remote sensing

MDPI

Article

A Dual-Generator Translation Network Fusing Texture and Structure Features for SAR and Optical Image Matching

Han Nie [1], Zhitao Fu [1,*], Bo-Hui Tang [1,2], Ziqian Li [1], Sijing Chen [1] and Leiguang Wang [3]

[1] Faculty of Land and Resources Engineering, Kunming University of Science and Technology, Kunming 650031, China; nie_han@stu.kust.edu.cn (H.N.); tangbh@kust.edu.cn (B.-H.T.); liziqian1012@stu.kust.edu.cn (Z.L.); chensijing@stu.kust.edu.cn (S.C.)
[2] State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[3] Institute of Big Data and Artificial Intelligence, Southwest Forestry University, Kunming 650024, China; leiguangwang@swfu.edu.cn
* Correspondence: zhitaofu@kust.edu.cn

Abstract: The matching problem for heterologous remote sensing images can be simplified to the matching problem for pseudo homologous remote sensing images via image translation to improve the matching performance. Among such applications, the translation of synthetic aperture radar (SAR) and optical images is the current focus of research. However, the existing methods for SAR-to-optical translation have two main drawbacks. First, single generators usually sacrifice either structure or texture features to balance the model performance and complexity, which often results in textural or structural distortion; second, due to large nonlinear radiation distortions (NRDs) in SAR images, there are still visual differences between the pseudo-optical images generated by current generative adversarial networks (GANs) and real optical images. Therefore, we propose a dual-generator translation network for fusing structure and texture features. On the one hand, the proposed network has dual generators, a texture generator, and a structure generator, with good cross-coupling to obtain high-accuracy structure and texture features; on the other hand, frequency-domain and spatial-domain loss functions are introduced to reduce the differences between pseudo-optical images and real optical images. Extensive quantitative and qualitative experiments show that our method achieves state-of-the-art performance on publicly available optical and SAR datasets. Our method improves the peak signal-to-noise ratio (PSNR) by 21.0%, the chromatic feature similarity (FSIMc) by 6.9%, and the structural similarity (SSIM) by 161.7% in terms of the average metric values on all test images compared with the next best results. In addition, we present a before-and-after translation comparison experiment to show that our method improves the average keypoint repeatability by approximately 111.7% and the matching accuracy by approximately 5.25%.

Keywords: SAR-to-optical image translation; dual-generator; texture and structure fusing; SAR and optical image matching

Citation: Nie, H.; Fu, Z.; Tang, B.-H.; Li, Z.; Chen, S.; Wang, L. A Dual-Generator Translation Network Fusing Texture and Structure Features for SAR and Optical Image Matching. *Remote Sens.* **2022**, *14*, 2946. https://doi.org/10.3390/rs14122946

Academic Editors: Olga Sykioti, Gangyao Kuang and Xin Su

Received: 10 May 2022
Accepted: 17 June 2022
Published: 20 June 2022

1. Introduction

Different sensors can capture different features. In particular, synthetic aperture radar (SAR) images and optical images are widely used in map production [1]. Optical images conform to human vision, but are susceptible to objective factors such as cloud interference [2], whereas SAR images are immune to the imaging defects of optical images and have the advantages of all-weather acquisition, a long line of sight, and some level of penetration capability. Therefore, the fusion of optical and SAR images is widely used in pattern recognition [3], change detection [4], and landslide recognition [5]. However, a prerequisite for the fusion of SAR and optical images is high-accuracy matching. In recent decades, many feature matching methods for homologous images have been proposed, e.g., SIFT [6], SURF [7], ORB [8], and LoFTR [9]. The LoFTR method mainly focuses on dense

matching of weakly textured regions of homologous images. However, these methods are applicable to homologous image matching, but not to SAR and optical image matching because NRDs are not considered. Recently, to address severe NRDs between SAR and optical images, Cui et al. [10] implemented MAP-Net by introducing spatial pyramid aggregated pooling (SPAP) and an attention mechanism to improve the matching precision of optical and SAR images. Li et al. [11] proposed the radiation-variation insensitive feature transform (RIFT) for different types of images. Cui et al. [12] extended scale invariance based on RIFT, but their method was more sensitive to noise, and Li et al. [13] proposed the locally normalized image feature transform (LNIFT) using a local normalization filter to convert images of different modalities into the same intermediate modality, turning the multimodal image matching problem into a homogenous matching problem, and making different modalities similar to improve the matching accuracy. In recent years, deep learning has been successfully introduced into the field of remote sensing image processing for applications such as image matching [14], image fusion [15], and image translation [16]. It is noteworthy that generative adversarial networks (GANs) can better convert multimodal image matching problems into homologous matching problems. Many researchers have implemented matching between SAR and optical images based on SAR-to-optical translation. Quan, D. [17] proposed a generative matching network (GMN) to generate a corresponding simulated optical image for a real SAR image or a pseudo-SAR image for a single optical image, and then input these matched pairs into a matching network to infer whether they matched, achieving improved performance in SAR–optical image matching. Merkle et al. [18] jointly implemented the translation of single-polarization SAR images into optical images by means of a conditional generative adversarial network (CGAN) and verified the possibility of using the transformed pseudo-optical images for image matching. A k-means clustering-guided generative adversarial network (KCG-GAN) [19] has also been proposed for use in SAR and optical image matching, and the results showed that the quality of SAR-to-optical translation limits the matching accuracy between SAR and optical images. Therefore, the key question that needs to be urgently answered is how to design a high-precision SAR-to-optical translation method to enhance the SAR–optical matching performance.

In recent decades, many researchers have proposed methods, which are mainly based on image enhancement algorithms and pseudo-colour encoding algorithms, for SAR–optical translation. In the field of image enhancement, a wavelet transform-based method was used for SAR image denoising to achieve SAR image enhancement from the perspective of noise suppression, but it was found that there was a possibility of increasing the amount of other types of clutter [20]. By introducing visualization algorithms to map high-dynamic-range SAR amplitude values to low-dynamic-range displays via reflectivity distortion, entropy maximization can be preserved to improve the visual quality of SAR images by maximizing the display information [21], and an adaptive two-scale enhancement method can be used to visualize all greyscale information and enhance local target peaks [22]. However, the previous approaches enhance SAR images by means of visualization methods that cannot effectively resolve differences caused by nonlinear radiation distortion (NRDs). In the field of pseudo-colour coding, the pixels of SAR images are mainly encoded to make them as similar as possible to those of optical images [23–25]; however, a greyscale image is obtained instead of a three-channel image, and because the results are highly dependent on the specifics of the model, the performance may decline in practical use. The images processed by image enhancement algorithms and pseudo colour encoding algorithms are enhanced in terms of visual features, but both types of algorithms ignore the NRDs differences between SAR images and optical images; consequently, large differences in structure and texture remain in the resulting pseudo-optical images compared to the real optical images. For the task of automatic image colorization, a deep learning model can be used to predict the pixel-by-pixel colour histogram suitable for the colouring task without structurally transformed image pairs [26]. In the field of SAR image processing, a convolutional neural network (CNN)-based approach has been used to convert

a single-polarization greyscale SAR image into a full-polarization image [27]. Moreover, generative adversarial networks (GANS) [28] are widely used for image translation. A dialectical GAN using conditional Wasserstein generative adversarial network–gradient penalty (WGAN-GP) loss functions has been applied to translate Sentinel-1 images into TerraSAR-X images [29]. Based on the proposal of a boundary equilibrium generative adversarial network (BEGAN) [30], an adversarial network was designed for SAR image generation, and it was demonstrated that the proposed method could improve the classification accuracy [31]. Many GAN-based methods have also been used in SAR-to-optical transformation, such as Pix2pix [32], CycleGAN [33], S-CycleGAN [34], and EPCGAN [16]. Pix2pix and CycleGAN can both be used for SAR–optical translation, but they have certain drawbacks. With Pix2pix, the structure is vague, and some objects have missing structural information, whereas CycleGAN can retain structural information, but ignores land cover information; accordingly, S-CycleGAN combines the advantages of CycleGAN, preserving both land cover information and structural information. He, W. [35] proposed a model combining residual networks and CGANs that can simulate optical images from multitemporal SAR images. However, there is a major problem with such methods; they usually rely on a network structure designed for optical image transformation, with only simple modifications, which is not applicable for SAR–optical translation because of the differences between the imaging principles of SAR images and optical images. Based on this understanding, a feature-guided method based on a discrete cosine transform (DCT) loss has been proposed [36], and edge information has been used to guide SAR–optical translation to obtain pseudo-optical images with better edge information [37]. Similarly, EPCGAN considers the edge blurring problem for pseudo-optical images, and uses gradient information to preserve the edge information in generated pseudo-optical images. The pseudo-optical images obtained in this way contain better structural information, but a situation may arise in which structure and texture features cannot be effectively fed back, resulting in poor and unrealistic imaging effects due to the inability to achieve deep fusion of the structure and texture features. Inspired by [38], in which more natural image inpainting results were obtained by means of a two-branch network, we also treat SAR-to-optical image translation as consisting of two complementary subtasks, namely, texture translation and structure translation, considering the NRDs of SAR images. We reduce the gap between pseudo-optical and real optical images by introducing a spatial-domain loss function and frequency-domain loss function, and thus, obtain pseudo-optical translation results with high accuracy (see Figure 1).

Figure 1. *Cont.*

Figure 1. High-quality image translation results obtained with our method. The pseudo-optical image is the image generated from the SAR image through our method.

In this paper, we propose a dual-generator translation network that fuses texture and structure features to obtain enhanced pseudo-optical images for SAR–optical matching. The proposed network consists of dual generators, bidirectional gated feature fusion (Bi-GFF) [38], and contextual feature aggregation (CFA) [39] modules and discriminators. First, the input SAR image is decomposed into structure and texture features based on the Canny edge detection algorithm [40]. Then, the structure features and greyscale map are input into the structure encoder, the SAR image is input into the texture encoder, and the feature maps of different dimensions from the texture encoder and structure encoder are stitched together to join the structure and texture decoder to obtain texture features and structure features, which are then fused and refined by the Bi-GFF module and CFA module. Finally, a frequency-domain loss function (focal frequency loss [41]) and a spatial-domain loss function (mean square error) are introduced to reduce the gap between the pseudo-optical and real optical images during the learning process. We present comparative experiments and ablation experiments conducted on the same dataset. The experimental results show that the proposed method yields images with clearer textures and structures that are used to achieve better evaluation results that exhibit better visual properties than the results of Pix2pix [32], CycleGAN [33], S-CycleGAN [34], and EPCGAN [16].

