the stable characteristic to extract the detailed information of SAR images. RDB mainly contains dense connection layers, local feature fusion and local residual learning, as shown in Figure 3.

There are N_c convolutional layers within each RDB, and each convolutional layer except for the last layer is followed by a ReLU. The dense connection is reflected in the interconnection between the convolutional layers, i.e., the input of the nth convolutional layer in RDB is the output of the previous (n − 1) convolutional layers connected to the

input of RDB. Therefore, the output $F_{R(i,n)}$ of the nth convolutional layer of the ith RDB can be expressed as:

$$F_{R(i,n)} = \sigma\left(W_{i,n}\left[F_{R(i-1)}, F_{R(i,1)}, \ldots, F_{R(i,n-1)}\right] + b_{i,n}\right) \tag{23}$$

where σ denotes the linear activation unit, and $W_{i,n}$ and $b_{i,n}$ denote the weight and bias of the nth convolutional layer within the ith RDB, respectively.

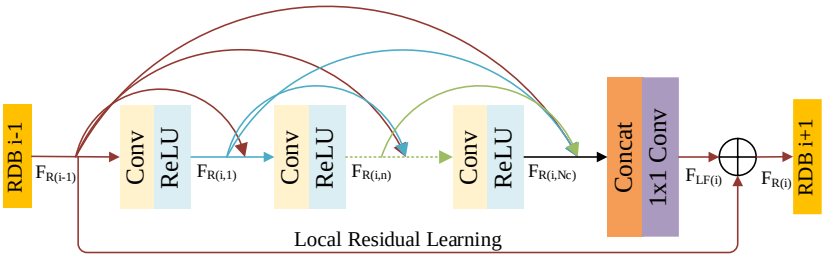

Figure 3. The architecture of residual dense block (RDB).

Then, local feature fusion is performed. The output of all convolutional layers within the current ith RDB and the input of RDB are connected into a 1×1 convolutional block, and the output is obtained as the feature $F_{LF(i)}$:

$$F_{LF(i)} = f_{conv,i}\left(\left[F_{R(i-1)}, F_{R(i,1)}, F_{R(i,2)}, \ldots, F_{R(i,N_c)}\right]\right) \tag{24}$$

where $f_{conv,i}(\bullet)$ denotes the last convolution operation of the ith RDB. In this way, the local features of the previous RDB flow into the next RDB by direct concatenation, which can greatly reduce the number of features and achieve the full utilization of features.

Finally, local residual learning is utilized to better improve the information flow. The local feature $F_{R(i)}$ of the output of the ith RDB can be expressed as:

$$F_{R(i)} = F_{R(i-1)} + F_{LF(i)} \tag{25}$$

4.3. Sub-Pixel Convolution

In super-resolution tasks, the commonly used image up-sampling methods are interpolation, deconvolution, and sub-pixel convolution [37]. The parameters of interpolation-based up-sampling methods are not obtainable by learning, while the deconvolution tends to introduce a checkerboard effect to the image. In order to ensure the trainability of the reconstruction parameters and achieve efficient up-sampling, sub-pixel convolution is applied here.

Sub-pixel convolution, also called pixel shuffling, exploits a channel-to-space conversion method to achieve spatial magnification by rearranging the pixels in multiple channels of the feature map, as shown in Figure 4. Suppose we need to up-sample the feature map $F_{in} \in \mathbb{R}^{W/r \times H/r \times cr^2}$ by a factor of r, where W/r, H/r, and cr^2 denote the width, height, and the number of channels of the feature map, respectively. Sub-pixel convolution is used to obtain the output:

$$F_{out}(x, y, c) = F_{in}(\lfloor x/r \rfloor, \lfloor y/r \rfloor, c \times r \times \mathrm{mod}(y, r) + c \times \mathrm{mod}(x, r) + c) \tag{26}$$

where $F_{out}(x, y, c)$ is the value at position (x, y) on the cth channel of the output, $\mathrm{mod}(\bullet)$ denotes the remainder operation, and $\lfloor \bullet \rfloor$ represents the function to find the maximum integer that does not exceed the given input. For example, a feature map of size $32 \times 32 \times 4$ is up-sampled twice, the number of feature channels is reduced after pixel shuffling, and the elements are rearranged to obtain an output of $64 \times 64 \times 1$.

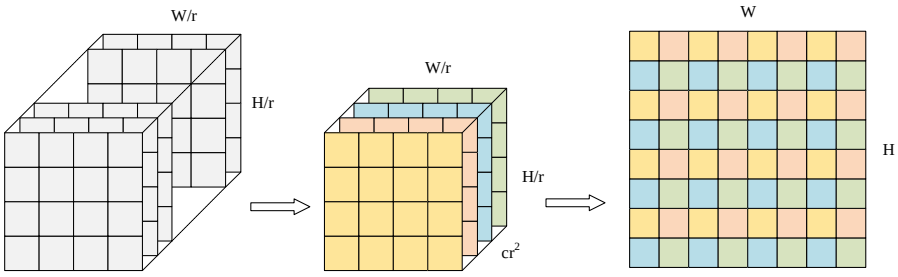

Figure 4. The sub-pixel convolutional operation.

