MS-GAN: Learn to Memorize Scene for Unpaired SAR-to-Optical Image Translation

Zhe Guo, Zhibo Zhang, Qinglin Cai, Jiayi Liu, Yangyu Fan, and Shaohui Mei, *Senior Member, IEEE*

***Abstract*—Synthetic aperture radar (SAR) and optical sensing are two important means of Earth observation. SAR-to-optical image translation (S2OIT) can integrate the advantages of both and assist SAR image interpretation under all-day and all-weather conditions. The existing S2OIT methods generally follow a paired training paradigm, which is difficult when dealing with the unpaired S2OIT application scenarios. Moreover, the generator and discriminator in current S2OIT methods have insufficient scene memory for SAR images, resulting in regional landform deformation in the generated images. To address these issues, we propose a novel generative adversarial network capable of memorizing scene for unpaired S2OIT called MS-GAN. The cycle learning framework based on cycle generative adversarial network for unpaired S2OIT is designed to construct the translation mapping between unpaired SAR and optical images. The multiscale representation generator is constructed for multiscale fusion and utilization of scene features of SAR images. The proposed multireceptive field discriminator has the ability to enhance scene memory and generate higher quality optical images in different landforms. In addition, the designed subbands shrinkage denoising module can further suppress the effect of speckle noise in SAR images on the quality of the generated results. Extensive experiments conducted on three challenging datasets SEN1-2, WHU-SEN-City, and QXS-SAROPT demonstrate that the proposed MS-GAN outperforms the state-of-the-art methods on both subjective and objective evaluation metrics.**

***Index Terms*—Synthetic aperture radar (SAR)-to-optical image translation (S2OIT), cycle generative adversarial network (CycleGAN), scene memory, multi-scale fusion, multireceptive field.**

I. Introduction

With the continuous development of space remote sensing detection technology, synthetic aperture radar (SAR) and optical sensing images have a wide range of application needs in land planning, environmental monitoring, resource prospection, military reconnaissance and other fields. SAR is a kind of active remote sensing, which can be used under all-day and all-weather conditions. But SAR images suffer from geometric distortion and speckle noise, which seriously affect the visual effect of SAR imaging [1]. Without prior knowledge, it is difficult for nonexperts to visually identify land cover types from SAR images. In contrast, the optical images contain high spectral resolution, where the visible light range is more in line with human visual perception, and the susceptibility of optical images to severe weather effects such as clouds and fog can be compensated by SAR images [2]. Therefore, combining the advantages of these two images and translate SAR images into corresponding optical images can improve the visual effect of SAR images, and also reduce the cost of interpreting SAR images [3]. The basic idea of the SAR-to-optical image translation (S2OIT) is to pass the SAR images through a deep learning model and learn the mapping relationship from the SAR image domain to the optical image domain based on the features of the SAR and optical images in the training set as a guidance condition [4]. The imaging principles of SAR and optical images are fundamentally different, in terms of measurement methods, wavelengths, detection instruments, and viewing angles [3]. SAR images mainly represent the structure and dielectric properties of the observed target, and the spectral information is insufficient, while optical images are rich in spectral information, which facilitates the visual interpretation of the landscape. However, SAR and optical images still have the essential connection. The acquisition of SAR and optical remote sensing images is a function of the reflection or scattering characteristics of surface environment to electromagnetic waves [3], and although the spectra and colors of SAR and optical images covering the same scene are different, they describe basically the same features, such as spatial locations, shapes, and types of scene, which are unchanged, and these intrinsic common characteristics are the feasibility principle for the S2OIT. Due to the rapid development of natural image-to-image translation (I2IT) based on deep learning [5], [6], researchers have begun to pay attention to S2OIT [4]. However, most of the current research on S2OIT generally follow a paired training paradigm, which is difficult when dealing with the unpaired S2OIT task with more flexible application scenarios.

Cycle generative adversarial network (CycleGAN) [7], as an unsupervised I2IT model for natural images, brings new inspiration for the task of unpaired S2OIT. Currently, most of the research on unpaired S2OIT is designed with CycleGAN as the backbone network, focusing on the unsupervised SAR image translation for image processing tasks, such as cloud and fog removal [8], and change detection [9], etc. Recently, Yang et al. [10] proposed a fine-grained generative adversarial network FG-GAN for the unpaired S2OIT, which enhances the

Manuscript received 23 January 2024; revised 12 March 2024, 11 April 2024, and 14 May 2024; accepted 2 June 2024. Date of publication 10 June 2024; date of current version 18 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62071384 and in part by the Key Research and Development Project of Shaanxi Province under Grant 2023-YBGY-239. *(Corresponding author: Zhe Guo.)*