Specifically, the major contributions of this paper are as follows:

1. We propose a dual-generator translation network that fuses texture and structure features to improve the matching of SAR images with optical images. The proposed network includes both structure and texture generators, and the structure and texture features are coupled with each other by these dual generators to obtain high-quality pseudo-optical images.
2. We introduce spatial-domain and frequency-domain loss functions to reduce the gap between pseudo-optical images and real optical images, and present ablation experiments to prove the superiority of our approach.
3. To demonstrate the superiority of the proposed algorithm, we select training and test data from public datasets, and we present keypoint detection and matching experiments for comparisons between pseudo-optical images and real optical images and between real optical images and SAR images before and after translation.

The remainder of this paper is organized as follows. The proposed dual-generator translation network fusing texture and structure features for SAR–optical image translation is introduced in Section 2. We present the experimental results and matching applications in Section 3. A discussion is provided in Section 4. Finally, the conclusions are summarized in Section 5.

2. Methods

In this section, we introduce the proposed dual-generator translation network fusing texture and structure features for SAR–optical image matching. As illustrated in Figure 2, the dual generators provide feedback to each other to obtain the structure and texture features, which are fused by the Bi-GFF and CFA modules. In the following subsections, we present the details of the generators, discriminator, and loss functions.

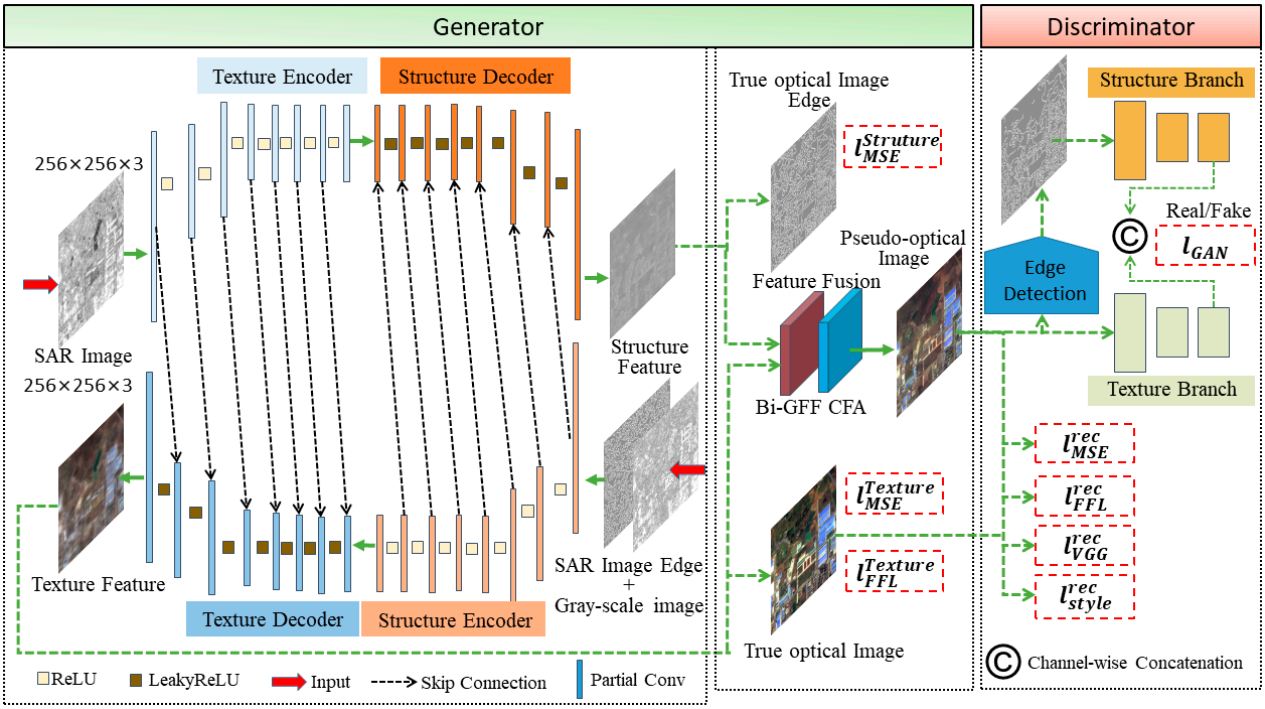

Figure 2. The generators and discriminator of our network. **Generators**: The SAR-to-optical translation process is divided between two generators, i.e., a structure generator and a texture generator, which borrow each other's depth features, and the Bi-GFF and CFA modules are used to refine and fuse the features from these structure and texture reconstruction branches to form the final pseudo-optical image. **Discriminator**: The texture branch guides texture generation, and the structure branch guides structure generation.

2.1. Generators

As shown in Figure 2, the generator part of the SAR-to-optical translation network is divided into two generators, namely, a structure generator and a texture generator, which are based on U-Net variants [42], where final features from the structure encoder and multilevel features from the texture encoder are added to the texture decoder via a skip connection, and final features from the texture encoder and multilevel features from the structure encoder are added to the structure decoder via a skip connection. We also show the structural details of the texture and structure generators in Table 1. In the encoder stage, the SAR image to be translated is passed to the texture encoder, and the greyscale image and edge structure image of the SAR image to be translated are passed to the structure encoder. In the decoder stage, the structure features from the structure encoder are used as constraints in the texture decoder, and the texture features from the texture encoder are used as constraints in the structure decoder. This coupled dual-generator structure ensures good complementarity between the structure and texture features. Compared with normal convolutional layers, partial convolutional layers can better capture the information of irregular boundaries [42]; accordingly, considering the severe scattering noise, NRD, and large irradiance differences between optical and SAR images, we also use partial convolutional layers instead of normal convolutional layers. In addition, we add skip

connections in the CFA module to join together low-level and high-level features during the fusion of the structure and texture features to ensure robust prediction results.

Table 1. Details of the texture and structure generator architecture. PConv is defined as a partial convolutional layer with the specified filter size, stride, and padding. Concat indicates that structure features and texture features are connected by a skip connection.

Module Name	Filter Size	Channel	Stride	Padding	Nonlinearity
Texture/Structure (T/S) Encoder					
T/S Input		3/2			
T/S Encoder PConv1	7×7	64	2	3	ReLU
T/S Encoder PConv2	5×5	128	2	2	ReLU
T/S Encoder PConv3	5×5	256	2	2	ReLU
T/S Encoder PConv4	3×3	512	2	1	ReLU
T/S Encoder PConv5	3×3	512	2	1	ReLU
T/S Encoder PConv6	3×3	512	2	1	ReLU
T/S Encoder PConv7	3×3	512	2	1	ReLU
Texture Decoder					
S Encoder-PConv7		512	-	-	-
Concat (S Encoder-PConv7, T Encoder-PConv6)		512 + 512	-	-	-
T Decoder PConv8	3×3	512	1	1	LeakyReLU
Concat (T Decoder PConv8, T Encoder-PConv5)		512 + 512	-	-	-
T Decoder PConv9	3×3	512	1	1	LeakyReLU
Concat (T Decoder PConv9, T Encoder-PConv4)		512 + 512	-	-	-
T Decoder PConv10	3×3	512	1	1	LeakyReLU
Concat (T Decoder PConv10, T Encoder-PConv3)		512 + 256	-	-	-
T Decoder PConv11	3×3	256	1	1	LeakyReLU
Concat (T Decoder PConv11, T Encoder-PConv2)		256 + 128	-	-	-
T Decoder PConv12	3×3	128	1	1	LeakyReLU
Concat (T Decoder PConv12, T Encoder-PConv1)		128 + 64	-	-	-
T Decoder PConv13	3×3	64	1	1	LeakyReLU
Concat (T Decoder PConv13, T Input)		64 + 3	-	-	-
Texture Feature	3×3	64	1	1	LeakyReLU
Structure Decoder					
T Encoder-PConv7		512	-	-	-
Concat (T Encoder-PConv7, S Encoder-PConv6)		512 + 512	-	-	-
S Decoder PConv14	3×3	512	1	1	LeakyReLU
Concat (S Decoder PConv14, T Encoder-PConv5)		512 + 512	-	-	-
S Decoder PConv15	3×3	512	1	1	LeakyReLU
Concat (S Decoder PConv15, T Encoder-PConv4)		512 + 512	-	-	-
S Decoder PConv16	3×3	512	1	1	LeakyReLU
Concat (S Decoder PConv16, T Encoder-PConv3)		512 + 256	-	-	-
S Decoder PConv17	3×3	256	1	1	LeakyReLU
Concat (S Decoder PConv17, T Encoder-PConv2)		256 + 128	-	-	-
S Decoder PConv18	3×3	128	1	1	LeakyReLU
Concat (S Decoder PConv18, T Encoder-PConv1)		128 + 64	-	-	-
S Decoder PConv19	3×3	64	1	1	LeakyReLU
Concat (S Decoder PConv19, S Input)		64 + 2	-	-	-
Structure Feature	3×3	64	1	1	LeakyReLU

After the texture and structure generators have obtained their respective features, the Bi-GFF module is applied to fuse the structure and texture features to enhance their consistency, and then, the CFA module is applied to further refine the generated pseudo-optical image.

Bidirectional Gated Feature Fusion (Bi-GFF): This module follows the structure and texture generators, and implements information exchange between the structure and texture features, as shown in Figure 3. The texture features are denoted by f_t, the structure features are denoted by f_s, and the features after information exchange can be expressed as:

$$\acute{f}_s = f_s \oplus (W_s(Concat(f_t, f_s)) \otimes f_t) \tag{1}$$

$$\acute{f}_t = f_t \oplus (W_t(Concat(f_t, f_s)) \otimes f_s) \tag{2}$$

where $\oplus$ denotes elementwise addition, $\otimes$ denotes elementwise multiplication, and W_s and W_t denote the convolutional layer mapping functions with a convolutional kernel of 3.