5. Experiments and Results

5.1. Dataset and Setting

DIV2K [40] is a common dataset for super-resolution tasks in optics, containing a total of 1000 high-resolution images, of which 800 were used for training, 100 for validation and 100 for testing. Therefore, according to the strategy in Section 3.3, we used 800 training images from DIV2K to construct the dataset for training FRSR-Net and used five validation images during the training process. In the down-sampling, scaling factors of two, three, and four were applied, respectively. For testing, we selected several standard test datasets for the super-resolution task, including Set14 [40], Manga109 [40], and some images from the test set of DIV2K. Among them, set14 contains 14 optical images and Manga109 consists of 109 manga, which are commonly used as test datasets in super-resolution tasks. The test datasets were prepared by the same processing steps as the training dataset.

For training SPCE-Net and RPM-Net, we selected a simulated video-SAR dataset and a real video-SAR dataset. The real video-SAR dataset contains 500 images, and the simulated video-SAR data contains 500 images. The networks were trained with these two datasets separately.

To test the registration performance of our S^2-PCM, we selected three test datasets containing a video-SAR simulation dataset and two video-SAR real datasets. Each test dataset contains 50 continuous frames. In the later subsection, we refer to the two real datasets as real dataset 1 and real dataset 2, respectively.

In the super-resolution experiment, the number of iterations was set to 4, and the number of RDBs was 8 in FRSR-Net. In the registration experiments, the scaling factor of super-resolution was set to 3, and the registration of each dataset was taken by registering the later images to the initial frame. The whole experiment was implemented with the Pytorch open-source framework and trained with the Intel i7-8700 CPU and NVIDIA GTX-1080 (8G) GPU hardware platform.

5.2. Evaluation Metrics

5.2.1. Super-Resolution Evaluation Metrics

We used the widely used peak signal-to-noise ratio (PSNR) [45] and structural similarity (SSIM) [46] to evaluate the super-resolution performance. PSNR is defined as follows:

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSD}\right) \tag{27}$$

where MAX_I is the maximum pixel value in the image, and MSD is the mean squared difference between the two images.

SSIM is a metric that measures the degree of similarity between images. It is more consistent with the human eye's judgment of image quality compared with PSNR, which is defined as follows:

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{28}$$

where μ_x, μ_y are the means of x and y, σ_x and σ_y are the standard deviations of x and y, σ_{xy} is the covariance of x and y, and C_1 and C_2 are constants.

5.2.2. Registration Evaluation Metrics

In order to quantitatively evaluate the registration performance of S^2-PCM, the following evaluation metrics [47,48] were applied. Firstly, we used the symbols M and N to denote the random variables of the statistical features of the reference image and the registered image, respectively.

1. Pearson correlation coefficient (PCC):

$$PCC = \frac{1}{l-1}\sum_{i=1}^{l}\left(\frac{m_i-\mu_M}{\sigma_M}\right)\left(\frac{n_i-\mu_N}{\sigma_N}\right) \tag{29}$$

where m_i and n_i denote the realizations of the random variables M and N, l is the length of the random variable, μ_M, μ_N are the means of M and N, and σ_M and σ_N are the standard deviations of M, N.

2. Mean squared differences (MSD):

$$MSD = \frac{1}{l}\sum_{i=1}^{l}(m_i-n_i)^2 \tag{30}$$

where m_i, n_i, and l are the same as in Equation (29).

3. Mutual information (MI):

$$MI = I(M,N) = \sum_{m\in M}\sum_{n\in N} p_{M,N}(m,n)\log\frac{p_{M,N}(m,n)}{p_M(m)p_N(n)} \tag{31}$$

where $p_M(m)$ and $p_N(n)$ denote the one-dimensional probability densities of the normalized histograms of M and N, respectively. $p_{M,N}(m,n)$ denotes the two-dimensional joint probability densities of the normalized joint histograms of M and N.

4. Normalized mutual information (NMI):

$$\begin{aligned} NMI &= \frac{H(M)+H(N)}{H(M,N)} \\ H(M,N) &= -\sum p_{M,N}(m,n)\log(p_{M,N}(m,n)) \\ H(M) &= -\sum p_M(m)\log(p_M(m)) \\ H(N) &= -\sum p_N(n)\log(p_N(n)) \end{aligned} \tag{32}$$

where $p_M(m)$, $p_N(n)$, and $p_{M,N}(m,n)$ are the same as in the above equation, *H(M)* and *H(N)* are the entropies of M and N, respectively, and *H(M, N)* is the joint entropy of the pair *(M, N)*.

5. Entropy correlation coefficient (ECC):

$$ECC = \sqrt{\frac{2I(M,N)}{H(M)+H(N)}} \tag{33}$$

where *I(M, N)*, *H(M)*, and *H(N)* are given by the above equation.

6. SSIM is given in Section 5.2.1.

Among the above evaluation metrics, MI, NMI, and ECC are information theory-based evaluation methods, and MSD, PCC, and SSIM are statistical-based evaluation methods.

5.3. Analysis of Super-Resolution Performance

In this sub-section, we verify and analyze the super-resolution performance of FRSR-Net. We compare FRSR-Net with two other super-resolution methods. One is the traditional Bicubic algorithm [43], and the other is the deep learning-based RDN [40]. We train RDN with the same dataset.