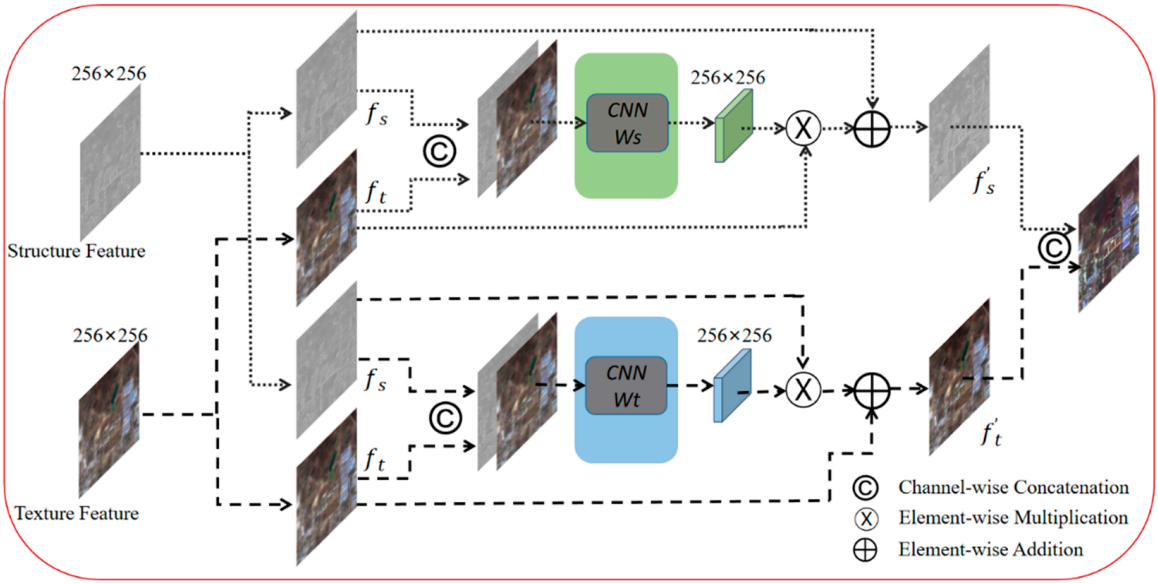

Figure 3. Structural diagram of the Bi-GFF module, in which deep fusion of texture and structure features is performed.

Finally, $\acute{f}_s$ and $\acute{f}_t$ are fused at the channel level to obtain the fused features:

$$f = Concat\left(\acute{f}_s, \acute{f}_t\right) \tag{3}$$

Contextual Feature Aggregation (CFA): As shown in Figure 4, the CFA module is introduced to determine which information in the SAR image contributes to SAR-to-optical translation, thereby enhancing the correlation between image features, and ensuring the overall consistency of the image.

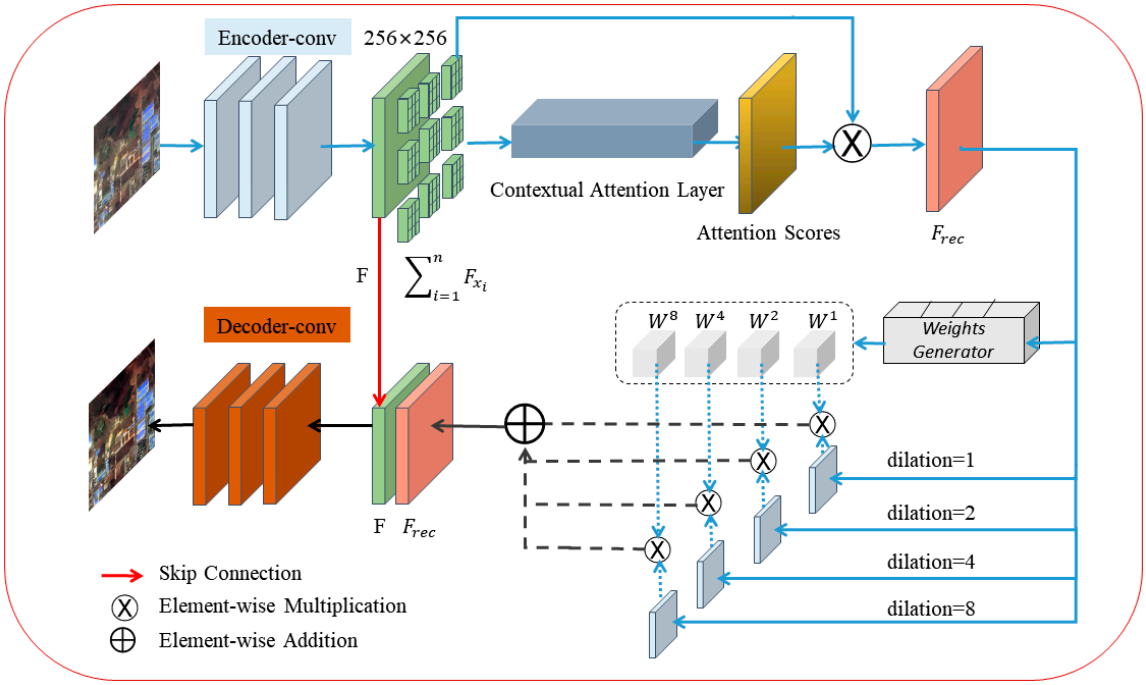

Figure 4. Structural diagram of the CFA module, which effectively models long-term spatial dependence through multiscale information.

First, the feature map F is divided into $\sum_{i=1}^{n} F_{x_i}$ 3×3 patches after being encoded in convolutional layers, and attention scores are obtained by a contextual attention layer, which calculates the cosine similarity between each pair of patches and applies the softmax function to this similarity to obtain the corresponding attention score. Then, the attention scores are multiplied by the 3×3 patches to obtain a reconstructed feature map. The contextual attention layer is defined as:

$$F = EncoderConv(F) \tag{4}$$

$$s_{Attention} = \frac{exp(\left\langle \frac{F_{x_i}}{\|F_{x_i}\|_2}, \frac{F_{x_j}}{\|F_{x_j}\|_2} \right\rangle)}{\sum_{i=1}^{n} exp(\left\langle \frac{F_{x_i}}{\|F_{x_i}\|_2}, \frac{F_{x_j}}{\|F_{x_j}\|_2} \right\rangle)} \tag{5}$$

$$F_{rec} = \sum_{i=1}^{n} F_{x_i} * s_{Attention} \tag{6}$$

where F_{x_i} and F_{x_j} denote the i-th and j-th patches, respectively; $s_{Attention}$ denotes the attention scores; and F_{rec} denotes the reconstructed feature map.

Then, multiscale semantic features are captured from the reconstructed feature map by using different dilation rates:

$$f_{rec}^{k} = Conv_k(F_{rec}) \tag{7}$$

where $Conv_k(\,.\,)$ denotes the k-th dilated convolution layer, $k \in \{1, 2, 4, 8\}$.

A weight generator module is defined to produce pixel-level prediction maps, which are split into four weight modules:

$$W^1, W^2, W^4, W^8 = Slice(W) \tag{8}$$

$$F_{rec} = (F_{rec}^1 \otimes W^1) \oplus (F_{rec}^2 \otimes W^2) \oplus (F_{rec}^4 \otimes W^4) \oplus (F_{rec}^8 \otimes W^8) \tag{9}$$

Finally, we use skip connections to splice F and $f1$ to prevent semantic information from being lost.

$$\acute{F} = DecoderConv(Concat(F_{rec}, F)) \tag{10}$$

2.2. Discriminator

The discriminator [38] distinguishes pseudo-optical images from real optical images by means of two branches: a texture branch and a structure branch. The last layer of the discriminator uses the sigmoid nonlinear activation function, and the structure branch has the same architecture as the texture branch. In the structure branch, the mapping for edge detection is first obtained by a residual network module [43] and a convolutional layer with a kernel size of 1. Then, the structure features are obtained by splicing with greyscale features. In the texture branch, the pseudo-optical image is directly mapped to obtain texture features, and finally, it is stitched to compute the adversarial loss. In addition, we apply spectral normalization [44] in the network to effectively solve the instability problem during network training.

2.3. Loss Functions

The algorithm proposed in this paper includes two generators, a texture generator G_t and a structure generator G_s, and a discriminator to learn the translation from the SAR image domain $\{x_i\}_{i=1}^{n} \in X$ to the optical image domain $\{y_i\}_{i=1}^{n} \in Y$. The original SAR image x_i^{SAR}, the greyscale image x_i^{gray} of the SAR image, and the edge image x_i^{Edge} of the SAR image are passed to the generators to generate the texture features f_t through the texture generator and the structure features f_s through the structure generator. The texture and structure features are then fused by the Bi-GFF module β_{Bi-GFF} and the CFA module σ_{CFA} to obtain the pseudo-optical image Y_{pseudo}. The discriminator D is similarly divided into two branches, i.e., a structure branch D_s and a texture branch D_t. The edge structure image Y_{pseudo}^{Edge} obtained through edge detection convolution is input into the structure branch of the discriminator, the generated pseudo-optical image is input into the texture branch of the discriminator, and finally, the features from the two branches are concatenated in the channel dimension to distinguish a real optical image Y_{real} from a generated pseudo-optical image Y_{pseudo}:

The generator is defined as:

$$Y_{pseudo} = \sigma_{CFA}\left(\beta_{Bi-GFF}\left(\left\{f_t, f_s = \left(G_t\left(x_i^{SAR}\right), G_s\left(x_i^{\text{gray}}, x_i^{\text{Edge}}\right)\right)\right\}\right)\right) \tag{11}$$

The discriminator is defined as:

$$Real/Fake = D\left(\left\{D_t\left(Y_{pseudo}\right),\ D_s\left(Y_{pseudo}^{Edge}, Y_{pseudo}^{Gray}\right)\right\}\right) \tag{12}$$

where $(\ .)$ denotes the projection function implemented by the convolutional layer and $\{\ .\}$ denotes concatenation in the channel dimension.

Reconstruction Loss: We define the reconstruction loss in terms of the differences between a real optical image and the corresponding pseudo-optical image obtained after the Bi-GFF and CFA modules have fused the structure and texture features.

(1) The mean square error (*MSE*) loss function is adopted to reduce the difference in the spatial domain between the pseudo-optical and real optical images at the pixel level. This loss function has the following form:

$$l_{MSE}^{rec} = \mathbb{E}\left[\|Y_{pseudo} - Y_{real}\|_2\right] \tag{13}$$

(2) The focal frequency loss (*FFL*) function is adopted to reduce the difference between the pseudo-optical and real optical images in the frequency domain, and to reduce the artefacts in the pseudo-optical image. The *FFL* function was proposed in [41]. We first use the 2D discrete Fourier transform (DTF) to separately adjust the frequency representations

of the pseudo-optical image and the real optical image, dividing each frequency value by $\sqrt{HW}$ for standard orthogonalization to obtain a smooth gradient, and adjusting the spatial frequency weights of each image by means of a dynamic spectral weight matrix $w(u,v)$. Then, the *FFL* function can be expressed as:

$$F(u,v)^{Y}_{pseudo} = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} f(x,y)\cdot e^{-i2\pi(\frac{ux}{H}+\frac{vy}{W})} \tag{14}$$

$$F(u,v)^{Y}_{real} = \sum_{x1=0}^{H-1}\sum_{y1=0}^{W-1} f(x_1,y_1)\cdot e^{-i2\pi(\frac{ux}{H}+\frac{vy}{W})} \tag{15}$$

$$l^{rec}_{FFL} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} w(u,v)\left|F(u,v)^{Y}_{pseudo} - F(u,v)^{Y}_{real}\right|^2 \tag{16}$$

where $F(u,v)^{Y}_{pseudo}$ denotes a frequency value in the pseudo-optical image, $F(u,v)^{Y}_{real}$ denotes the corresponding frequency value in the real optical image, (u, v) represents the coordinates of a spatial frequency in the frequency spectrum, H × W denotes the size of the image, (x, y) denotes the coordinates of an image pixel in the spatial domain, $f(x, y)$ is the corresponding pixel value, and $w(u, v)$ denotes the dynamic spectral weight matrix.

(3) The VGG loss is used for the perceptual loss of the pseudo-optical and real optical images in terms of high-level semantic information. The pseudo-optical and real optical images are input into the VGG model pretrained on ImageNet [45] to obtain their high-level semantic information. The VGG loss can then be expressed as:

$$l^{rec}_{VGG} = \mathbb{E}\left[\sum_{i}^{3}\|\varphi^{i}_{VGG}\left(Y_{pseudo}\right) - \varphi^{i}_{VGG}(Y_{real})\|_1\right] \tag{17}$$

where $\varphi^{i}_{VGG}(.)$ denotes the projection function of the i-th pooling layer of the pretrained VGG network model.

(4) A style loss is used to ensure that SAR images are translated into pseudo-optical images with the same style as real optical images. The style loss can be expressed as:

$$l^{rec}_{Style} = \mathbb{E}\left[\sum_{i}^{3}\|\mu^{i}_{VGG}\left(Y_{pseudo}\right) - \mu^{i}_{VGG}(Y_{real})\|_1\right] \tag{18}$$

where $\mu^{i}_{VGG}(.) = \varphi^{i}_{VGG}(.)^{T}$, with $\varphi^{i}_{VGG}(.)$ denoting the Gram matrix constructed from the activation map φ^{i}_{VGG}. We choose to use the style loss [46] as demonstrated by Sajjadi et al. [47], based on its effectiveness in eliminating checkerboard artefacts.

Adversarial Loss: We define the adversarial loss in terms of a criterion for similarity evaluation between the pseudo-optical image and the real image.

(1) The GAN loss function is adopted to ensure that the generated pseudo-optical image is as close as possible to a real optical image. The pseudo-optical image and the corresponding real optical image are passed into the structure and texture branches, respectively, of the discriminator to ensure the consistency of the structure and texture. The GAN loss can be expressed as:

$$l_{GAN} = min_G max_D \mathbb{E}\left[\log D\left(Y_{real}, Y^{Edge}_{real}\right)\right] + \mathbb{E}\left[\log 1 - D\left(Y_{pseudo}, Y^{Edge}_{pseudo}\right)\right] \tag{19}$$

Structure Loss: We define the structure loss by comparing the structure features generated by the structure generator with the structure features of the real optical image.

(1) The *MSE* loss function is adopted to ensure that the structure features generated by the structure generator are close to those of a real optical image. The texture *MSE* loss can be expressed as:

$$l^{Struture}_{MSE} = \mathbb{E}\left[\|f_s - Y^{Edge}_{pseudo}\|_2\right] \tag{20}$$

Texture Loss: Distinct from the reconstruction loss, we define the texture loss by comparing the texture features generated by the texture generator with the texture features of the real optical image.

(1) The *MSE* loss function is adopted to ensure that the texture features generated by the texture generator are close to those of a real optical image. The texture *MSE* loss can be expressed as:

$$l_{MSE}^{Texture} = \mathbb{E}\left[\|f_t - Y_{pseudo}\|_2\right] \tag{21}$$

(2) The *FFL* function is adopted to reduce the differences between the texture features generated by the texture generator and those of a real optical image in the frequency domain, and to reduce the artefacts in the texture features:

$$F(u,v)_{f_t}^{Y} = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} f(x,y)\cdot e^{-i2\pi(\frac{ux}{H}+\frac{vy}{W})} \tag{22}$$

$$F(u,v)_{\text{real}}^{Y} = \sum_{x1=0}^{H-1}\sum_{y1=0}^{W-1} f(x_1,y_1)\cdot e^{-i2\pi(\frac{ux}{H}+\frac{vy}{W})} \tag{23}$$

$$l_{FFL}^{Texture} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} w(u,v)\left|F(u,v)_{f_t}^{Y} - F(u,v)_{real}^{Y}\right|^2 \tag{24}$$

where $F(u,v)_{f_t}^{Y}$ denotes the frequency value of the texture features generated by the texture generator, $F(u,v)_{real}^{Y}$ denotes the corresponding frequency value of the real optical image, (u, v) represents the coordinates of the spatial frequency in the frequency spectrum, H × W denotes the size of the image, (x, y) denotes the coordinates of an image pixel in the spatial domain, $f(x, y)$ is the corresponding pixel value, and $w(u, v)$ denotes the dynamic spectral weight matrix.

In summary, the total loss is written as:

$$\begin{aligned} L = {} & \lambda_1(l_{MSE}^{rec}) + \lambda_2(l_{FFL}^{rec}) + \lambda_3 l_{VGG}^{rec} + \lambda_4 l_{Style}^{rec} + \lambda_5 l_{GAN} \\ & + \lambda_6\left(l_{MSE}^{Struture} + l_{MSE}^{Texture}\right) + \lambda_7\left(l_{FFL}^{Texture}\right) \end{aligned} \tag{25}$$

where λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, and λ_7 are weighting coefficients of the loss functions, and the values set for our experiments are $\lambda_1 = 10$, $\lambda_2 = 50$, $\lambda_3 = 0.1$, $\lambda_4 = 250$, $\lambda_5 = 0.1$, $\lambda_6 = 1$, and $\lambda_7 = 5$.

3. Experiments

To demonstrate the effectiveness of our method, comparative experiments with Pix2pix [32], CycleGAN [33], S-CycleGAN [34], and EPCGAN [16] are presented. The results of qualitative visualizations show that our method achieves the best results in terms of both structure and texture. In quantitative experiments, three image quality assessment (IQA) metrics are used, namely, the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM) [48], and the chromatic feature similarity (FSIMc) [49]. A higher PSNR indicates higher image quality, and the SSIM and FSIMc reflect the similarity between the pseudo-optical and real optical images, taking a value of 1 if the two images are identical. Experiments show that our method improves the PSNR by 21.0%, the FSIMc by 6.9%, and the SSIM by 161.7% in terms of the average metric values on all test images compared with the next best results, and the considerable SSIM improvement, in particular, proves the superiority of our dual-generator translation network in producing pseudo-optical images with better structure features.

We also present ablation experiments to demonstrate the effectiveness of the adopted loss functions. The superiority of the loss functions is demonstrated by qualitative visualization results that show the gradual texture and structure enhancement of the pseudo-optical images. In addition, quantitative experimental results show that adding the *MSE* loss function to the method presented in [38] can improve the PSNR by 2.3%, the FSIMc by 1.5%, and the SSIM by 13.9% in terms of the average metric values on all test images, whereas

adding the *FFL* function can similarly improve the PSNR by 4.6%, the FSIMc by 0.7%, and the SSIM by 6.9%) on average, thus proving the superiority of these loss functions.

To verify the improvement in the matching performance, the translation results of our network are applied in keypoint detection and matching experiments, and experiments are presented to compare the performance before and after image translation. These experiments show that our method can improve the overall repeatability of image keypoint detection before and after translation by 111.7% on average for different keypoint detection methods and Euclidean distance thresholds, and the matching performance is also greatly improved. In the following subsections, we give the details of the comparative, ablation, and matching experiments.

3.1. Implementation Details

3.1.1. Datasets

The SEN1-2 datasets [50] contain 282384 optical and SAR images from all parts of the world and all meteorological seasons, and the size of each image block is 256 × 256 pixels. These data can be used for SAR-to-optical translation tasks. Similar to the research in [16], we selected a training set consisting of 2100 images from SEN1-2 that depict many kinds of land cover, such as mountains, forests, lakes, rivers, buildings, farmland, and roads. Figure 5 shows some of the image blocks selected as training samples. At the same time, we selected 222 images as the test set. The test images from SEN-1 and SEN-2 were only used to validate the performance of the proposed method, and were not used during training. The test SAR images were used for SAR-to-optical translation. The optical image blocks in our test set were only used to calculate the image quality of the results of the proposed method for SAR-to-optical translation. Figure 6 shows some of the image blocks selected as test samples. The same dataset was used for retraining in all comparative experiments to effectively evaluate the robustness of the network on the same dataset.

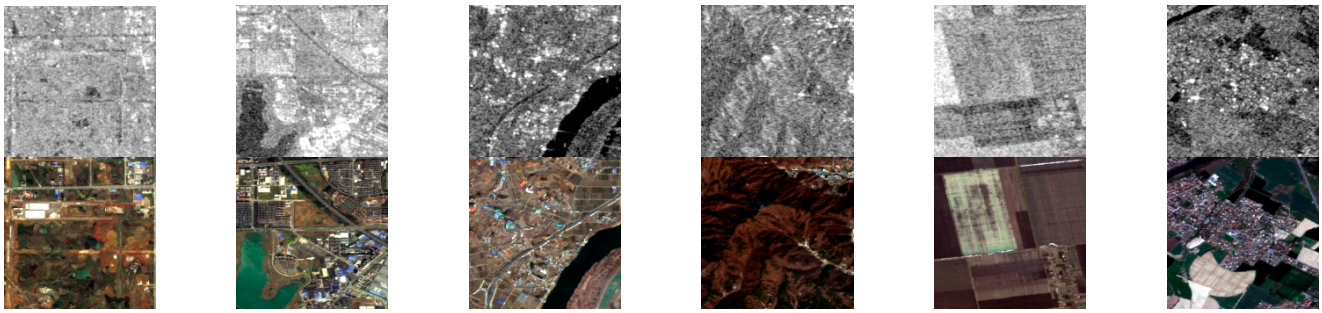

Figure 5. Some examples of training data. First row: SAR image blocks. Second row: optical image blocks.

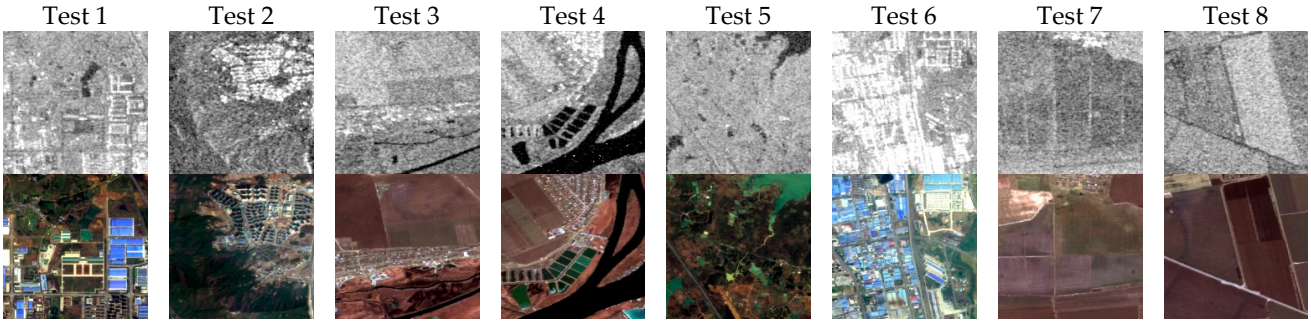

Figure 6. Some of the image blocks were selected as test samples. First row: SAR image blocks. Second row: optical image blocks.