Firstly, we verify the super-resolution performance with three test datasets, and the experimental results of super-resolution with different scaling factors are shown in Table 1. It can be seen from Table 1 that all the evaluation metrics of FRSR-Net are superior to the rest of the compared methods, and it achieves the best performance. The test results of RDN are the second, and Bicubic is the worst.

Table 1. Average PSNR/SSIM values.

Dataset	Scaling Factor	Bicubic	RDN	FRSR-Net
Set14	2	16.8536/0.1850	23.1123/0.5920	**23.3392/0.6015**
	3	16.3669/0.1696	21.5473/0.5301	**21.7953/0.5406**
	4	16.0525/0.1734	20.4009/0.4907	**20.6160/0.4986**
Manga109	2	15.2934/0.1592	21.8103/0.7003	**22.3246/0.7189**
	3	14.7603/0.1473	20.0737/0.6393	**20.3832/0.6554**
	4	14.4211/0.1583	18.9417/0.6042	**19.1847/0.6156**
DIV2K	2	17.3878/0.1789	25.2677/0.6623	**25.5155/0.6679**
	3	17.0843/0.1766	23.9786/0.6190	**24.1894/0.6245**
	4	16.9035/0.1918	23.0580/0.5954	**23.2885/0.6007**

FRSR-Net takes full advantage of the features of different levels of RDB and adopts a recurrence structure to refine the low-level features with the high-level output features, which results in better reconstruction performance. Since Bicubic is not able to learn the mapping from low-resolution images to high-resolution images, it has the worst test results. Due to the use of the recurrence structure, FRSR-Net applies a smaller number of RDBs compared with RDN, which results in a smaller number of parameters for the network. With a super-resolution scaling factor of three, the number of network parameters of FRSR-Net is about 1.7 million, and that of RDN is about 3.6 million. It can be seen that the number of parameters used in our method is less than 50% of that of RDN.

The calculation of the above metrics requires noise-free high-resolution images as reference, but it is not possible to obtain real SAR images without the influence of noise. In order to verify the super-resolution performance of our method on real SAR images, a real SAR image is applied to super-resolution with a scaling factor of three. We qualitatively compare the super-resolution results of the FRSR-Net with the above comparison methods, as shown in Figure 5.

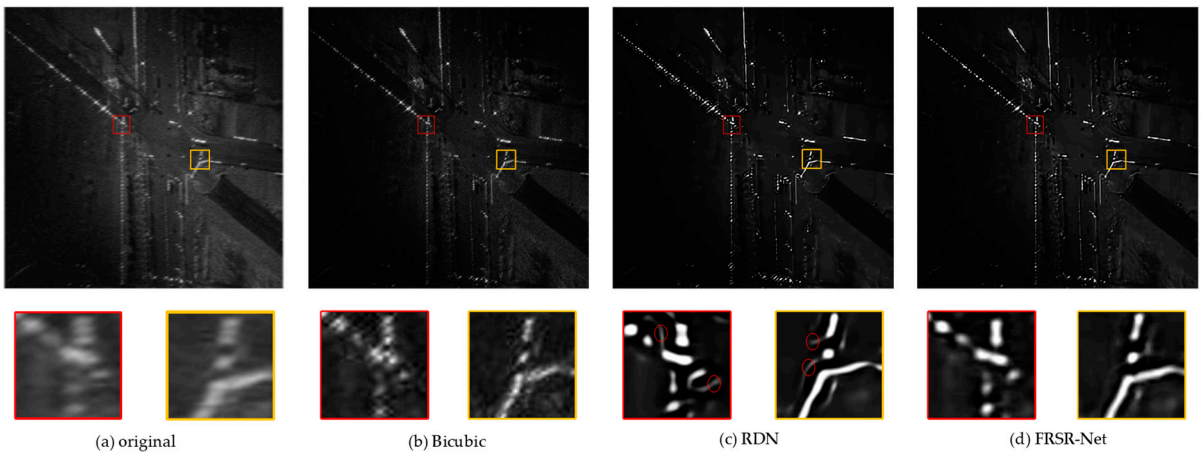

Figure 5. Visual results of super-resolution.

It can be seen from Figure 5 that, compared with the remaining two methods, our method achieves the best visual effect. Specifically, some details of the image are recovered in the results, the edges of the image are sharpened, and the speckle noise is suppressed. Compared with the original image, the RDN recovered image detail is not accurate enough, and it generates some additional textures, as shown by the red circles in the enlarged subfigure. The super-resolution effect of Bicubic is the worst, which only enlarges the image and cannot recover image details and suppress speckle noise.

5.4. Analysis of Registration Performance

In this sub-section, we verify the registration performance of our S^2-PCM on a simulated dataset and real dataset, respectively. We compare our S^2-PCM with three traditional algorithms, including SIFT, SAR-SIFT and PSO-SIFT [49]. We match the feature points using the nearest neighbor distance ratio (NNDR) with Euclidean distance and eliminate all incorrectly matched feature points using the Fast Sample Consensus (FSC) algorithm [50], which divides the candidate objects in Random Sample Consensus (RANSAC) [51] into two parts: the sample set with a high correct rate and the consensus set with large correct matches. Compared with RANSAC, FSC can achieve a greater number of correct feature point matching with fewer iterations.