3.1.2. Training Details

Our experiments were based on the deep learning framework PyTorch [51]. The GPU used was an NVIDIA GTX3090, and the experimental operating system environment was Windows 10. During training, the Adam optimization algorithm [52] was used to train the network model. First, a learning rate of 2×10^{-4} was used in the initial training stage; then, once the model stabilized, the learning rate was adjusted to 5×10^{-5} to fine tune the model parameters, and the gradient of the batch normalization layer was frozen at the same time. The generator learning rate was 10 times that of the discriminator learning rate, which was 2×10^{-5}. It took 10 h to train our model on the dataset using a batch size of 1. In the comparative experiments, the Adam optimizer with $\beta 1 = 0.5$ and $\beta 2 = 0.999$ was used for the optimization of the Pix2pix, CycleGAN, S-CycleGAN, and EPCGAN models. Specifically, the generators and discriminators were trained using the Adam optimizer for 200 epochs, and at 100 epochs, the learning rate began to be linearly reduced to 0.

3.2. A Comparison of Textural and Structural Information

To demonstrate the superiority of our proposed method, Figure 7 presents a visual comparison of the different SAR-to-optical translation methods. We can see that, as is evident from Test 2 in Figure 7, Pix2pix fails to preserve the details in the pseudo-optical image, with some shift in style and poor results in terms of both texture and structure. The CycleGAN results are visually superior to those of Pix2pix, containing better texture and structure information, but the details of the structure are blurred and unclear. S-CycleGAN produces images that are relatively clear in texture, but the structure features are distorted. EPCGAN adds gradient information branching on the basis of S-CycleGAN; as a result, its structure information is improved, but the gradient information and the backbone network cannot be fully coupled, and structure or texture features may be sacrificed to some extent to fit the model, which leads to poor results. In contrast, our proposed method yields the best texture and structure information in Tests 1–4.

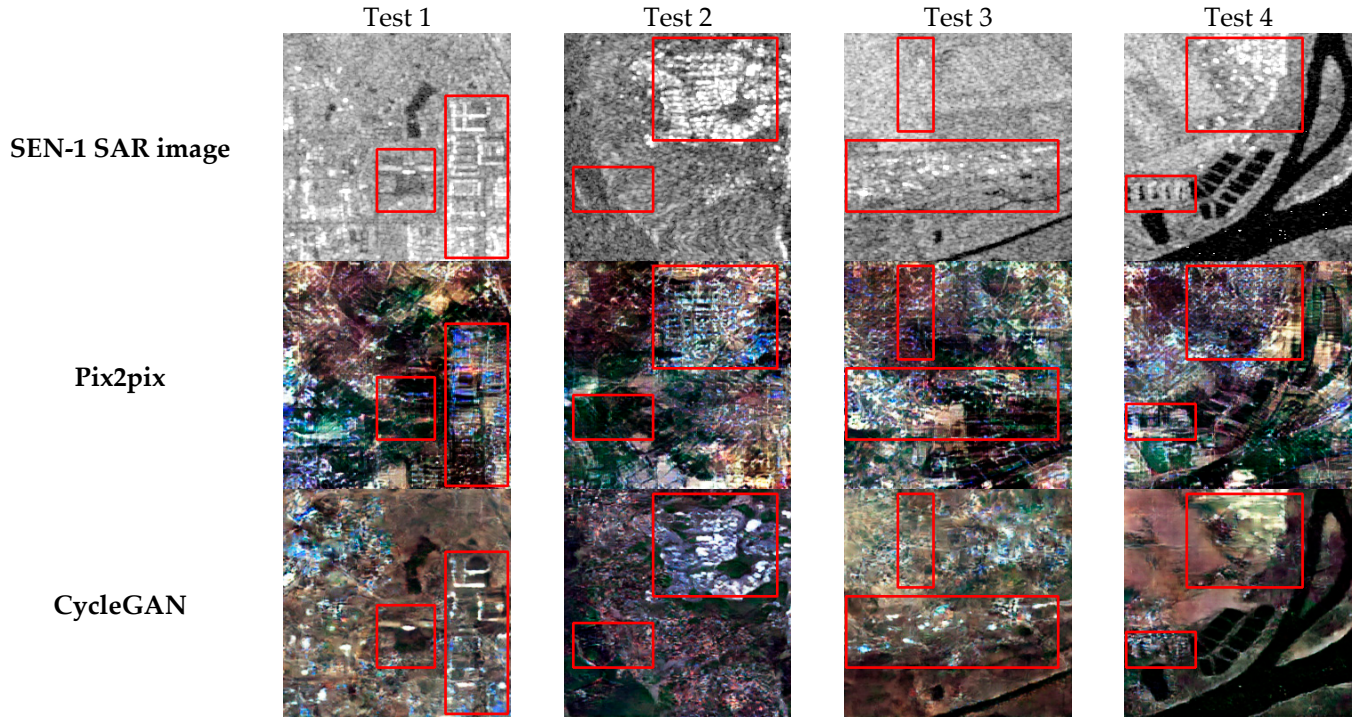

Figure 7. *Cont.*

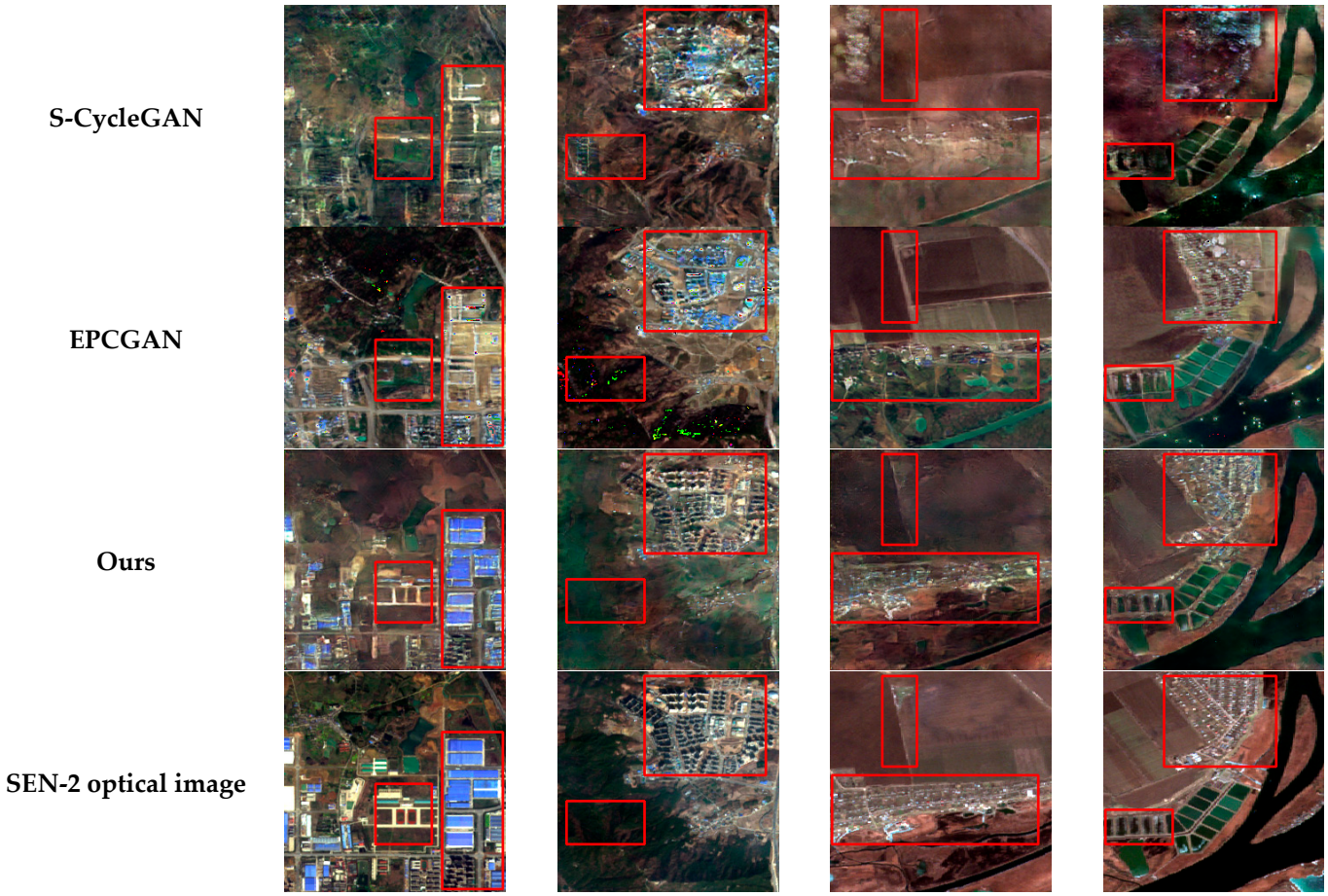

Figure 7. A visual comparison of different SAR-to-optical methods. The size of all images is 256 × 256.

3.3. Results and Analysis

We quantitatively evaluated Pix2pix [32], CycleGAN [33], S-CycleGAN [34], EPC-GAN [16], and our method using the Test 1–4 data to demonstrate the superiority of our method. In addition, we computed the mean values of the PSNR, SSIM, and FSIMc metrics for 222 pairs of test images to further demonstrate the strong robustness of our method. The experimental results show that our method achieves large improvements in Tests 1–4, improving the PSNR by 21.0%, the FSIMc by 6.9%, and the SSIM by 161.7% in terms of the average metric values on all test images compared with the next best results, as shown in Table 2.

The main reasons for these findings are as follows. The performance of Pix2pix is relatively weak because it has less constrained loss conditions, whereas CycleGAN, which is based on the idea of cycles with more strongly constrained SAR-to-optical translation, produces better texture and structure features by virtue of the addition of cyclic loss functions. However, CycleGAN still faces problems with image translation and artefacts. S-CycleGAN improves the CycleGAN network by adding an *MSE* loss function to solve the above problems. However, the NRDs-induced differences between SAR and optical images still lead to poor structural information in the generated pseudo-optical images. To address this shortcoming, EPCGAN incorporates gradient information to guide the SAR-to-optical translation process, thereby improving the structural similarity and visual effect. Compared with the other methods, our proposed dual-generator translation network fuses texture and structure information to achieve considerable enhancement of both the texture and structure features, thus achieving the best performance in these tests.

Table 2. IQA results for different methods obtained by averaging the evaluation metrics over 222 pairs of images in the test set. The best values for each evaluation index are shown in bold.

IQA	DATA	Pix2pix	CycleGAN	S-CycleGAN	EPCGAN	Ours
PSNR	Test 1	11.2090	11.8234	13.6493	13.1767	**19.0867**
	Test 2	12.0022	13.0566	14.3403	15.7064	**20.2105**
	Test 3	13.1578	12.8568	16.5878	16.6274	**20.1606**
	Test 4	11.6502	14.7996	14.1882	16.0283	**20.2383**
	all_test (Average)	11.2212	12.0270	14.4801	14.7253	**17.8228**
FSIMc	Test 1	0.5962	0.6062	0.6262	0.6210	**0.7357**
	Test 2	0.5980	0.6071	0.6712	0.6623	**0.7736**
	Test 3	0.5222	0.5622	0.6651	0.6859	**0.7793**
	Test 4	0.5383	0.7011	0.6942	0.7005	**0.7837**
	all_test (Average)	0.5719	0.6055	0.6699	0.6611	**0.7167**
SSIM	Test 1	0.0711	0.0413	0.0566	0.0746	**0.4574**
	Test 2	0.0825	0.0642	0.0909	0.1318	**0.4586**
	Test 3	0.0567	0.0601	0.2367	0.2326	**0.4911**
	Test 4	0.0616	0.2023	0.1697	0.2042	**0.4263**
	all_test (Average)	0.0528	0.0533	0.1204	0.1264	**0.3308**

3.4. Ablation Experiment

In our proposed method, we incorporated multiple loss functions to improve the translation quality. To demonstrate the superiority of our method, we conducted ablation experiments to demonstrate the effects of the different loss functions, as manifested in the variations of the three evaluation metrics. The gradual improvement in the texture and structure of the pseudo-optical images, as shown in the qualitative visualization results in Figure 8, proves the superiority of adding these loss functions. The quantitative experimental results further show that adding the *MSE* loss function can improve the PSNR by 2.3%, the FSIMc by 1.5%, and the SSIM by 13.9% in terms of the average metric values on all test images, whereas adding the *FFL* function can further improve the PSNR by 4.6%, the FSIMc by 0.7%, and the SSIM by 6.9% on average, as shown in Table 3.

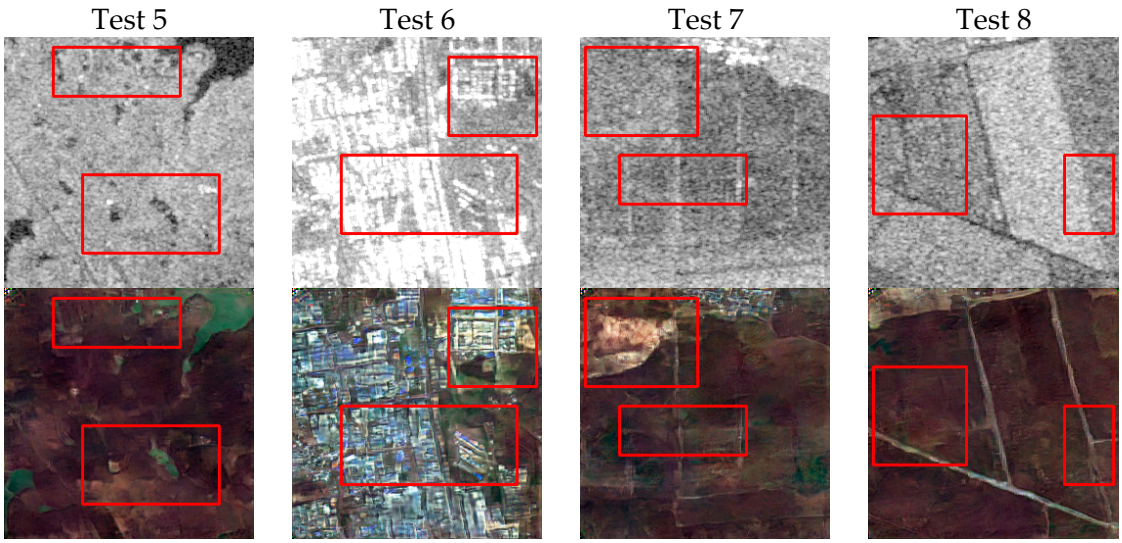

Figure 8. *Cont.*

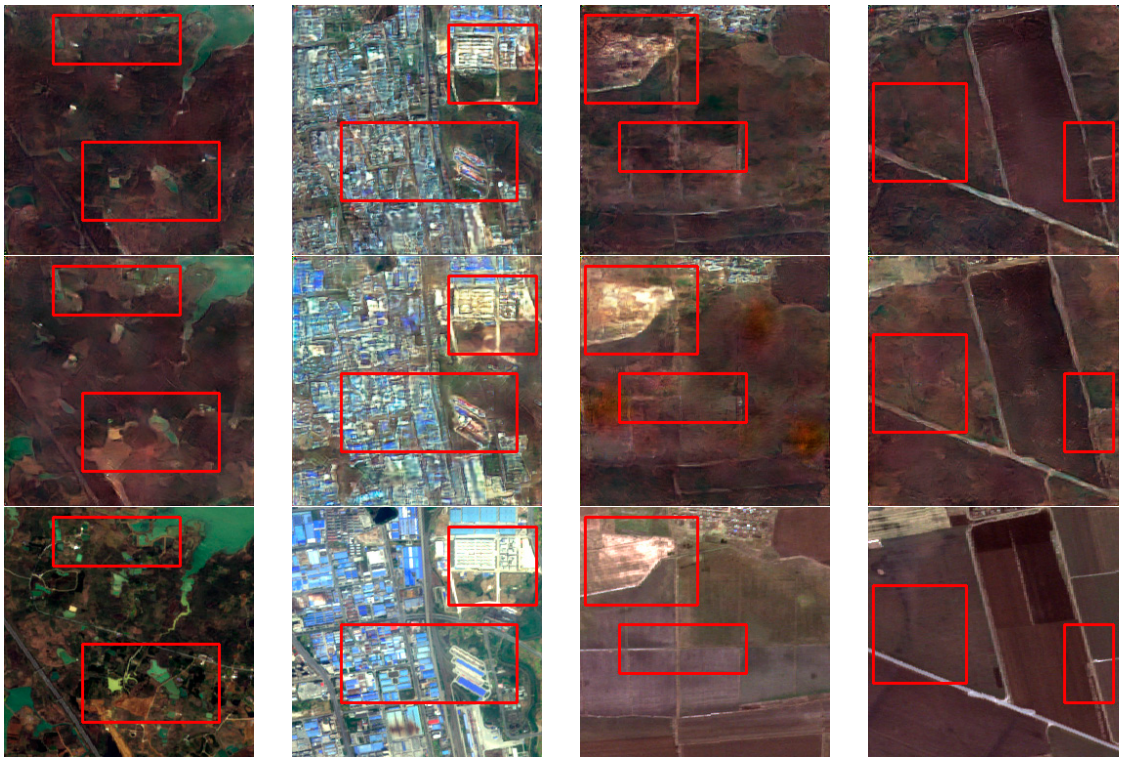

Figure 8. A visual comparison of the results from the ablation experiment. (1) SEN-1 SAR image. (2) Guo [38]. Adapted with permission from Ref. [38]. 2021, Xiefan Guo. (3) Ours (+*MSE* Loss). (4) Ours (+*MSE* loss +*FFL* Loss). (5) SEN-2 optical image.

Table 3. IQA results in the ablation experiment obtained by averaging the evaluation metrics over 222 pairs of images in the test set. The best values for each evaluation index are shown in bold.

IQA	DATA	Guo [38]	Ours (+*MSE* Loss)	Ours (+*MSE* Loss +*FFL* Loss)
PSNR	Test 5	22.6563	22.8161	**23.3868**
	Test 6	13.0547	13.8798	**16.4298**
	Test 7	16.4138	16.5352	**19.4805**
	Test 8	22.5349	22.7283	**24.8288**
	all_test (Average)	16.6327	17.0269	**17.8228**
FSIMc	Test 5	0.7227	0.7255	**0.7271**
	Test 6	0.6789	0.7026	**0.7231**
	Test 7	0.7745	0.7837	**0.8068**
	Test 8	0.7747	0.7734	**0.7824**
	all_test (Average)	0.7007	0.7117	**0.7167**
SSIM	Test 5	0.3237	0.3344	**0.3756**
	Test 6	0.2493	0.3038	**0.3818**
	Test 7	0.3534	0.4079	**0.4459**
	Test 8	0.4897	0.5259	**0.5364**
	all_test (Average)	0.2714	0.3092	**0.3308**

3.5. Matching Applications

To further verify the superiority of our proposed algorithm, the results of comparative experiments on the repeatability of keypoint detection and matching between real optical and generated pseudo-optical (O-PO) images and between real optical and real SAR (O-S) images with different Euclidean distance thresholds and different keypoint detection methods are shown in Table 4. The higher the keypoint repeatability is, the higher the likelihood that the extracted keypoints are correctly matched [53]. From the results, it can be

seen that our method can be applied for keypoint detection well. The smaller the Euclidean distance threshold is, the more obvious the improvement in keypoint repeatability. The average improvements for different keypoint detection methods are 58.4%, 77.6%, 104.1%, 138.6%, and 179.9% for Euclidean distance thresholds of 3.0, 2.5, 2.0, 1.5, and 1.0, respectively. Overall, the average repeatability of image keypoint detection before and after translation is improved by 111.7%, indicating that our translation method can substantially improve the number of potential matching points and further improve the root mean square error (RMSE) of correctly matched point pairs.

Table 4. O-PO: real optical images and generated pseudo-optical images; O-S: real optical images and real SAR images. Experimental comparison of O-PO and O-S image keypoint repeatability under different Euclidean distance thresholds and keypoint detection methods.