5.4.1. Registration Results of Simulation Dataset

The registration results of the simulated dataset are shown in Table 2. It can be seen that the proposed S^2-PCM method is superior to the compared methods with most of the metrics. However, due to the simplicity of the simulated scenes, the registration performance of SAR-SIFT and PSO-SIFT is acceptable with most image pairs, and the proposed method has only a slight superiority over them. Since the moving targets' shadows in the simulated SAR images are vivid and similar, we find that the three comparison methods regard the moving targets' shadows as feature descriptors. Since the moving targets' shadows are always moving, there are some error conversion relationships, which lead to the degradation of the registration performance.

Table 2. Comparison of SIFT, SAR-SIFT, PSO-SIFT and the proposed method on the simulated dataset.

	MI	**NMI**	**ECC**	**MSD**	**PCC**	**SSIM**
SIFT	4.0214	1.5011	0.8167	758.6585	0.5180	0.2684
SAR-SIFT	4.2285	1.5053	0.8192	622.5073	0.5566	0.2867
PSO-SIFT	4.2582	1.5031	0.8179	616.7051	0.5426	**0.2929**
Proposed	**4.2764**	**1.5058**	**0.8196**	**551.6772**	**0.5921**	0.2702

As shown in Figure 6, we select two representative images to show the registration results of our S^2-PCM. Figure 6c,f shows the matching results of the structural point clouds, where the red points denote the structural point clouds of the sensed images, the blue points are the structural point clouds of the reference image, and the green points are the point clouds after being matched. The green point cloud is completely overlapped with the blue point cloud and has a good registration result. Figure 6d,g is the checkerboard mosaicked images [24] after registration, and the brightness of the images is adjusted. The edge continuity between the registered image and the reference image can be seen in the figure, which indicates the excellent registration performance of our method.

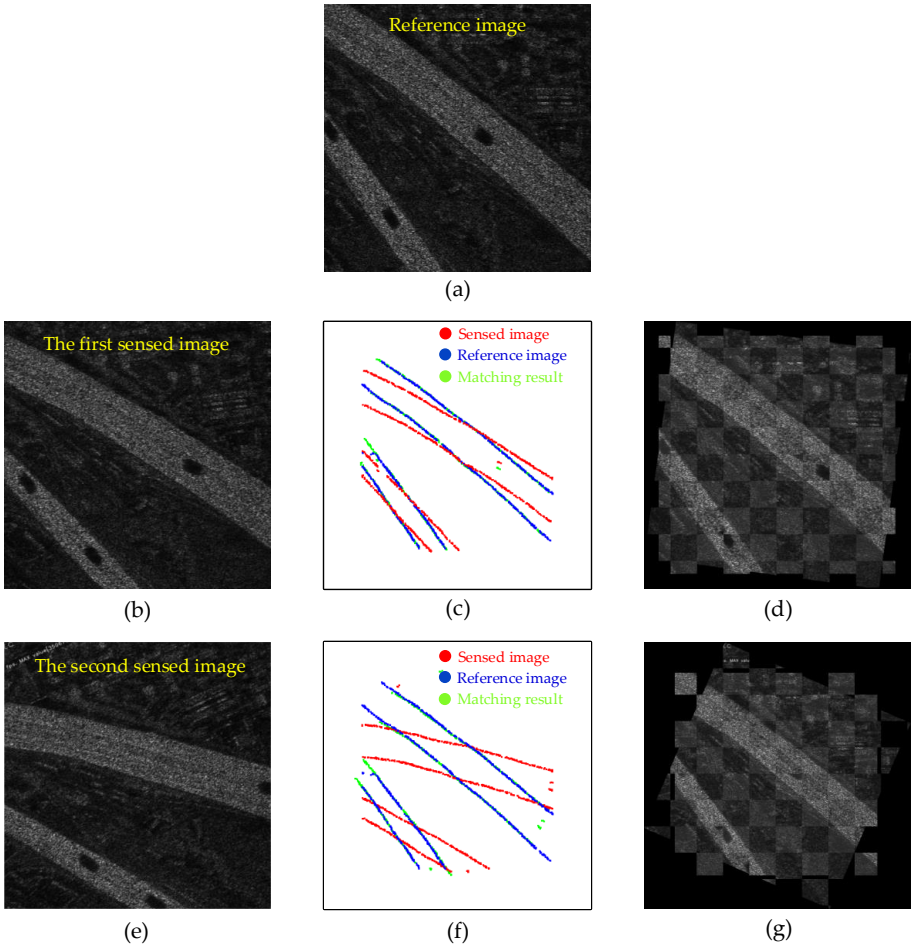

Figure 6. Registration results on simulated dataset. (**a**) Reference image. (**b**) The first sensed image. (**c**) Point cloud matching result of the first sensed image. (**d**) Checkerboard mosaicked image of the first sensed image. (**e**) The second sensed image. (**f**) Point cloud matching result of the second sensed image. (**g**) Checkerboard mosaicked image of the second sensed image.

5.4.2. Registration Results of Real Datasets

In the two real test datasets, the images' quality of real dataset 1 is better than that of real dataset 2, and some images of real dataset 2 appear defocused. Both have large speckle noise and brightness variation between images. A comparison of the registration results is shown in Tables 3 and 4. We can observe from the tables that the registration performance of our method is superior to other comparison methods on both datasets under MI, NMI, ECC, MSD, PCC, and SSIM evaluation metrics. Moreover, the registration performance on dataset 2 is significantly better than the comparison methods, which indicates that the S^2-PCM method is relatively robust to image defocusing. In summary, our method uses structural point clouds, which reflect the geometric characteristics of images, as matching features. Due to the enhancement of image details and suppression of speckle noise by super-resolution, the structural point clouds are extracted more accurately, which is beneficial for improving registration accuracy. In addition, the RPM-Net, which is robust to outliers and desensitized to initialization, is used to match the point clouds, which further improves registration accuracy.

As mentioned in reference [20,21], the calculation of the dominant direction of the SIFT algorithm is strongly influenced by speckle noise, which will affect the matching performance of the feature descriptors. During experiments, we find that due to the texture characteristic of the SAR image, there are sometimes only a few correctly matched feature point pairs (less than four feature point pairs) persevered by the SIFT-like algorithms after the FSC algorithm. In this case, the registration fails.

To demonstrate the registration performance of our S^2-PCM more intuitively, for each dataset, we select two representative sensed images and draw the registration results, as shown in Figures 7 and 8, where Figure 7 shows the results of real dataset 1 and Figure 8 shows the results of real dataset 2. The matching results of the point clouds are shown in Figures 7c,f and 8c,f. It can be seen that the green point cloud has almost been overlapped with the blue point cloud, which shows a good matching performance. Figures 7d,g and 8d,g are the checkerboard mosaicked images after registration, and we zoom in for the details in the image. From the figures, we can see that the edges of the image after registration and the reference image are continuous, and the regions are overlapped well, which indicates that our S^2-PCM method is able to achieve excellent registration performance.

Table 3. Comparison of SIFT, SAR-SIFT, PSO-SIFT and proposed method on real dataset 1.

	MI	NMI	ECC	MSD	PCC	SSIM
SIFT	3.6440	1.5976	0.8639	136.6799	0.6462	0.6936
SAR-SIFT	3.7013	1.6105	0.8698	118.1341	0.7014	0.7229
PSO-SIFT	3.7276	1.6138	0.8714	119.7662	0.7135	0.7269
Proposed	**3.8130**	**1.6271**	**0.8777**	**102.6517**	**0.7458**	**0.7544**

Table 4. Comparison of SIFT, SAR-SIFT, PSO-SIFT and proposed method on real dataset 2.

	MI	NMI	ECC	MSD	PCC	SSIM
SIFT	3.5571	1.5852	0.8572	187.7355	0.6203	0.6051
SAR-SIFT	3.7284	1.6057	0.8674	147.3779	0.7181	0.6564
PSO-SIFT	3.7644	1.5979	0.8640	149.6696	0.6639	0.6667
Proposed	**3.8850**	**1.6217**	**0.8753**	**111.4737**	**0.7586**	**0.7083**

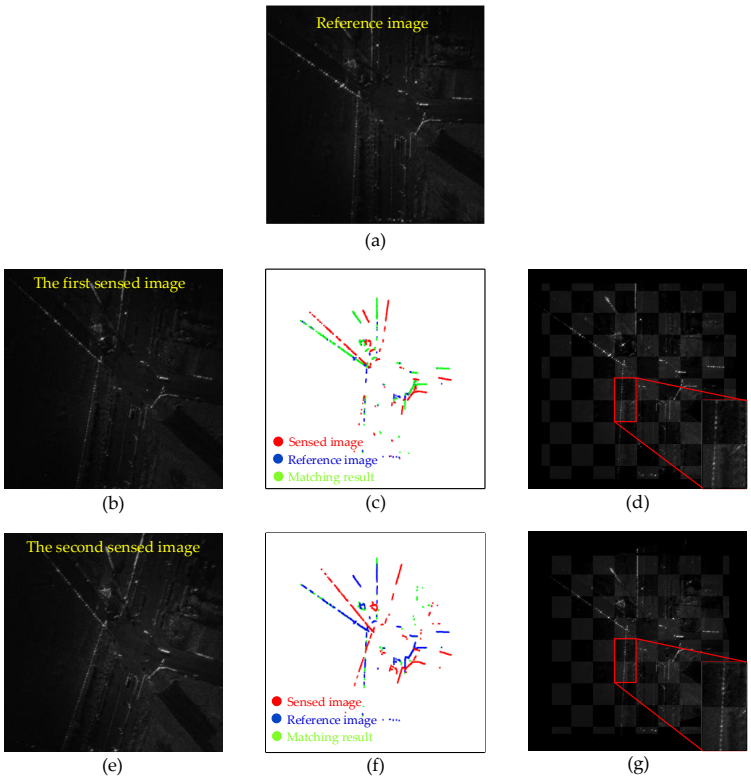

Figure 7. Registration results on real dataset 1. (**a**) Reference image. (**b**) The first sensed image. (**c**) Point cloud matching result of the first sensed image. (**d**) Checkerboard mosaicked image of the first sensed image. (**e**) The second sensed image. (**f**) Point cloud matching result of the second sensed image. (**g**) Checkerboard mosaicked image of the second sensed image.