Keypoint Repeatability	Keypoint Detection Methods	Optical Keypoint Number	PO/S Keypoint Number	Translation Mode	Euclidean Distance Threshold (L2) 3.0	2.5	2.0	1.5	1.0
Test 1 % Rep.	SuperPoint [54]	603	488/288	O-PO	56.64%	45.55%	30.25%	20.18%	9.93%
				O-S	26.23%	19.25%	11.07%	7.46%	4.09%
	Key.Net [55]	494	593/419	O-PO	40.66%	36.40%	28.07%	22.37%	15.29%
				O-S	19.27%	14.68%	9.20%	5.91%	3.72%
	SIFT [6]	564	600/597	O-PO	51.20%	42.27%	34.36%	23.19%	14.95%
				O-S	39.28%	29.80%	22.22%	13.61%	5.50%
	SURF [7]	530	517/542	O-PO	53.30%	46.03%	36.87%	26.93%	17.38%
				O-S	35.45%	24.63%	16.79%	10.26%	5.41%
	BRISK [56]	613	604/585	O-PO	67.05%	61.13%	49.96%	39.77%	24.65%
				O-S	39.23%	30.55%	21.70%	14.02%	6.51%
	Harris [57]	587	599/576	O-PO	50.08%	38.11%	27.65%	18.21%	9.10%
				O-S	40.24%	30.26%	18.57%	10.83%	4.64%
	PC-Harris [58]	532	458/581	O-PO	53.13%	43.23%	30.30%	19.80%	9.90%
				O-S	40.61%	29.11%	18.86%	9.88%	5.03%
	Hessian	569	563/588	O-PO	55.83%	44.52%	33.92%	22.61%	11.84%
				O-S	41.83%	33.36%	24.20%	12.96%	5.36%

We use the number of correct matches (NCM) and the RMSE to quantify the improvement in matching performance. The experimental results for the O-PO and O-S image matching performance on the images from Tests 1–4 under different matching methods are compared in Table 5. It can be seen that the pseudo-optical images obtained using our method for heterologous source image matching have high application value. In matching experiments using the radiation-variation insensitive feature transform (RIFT), as shown in Figure 9, there are improvements in both the NCM and RMSE. Specifically, the average NCM increases by 137 on the Test 1–4 data, and the average RMSE accuracy improves by 5.25%. The experimental results obtained using the scale-invariant feature transform (SIFT), position scale orientation SIFT (PSO-SIFT), LoFTR [9], and SAR-SIFT also show that our method successfully converts heterologous images that cannot be correctly matched into pseudo homologous images that can be correctly matched, providing a new approach for solving the problem of matching heterologous remote sensing images.

Table 5. O-PO: real optical images and generated pseudo-optical images; O-S: real optical images and real SAR images. Experimental comparison of O-PO and O-S matching performance under different matching methods.

Match Pairs	Translation Mode	LoFTR [9]		RIFT [11]		PSO-SIFT [59]		SAR-SIFT [60]		SIFT [6]	
		NCM	RMSE	NCM	RMSE	NCM	RMSE	NCM	RMSE	NCM	RMSE
Test 1	O-PO	638	0.1113	232	1.6275	75	1.5361	22	0.5712	29	0.6069
	O-S	0	/	86	1.8802	0	/	0	/	0	/
Test 2	O-PO	485	0.1861	328	1.7807	69	1.5554	18	0.5427	26	0.5795
	O-S	0	/	131	1.8214	0	/	0	/	0	/
Test 3	O-PO	335	0.6239	211	1.8689	20	1.3701	7	0.4694	14	0.5907
	O-S	0	/	91	1.9132	0	/	0	/	0	/
Test 4	O-PO	404	0.2441	265	1.8538	57	1.7190	8	0.6512	15	0.5801
	O-S	72	1.1113	180	1.9114	11	1.0988	0	/	8	0.4473

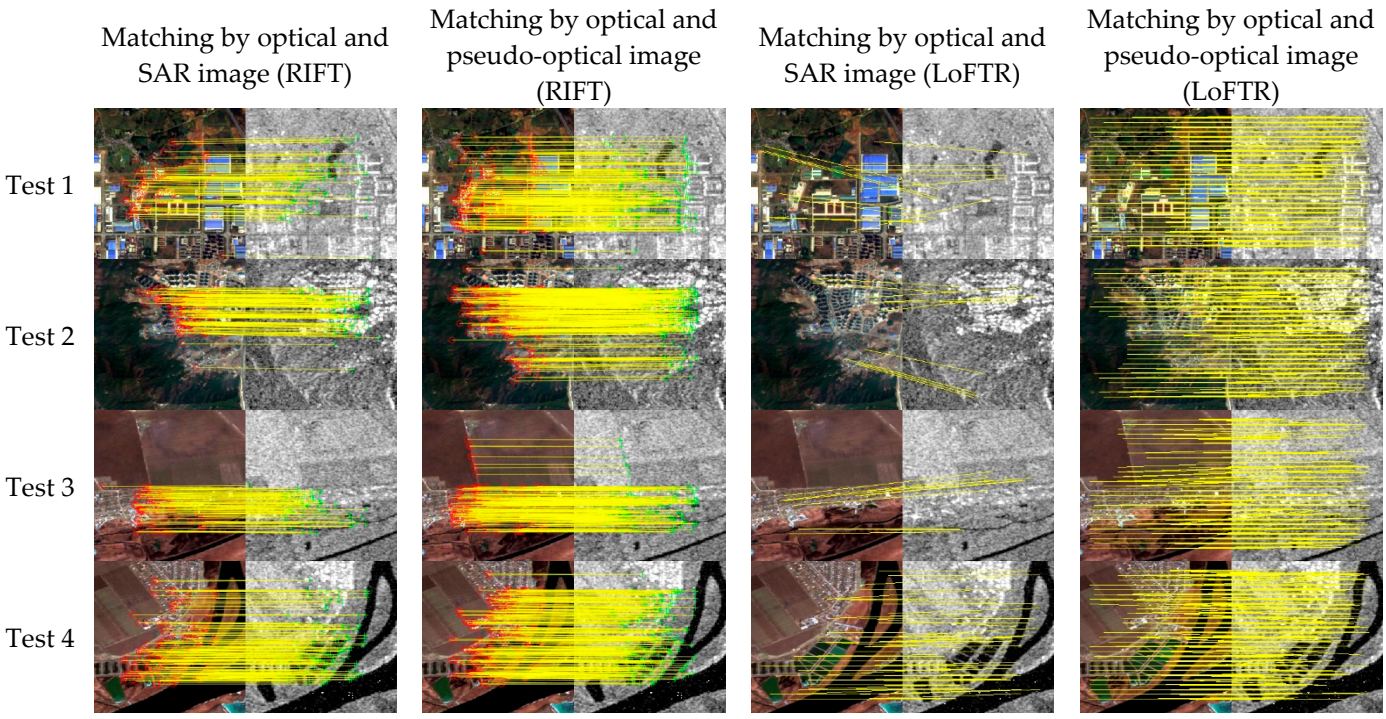

Figure 9. A visual comparison of the results of the matching experiment using RIFT and LoFTR.

4. Discussion

There are large NRDs between optical and SAR images in terms of both structure and texture, which pose a great challenge for optical–SAR image translation. In most image translation tasks, there is a strong connection between the source and target image domains; for example, the structural information may be identical, with the only differences appearing in texture and colour. In contrast, in optical–SAR image translation, both texture and structure information need to be considered. The existing methods all have the problem of favouring either texture features or structure features, and the two cannot influence each other, which can lead to serious structure or texture distortion; furthermore, simply increasing the gradient information of SAR images cannot yield good results. To overcome these challenges, our method considers the dual generation of texture and structure, and fuses the deep structure and texture features thus obtained to produce pseudo-optical images with clear structure and texture information (see Figure 7).

To qualitatively evaluate the structural features of the pseudo-optical images generated by our network, inspired by the research in [61], a fast Fourier transform (FFT)-accelerated

sum of squared differences (SSD) method is used to measure the similarity between the structural features of pseudo-optical images obtained via different methods and those of real optical images. The value of the SSD score plot indicates the offset between image pairs, and a smaller value indicates a higher similarity of their features [62]. In addition, it is noted that the SSD score map obtained with the maximum offset set to 8 pixels has dimensions of 17 × 17. For clarity of observation, we set the maximum offset value to 8 pixels. As shown in Figure 10, the structural features of the pseudo-optical image obtained using our proposed method have the highest similarity with the structural features of the real optical image. Because our network has two generators, one for texture and one for structure, the texture and structure features are coupled and provide feedback to each other to enhance the edge information. Moreover, we add an *MSE* loss function and an *FFL* function to reduce the difference between pseudo-optical and real optical images in both the spatial and frequency domains to achieve greater structural enhancement. Consequently, our method can produce pseudo-optical images with significantly improved structural features.

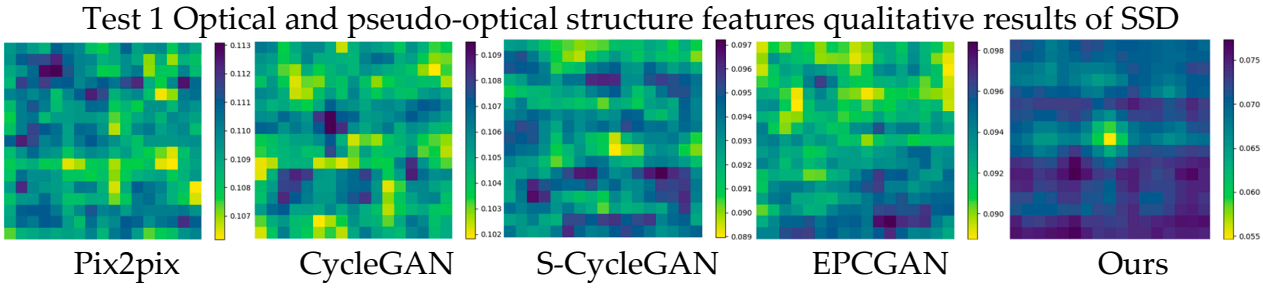

Figure 10. Test 1: qualitative similarity comparison of the structural features of the real optical image and the pseudo-optical images obtained using different SAR-to-optical image translation methods.

The use of our translation results in matching applications can substantially improve the matching performance for SAR and optical images, enhancing the number of potential matching points obtained through keypoint detection, and significantly improving the number of correctly matched point pairs. The practical value of high-precision SAR-to-optical translation for SAR–optical image matching has been explored, and the necessity of high-precision SAR-to-optical translation networks for image matching has been demonstrated. However, in the practical application of SAR images, motion error is a key problem that needs to be solved, and the presence of motion error will lead to unfocused SAR images [63]. Edge features and texture features cannot be obtained, which leads to the poor robustness of our method. In the future, we will work on improving the robustness of the algorithm to apply it in practical applications.