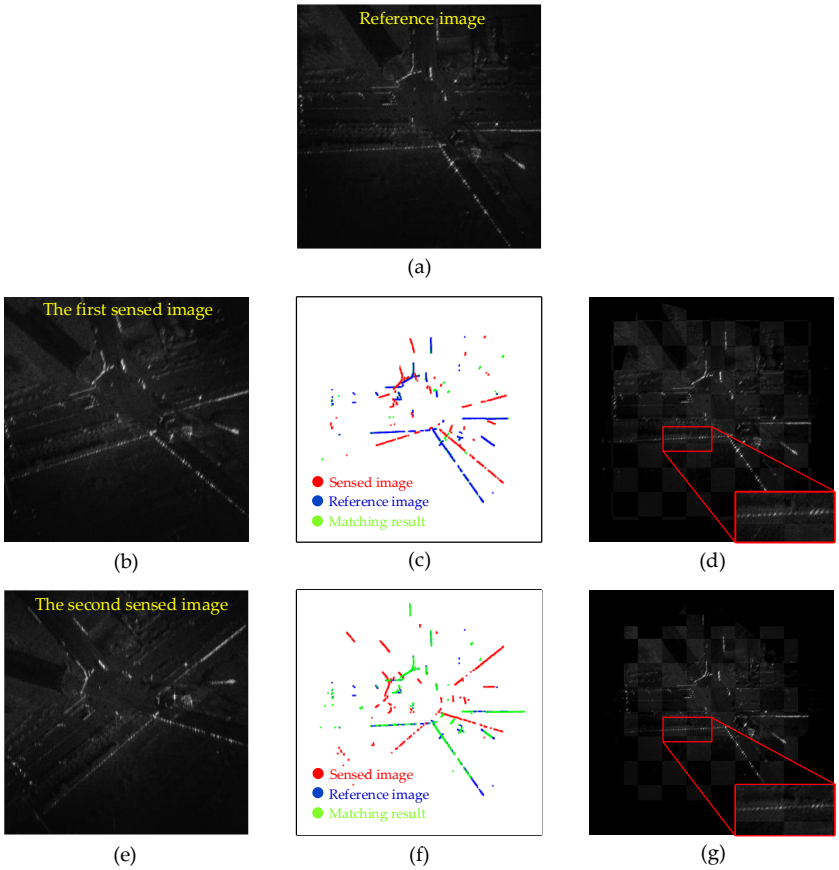

Figure 8. Registration results on real dataset 2. (**a**) Reference image. (**b**) The first sensed image. (**c**) Point cloud matching result of the first sensed image. (**d**) Checkerboard mosaicked image of the first sensed image. (**e**) The second sensed image. (**f**) Point cloud matching result of the second sensed image. (**g**) Checkerboard mosaicked image of the second sensed image.

5.5. Ablation Study

5.5.1. Super-Resolution Performance with Different Numbers of RDBs

In this sub-section, we analyze the effect of different numbers of RDBs on the super-resolution performance of FRSR-Net. With a super-resolution scaling factor of three, we perform experiments on three super-resolution test datasets, and Figure 9 shows the experimental results.

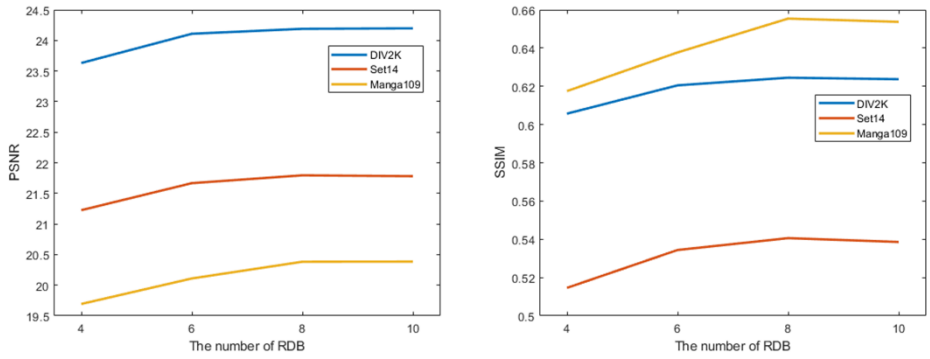

Figure 9. Test results for different numbers of RDBs. (**a**) PSNR metric. (**b**) SSIM metric.

It can be seen from Figure 9 that the PSNR and SSIM metrics increase as the number of RDBs increases, which suggests that increasing the number of RDBs is beneficial to improving the super-resolution performance of the network. The main reason is that more

RDBs indicates a deeper network with stronger feature extraction capacity. When the number of RDBs is greater than eight, we find that PSNR and SSIM increase slightly or even stop at a certain point. However, as the number of RDBs increases, the network is more complex and time-consuming. Therefore, we believe that when the number of RDBs is eight, there is a good balance between network performance and complexity.

5.5.2. Super-Resolution Registration Results with Different Scaling Factors

In this subsection, the effects of different super-resolution scaling factors on the registration performance are discussed. We reselect 30 real video-SAR images to compare the registration performance with different super-resolution scaling factors, and the results are shown in Table 5. From the table, it can be seen that the registration performance of video-SAR images after super-resolution is significantly better than that without super-resolution, indicating that the use of super-resolution is beneficial in improving the registration accuracy of images. This is mainly because the FRSR-Net highlights the edge features of the image and suppresses the speckle noise, which makes the extraction of the structural point cloud easier and more accurate. The best registration performance is achieved when the scaling factor of the super-resolution is three. A possible reason is that as the scaling factor increases, the performance of the super-resolution network becomes unstable.

Table 5. Comparison of the registration with different super-resolution scaling factors.

Scaling Factor	MI	NMI	ECC	MSD	PCC	SSIM
1	4.1025	1.6166	0.8734	134.2960	0.7652	0.7007
2	4.1840	1.6401	0.8834	98.3499	0.8570	0.7504
3	**4.1887**	**1.6413**	**0.8839**	**95.8823**	**0.8611**	**0.7530**
4	4.1779	1.6384	0.8827	101.3225	0.8480	0.7471

6. Conclusions

In this paper, the super-resolution structural point cloud matching (S2-PCM) framework is proposed for video-SAR inter-frame registration, which consists of FRSR-Net, SPCE-Net and RPM-Net. The main conclusions are as follows:

1. Compared with the classical SIFT-like algorithms, S^2-PCM has higher registration accuracy for video-SAR images under diverse evaluation metrics, such as MI, NMI, ECC, SSIM, etc.
2. By integrating feature recurrence structure and RDB, the proposed FRSR-Net can significantly improve the quality of video-SAR images and point cloud extraction accuracy. Combining FRSR-Net with S^2-PCM, we can obtain higher registration accuracy.
3. Increasing the number of RDBs is beneficial in improving the super-resolution performance of FRSR-Net. Experimental results show that when the number of RDBs is eight, an excellent balance between network complexity and performance is achieved.
4. The scaling factor has a significant effect on the results, and a reasonable super-resolution scale should be chosen. Too high a super-resolution scaling factor may lead to the unstable performance of FRSR-Net. Experimental results show that the highest registration accuracy can be obtained when the scaling factor is three.

Our future work will focus on two aspects. The first is to deal with the linear deformation in video-SAR images, which might lead to the RPM-Net failure. Furthermore, we will take on more endeavors on the problem of multi-source point cloud extraction and image registration.

Author Contributions: Conceptualization, Z.X. and J.S.; methodology, Z.X. and J.S.; software, Z.X. and Y.Z.; validation, Z.X. and X.Y.; investigation, W.G.; writing—original draft preparation, Z.X.; writing—review and editing, X.Y. and X.Z.; visualization, Z.X.; supervision, Y.Z. and W.G. funding acquisition, J.S. 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 Grants 61671113 and in part by the Multi-sensor Intelligent Fusion Detection and Recognition Seed Foundation under Grant ZZJJ202103-01.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [https://data.vision.ee.ethz.ch/cvl/DIV2K/], [http://www.manga109.org/en/] and [https://deepai.org/dataset/set14-super-resolution] (all accessed on 16 July 2022).

Acknowledgments: The authors would like to thank the Nanjing Research Institute of Electronics Technology for providing the airborne W-band video-SAR data.

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

References

1. Song, X.; Yu, W. Processing video-SAR data with the fast backprojection method. *IEEE Trans. Aerosp. Electron. Syst.* **2016**, *52*, 2838–2848. [CrossRef]
2. Yang, X.; Shi, J.; Zhou, Y.; Wang, C.; Hu, Y.; Zhang, X.; Wei, S. Ground moving target tracking and refocusing using shadow in video-SAR. *Remote Sens.* **2020**, *12*, 3083. [CrossRef]
3. Jun, S.; Long, M.; Xiaoling, Z. Streaming BP for non-linear motion compensation SAR imaging based on GPU. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 2035–2050. [CrossRef]
4. Chen, F.; Lasaponara, R.; Masini, N. An overview of satellite synthetic aperture radar remote sensing in archaeology: From site detection to monitoring. *J. Cult. Herit.* **2017**, *23*, 5–11. [CrossRef]
5. Zhou, Y.; Shi, J.; Wang, C.; Hu, Y.; Zhou, Z.; Yang, X.; Wei, S. SAR Ground Moving Target Refocusing by Combining mRe^3 Network and TVβ-LSTM. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 1–14.
6. Yang, X.; Shi, J.; Chen, T.; Hu, Y.; Zhou, Y.; Zhang, X.; Wu, J. Fast Multi-Shadow Tracking for Video-SAR Using Triplet Attention Mechanism. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
7. Ding, J.; Wen, L.; Zhong, C.; Loffeld, O. Video SAR moving target indication using deep neural network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7194–7204. [CrossRef]
8. Rui, J.; Wang, C.; Zhang, H.; Jin, F. Multi-Sensor SAR Image Registration Based on Object Shape. *Remote Sens.* **2016**, *8*, 923. [CrossRef]
9. Cui, S.; Xu, M.; Ma, A.; Zhong, Y. Modality-free feature detector and descriptor for multimodal remote sensing image registration. *Remote Sens.* **2020**, *12*, 2937. [CrossRef]
10. Fan, J.; Wu, Y.; Wang, F.; Zhang, P.; Li, M. New point matching algorithm using sparse representation of image patch feature for SAR image registration. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 1498–1510. [CrossRef]
11. Xing, C.; Qiu, P. Intensity-based image registration by nonparametric local smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2081–2092. [CrossRef]
12. Sarvaiya, J.N.; Patnaik, S.; Bombaywala, S. Image Registration by Template Matching Using Normalized Cross-Correlation. In Proceedings of the 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, Bangalore, India, 28–29 December 2009; pp. 819–822.
13. Mahmood, A.; Khan, S. Correlation-coefficient-based fast template matching through partial elimination. *IEEE Trans. Image process.* **2011**, *21*, 2099–2108. [CrossRef]
14. Kern, J.P.; Pattichis, M.S. Robust multispectral image registration using mutual-information models. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1494–1505. [CrossRef]
15. Suri, S.; Reinartz, P. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. *IEEE Trans. Geosci. Remote Sens.* **2009**, *48*, 939–949. [CrossRef]
16. Thévenaz, P.; Unser, M. Optimization of mutual information for multiresolution image registration. *IEEE Trans. Image Process.* **2000**, *9*, 2083–2099. [PubMed]
17. Lowe, D. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *20*, 91–110. [CrossRef]
18. Ke, Y.; Sukthankar, R.; Society, I.C. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, USA, 27 June–2 July 2004; pp. 506–513.
19. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. Sar-sift: A sift-like algorithm for sar images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *53*, 453–466. [CrossRef]
20. Xiang, Y.; Wang, F.; Wan, L.; You, H. An Advanced Rotation Invariant Descriptor for SAR Image Registration. *Remote Sens.* **2017**, *9*, 686. [CrossRef]