5. Conclusions

In this paper, considering the NRDs in the texture and structure features of optical and SAR images, we summarize the current methods of SAR-to-optical image translation and propose a dual-generator translation network fusing structural and texture features for SAR–optical image matching. Comparative experiments with the latest methods and ablation experiments are conducted to demonstrate that our method achieves superior performance in SAR-to-optical translation. Our method improves the PSNR by 21.0%, the FSIMc by 6.9%, and the SSIM by 161.7% in terms of the average metric values on all test images compared with the next best results. Furthermore, the ablation experiments demonstrate that our introduction of an *MSE* loss function and an FFL function can effectively reduce the spatial- and frequency-domain differences between pseudo-optical images and real optical images and enhance the visual quality of the generated pseudo-optical images, especially in regard to texture, structure, and color information. In addition, to further demonstrate the superiority of our method, comparative experiments of keypoint detection and matching in heterologous remote sensing images before and after translation are presented, and the results prove that the proposed high-precision image translation method can significantly

improve the matching performance for heterologous remote sensing images. Our method improves the average keypoint repeatability by approximately 111.7% and the matching accuracy by approximately 5.25%. In the future, we will strive to further improve the accuracy of our model and enhance its generalization ability.

Author Contributions: Conceptualization, H.N.; methodology, Z.F.; writing—review and editing, Z.L. and L.W.; data curation, S.C. and B.-H.T. 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 nos. 41961053 and 31860182. It was also supported by Yunnan Fundamental Research Projects under grant nos. 202101AT070102, 202101BE070001-037 and 202201AT070164.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The SEN1-2 dataset can be downloaded free of charge from the library of the Technical University of Munich according to the link in [50]. And our code is available at https://github.com/nh945/Translation_Matching.

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

References

1. Kulkarni, S.C.; Rege, P.P. Pixel Level Fusion Techniques for SAR and Optical Images: A Review. *Inf. Fusion* **2020**, *59*, 13–29. [CrossRef]
2. Li, Z.; Zhang, H.; Huang, Y. A Rotation-Invariant Optical and SAR Image Registration Algorithm Based on Deep and Gaussian Features. *Remote Sens.* **2021**, *13*, 2628. [CrossRef]
3. Tapete, D.; Cigna, F. Detection of Archaeological Looting from Space: Methods, Achievements and Challenges. *Remote Sens.* **2019**, *11*, 2389. [CrossRef]
4. Song, S.; Jin, K.; Zuo, B.; Yang, J. A novel change detection method combined with registration for SAR images. *Remote Sens. Lett.* **2019**, *10*, 669–678. [CrossRef]
5. Lacroix, P.; Gavillon, T.; Bouchant, C.; Lavé, J.; Mugnier, J.-L.; Dhungel, S.; Vernier, F. SAR and optical images correlation illuminates post-seismic landslide motion after the Mw 7.8 Gorkha earthquake (Nepal). *Sci. Rep.* **2022**, *12*, 6266. [CrossRef]
6. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
7. Bay, H.; Tuytelaars, T.; Gool, L.V. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
8. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
9. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8922–8931.
10. Cui, S.; Ma, A.; Zhang, L.; Xu, M.; Zhong, Y. MAP-Net: SAR and Optical Image Matching via Image-Based Convolutional Network with Attention Mechanism and Spatial Pyramid Aggregated Pooling. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1000513. [CrossRef]
11. Li, J.; Hu, Q.; Ai, M. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Trans. Image Process.* **2019**, *29*, 3296–3310. [CrossRef]
12. 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]
13. Li, J.; Xu, W.; Shi, P.; Zhang, Y.; Hu, Q. LNIFT: Locally Normalized Image for Rotation Invariant Multimodal Feature Matching. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3165940. [CrossRef]
14. Xiang, Y.; Jiao, N.; Wang, F.; You, H. A Robust Two-Stage Registration Algorithm for Large Optical and SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5218615. [CrossRef]
15. Tang, L.; Yuan, J.; Ma, J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Inf. Fusion* **2022**, *82*, 28–42. [CrossRef]
16. Guo, J.; He, C.; Zhang, M.; Li, Y.; Gao, X.; Song, B. Edge-Preserving Convolutional Generative Adversarial Networks for SAR-to-Optical Image Translation. *Remote Sens.* **2021**, *13*, 3575. [CrossRef]
17. Quan, D.; Wang, S.; Liang, X.; Wang, R.; Fang, S.; Hou, B.; Jiao, L. Deep generative matching network for optical and SAR image registration. In Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 23–17 July 2018; pp. 6215–6218.
18. Merkle, N.; Auer, S.; Muller, R.; Reinartz, P. Exploring the Potential of Conditional Adversarial Networks for Optical and SAR Image Matching. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1811–1820. [CrossRef]

19. Du, W.-L.; Zhou, Y.; Zhao, J.; Tian, X. K-means clustering guided generative adversarial networks for SAR-optical image matching. *IEEE Access* **2020**, *8*, 217554–217572. [CrossRef]
20. Odegard, J.E.; Guo, H.; Lang, M.; Burrus, C.S.; Wells, R.O., Jr.; Novak, L.M.; Hiett, M. Wavelet-based SAR speckle reduction and image compression. In *Algorithms for Synthetic Aperture Radar Imagery II*; SPIE Press: Bellingham, WA, USA, 1995; pp. 259–271.
21. Jiao, Y.; Niu, Y.; Liu, L.; Zhao, G.; Shi, G.; Li, F. Dynamic range reduction of SAR image via global optimum entropy maximization with reflectivity-distortion constraint. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2526–2538. [CrossRef]
22. Zhang, B.; Wang, C.; Zhang, H.; Wu, F. An adaptive two-scale enhancement method to visualize man-made objects in very high resolution SAR images. *Remote Sens. Lett.* **2015**, *6*, 725–734. [CrossRef]
23. Zhou, X.; Zhang, C.; Li, S. A perceptive uniform pseudo-color coding method of SAR images. In Proceedings of the 2006 CIE International Conference on Radar, Shanghai, China, 16–19 October 2006; pp. 1–4.
24. Li, Z.; Liu, J.; Huang, J. Dynamic range compression and pseudo-color presentation based on Retinex for SAR images. In Proceedings of the 2008 International Conference on Computer Science and Software Engineering, Wuhan, China, 12–14 December 2008; pp. 257–260.
25. Deng, Q.; Chen, Y.; Zhang, W.; Yang, J. Colorization for polarimetric SAR image based on scattering mechanisms. In Proceedings of the 2008 Congress on Image and Signal Processing, Sanya, China, 27–30 May 2008; pp. 697–701.
26. Larsson, G.; Maire, M.; Shakhnarovich, G. Learning representations for automatic colorization. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 577–593.
27. Wang, P.; Patel, V.M. Generating high quality visible images from SAR images using CNNs. In Proceedings of the 2018 IEEE Radar Conference (RadarConf18), Oklahoma City, OK, USA, 23–27 April 2018; pp. 0570–0575.
28. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *arXiv* **2014**, arXiv:1406.2661.
29. Ao, D.; Dumitru, C.O.; Schwarz, G.; Datcu, M. Dialectical GAN for SAR image translation: From Sentinel-1 to TerraSAR-X. *Remote Sens.* **2018**, *10*, 1597. [CrossRef]
30. Berthelot, D.; Schumm, T.; Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv* **2017**, arXiv:1703.10717.
31. Marmanis, D.; Yao, W.; Adam, F.; Datcu, M.; Reinartz, P.; Schindler, K.; Wegner, J.D.; Stilla, U. Artificial generation of big data for improving image classification: A generative adversarial network approach on SAR data. *arXiv* **2017**, arXiv:1711.02010.
32. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
33. 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.
34. Wang, L.; Xu, X.; Yu, Y.; Yang, R.; Gui, R.; Xu, Z.; Pu, F. SAR-to-optical image translation using supervised cycle-consistent adversarial networks. *IEEE Access* **2019**, *7*, 129136–129149. [CrossRef]
35. He, W.; Yokoya, N. Multi-temporal sentinel-1 and-2 data fusion for optical image simulation. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 389. [CrossRef]
36. Zhang, J.; Zhou, J.; Lu, X. Feature-guided SAR-to-optical image translation. *IEEE Access* **2020**, *8*, 70925–70937. [CrossRef]
37. Zhang, Q.; Liu, X.; Liu, M.; Zou, X.; Zhu, L.; Ruan, X. Comparative analysis of edge information and polarization on sar-to-optical translation based on conditional generative adversarial networks. *Remote Sens.* **2021**, *13*, 128. [CrossRef]
38. Guo, X.; Yang, H.; Huang, D. Image Inpainting via Conditional Texture and Structure Dual Generation. In Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 14134–14143.
39. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
40. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *6*, 679–698. [CrossRef]
41. Jiang, L.; Dai, B.; Wu, W.; Loy, C.C. Focal frequency loss for image reconstruction and synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13919–13929.
42. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.-C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv* **2018**, arXiv:1802.05957.
45. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
46. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
47. Sajjadi, M.S.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4491–4500.

48. 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]
49. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef]
50. Schmitt, M.; Hughes, L.H.; Zhu, X.X. The SEN1-2 dataset for deep learning in SAR-optical data fusion. *arXiv* **2018**, arXiv:1807.01569. [CrossRef]
51. Collobert, R.; Kavukcuoglu, K.; Farabet, C. Torch7: A matlab-like environment for machine learning. In Proceedings of the BigLearn, NIPS Workshop, Granada, Spain, 12–15 December 2011.
52. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
53. Ye, Y.; Shan, J.; Hao, S.; Bruzzone, L.; Qin, Y. A local phase based invariant feature for remote sensing image matching. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 205–221. [CrossRef]
54. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 224–236.
55. Barroso-Laguna, A.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Key. net: Keypoint detection by handcrafted and learned cnn filters. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5836–5844.
56. Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary robust invariant scalable keypoints. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2548–2555.
57. Harris, C.; Stephens, M. A combined corner and edge detector. In Proceedings of the Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988; pp. 10–5244.
58. Zhang, X.; Hu, Q.; Ai, M.; Ren, X. A Multitemporal UAV Images Registration Approach Using Phase Congruency. In Proceedings of the 2018 26th International Conference on Geoinformatics, Kunming, China, 28–30 June 2018; pp. 1–6.
59. 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.* **2016**, *14*, 3–7. [CrossRef]
60. Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; Tupin, F. SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 453–466. [CrossRef]
61. Zhang, H.; Lei, L.; Ni, W.; Tang, T.; Wu, J.; Xiang, D.; Kuang, G. Optical and SAR image matching using pixelwise deep dense features. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 6000705. [CrossRef]
62. Zhang, H.; Lei, L.; Ni, W.; Tang, T.; Wu, J.; Xiang, D.; Kuang, G. Explore Better Network Framework for High-Resolution Optical and SAR Image Matching. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4704418. [CrossRef]
63. Pu, W. SAE-Net: A Deep Neural Network for SAR Autofocus. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5220714. [CrossRef]