21. Fan, B.; Wu, F.; Hu, Z. Aggregating gradient distributions into intensity orders: A novel local image descriptor. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 2377–2384.
22. Wang, S.; Quan, D.; Liang, X.; Ning, M.; Guo, Y.; Jiao, L. A deep learning framework for remote sensing image registration. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 148–164. [CrossRef]
23. 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, Boston, MA, USA, 7–12 June 2015; pp. 3279–3286.
24. Mao, S.; Yang, J.; Gou, S.; Jiao, L.; Xiong, T.; Xiong, L. Multi-Scale Fused SAR Image Registration Based on Deep Forest. *Remote Sens.* **2021**, *13*, 2227. [CrossRef]
25. Besl, P.J.; McKay, N.D. Method for registration of 3-D shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*; SPIE: Bellingham, WA, USA, 1992; pp. 586–606.
26. Yang, J.; Li, H.; Campbell, D.; Jia, Y. Go-ICP: A globally optimal solution to 3D ICP point-set registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2241–2254. [CrossRef]
27. Wang, Y.; Solomon, J.M. Deep closest point: Learning representations for point cloud registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3523–3532.
28. Aoki, Y.; Goforth, H.; Srivatsan, R.A.; Lucey, S. Pointnetlk: Robust & efficient point cloud registration using pointnet. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019; pp. 7163–7172.
29. Sarode, V.; Li, X.; Goforth, H.; Aoki, Y.; Srivatsan, R.A.; Lucey, S.; Choset, H. Pcrnet: Point cloud registration network using pointnet encoding. *arXiv* **2019**, arXiv:1908.07906.
30. Yew, Z.J.; Lee, G.H. Rpm-net: Robust point matching using learned features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11824–11833.
31. Gold, S.; Rangarajan, A.; Lu, C.P.; Pappu, S.; Mjolsness, E. New algorithms for 2D and 3D point matching: Pose estimation and correspondence. *Pattern Recognit.* **1998**, *31*, 1019–1031. [CrossRef]
32. Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Stat.* **1964**, *35*, 876–879. [CrossRef]
33. Kirkpatrick, S.; Gelatt, C.D., Jr.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680. [CrossRef]
34. Papadopoulo, T.; Lourakis, M.I. Estimating the jacobian of the singular value decomposition: Theory and applications. In *European Conference on Computer Vision*; Springer: Berlin, Heidelberg, 2000; pp. 554–570.
35. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
36. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016; pp. 391–407.
37. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1874–1883.
38. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1646–1654.
39. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 June 2017; pp. 136–144.
40. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
41. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
42. Slabaugh, G.G. Computing Euler angles from a rotation matrix. *Retrieved August* **1999**, *6*, 39–63.
43. De Boor, C. Bicubic spline interpolation. *J. Math. Phys.* **1962**, *41*, 212–218. [CrossRef]
44. Shamir, R.R.; Duchin, Y.; Kim, J.; Sapiro, G.; Harel, N. Continuous dice coefficient: A method for evaluating probabilistic segmentations. *arXiv* **2019**, arXiv:1906.11031.
45. Goodman, J.W. Statistical properties of laser speckle patterns. In *Laser Speckle and Related Phenomena*; Springer: Berlin/Heidelberg, Germany, 1975; pp. 9–75.
46. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
47. Penney, G.P.; Weese, J.; Little, J.A.; Desmedt, P.; Hill, D.L. A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE Trans. Med. Imaging* **1998**, *17*, 586–595. [CrossRef] [PubMed]
48. Razlighi, Q.R.; Kehtarnavaz, N.; Yousefi, S. Evaluating similarity measures for brain image registration. *J. Vis. Commun. Image Represent.* **2013**, *24*, 977–987. [CrossRef] [PubMed]

49. Ma, W.; Wen, Z.; Wu, Y.; Jiao, L.; Gong, M.; Zheng, Y. Remote sensing image registration with modified sift and enhanced feature matching. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 3–7. [CrossRef]
50. Wu, Y.; Ma, W.; Gong, M.; Su, L.; Jiao, L. A novel point-matching algorithm based on fast sample consensus for image registration. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 43–47. [CrossRef]
51. 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]