The authors are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: guozhe@nwpu.edu.cn; billz@mail.nwpu.edu.cn; qlcai@mail.nwpu.edu.cn; liujiayiqqq@mail.nwpu.edu.cn; fan_yangyu@nwpu.edu.cn; meish@nwpu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3411691

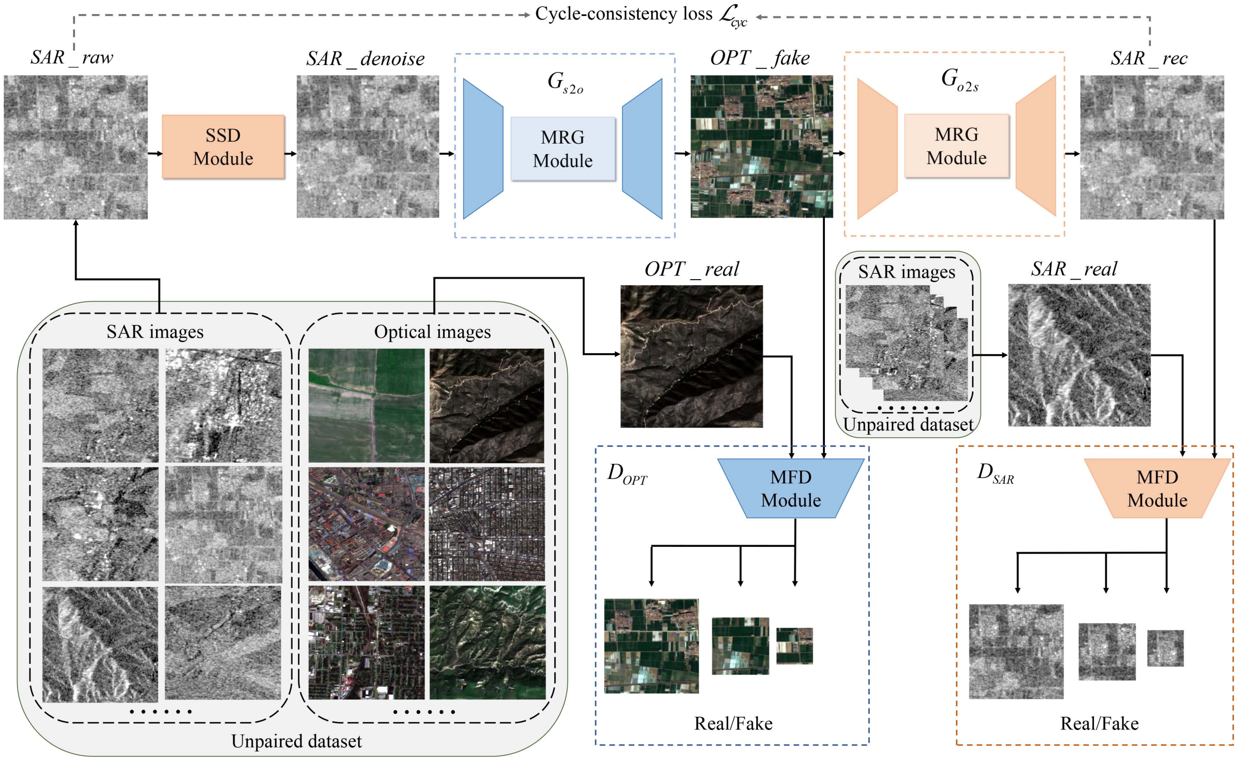

Fig. 1. Overall framework of MS-GAN model. The proposed framework can be divided into two cyclic processes, i.e., S2OIT and O2SIT. This figure only focuses on the S2OIT task of this article.

detailed information in generated optical images by improving the structure of generator and discriminator.

However, the existing unpaired S2OIT methods are designed from the perspective of natural images, without considering complex scene characterization of SAR images. Although several paired S2OIT methods improve the adaptation to SAR images by focusing on wavelet decomposition [11], color information [12], and hierarchical potential features [13], they are not applicable to unpaired S2OIT task due to the limitations of their paired training paradigm. In addition, the generator and discriminator in current S2OIT methods have insufficient feature extraction and discrimination ability for multiscale and multireceptive fields, and it is more difficult to memorize the scene effectively, so the generated images will have regional topographic deformation [14], which is not conducive for scene interpretation of different landforms. Furthermore, the speckle noise in SAR images may also cause color distortion and noisy color patches in the generated images.

In this article, we propose an end-to-end scene memory generative adversarial network (MS-GAN) under the cycle learning framework for unpaired S2OIT. In order to enhance the scene memorization capability for SAR images of different landforms, we redesign the structure of the generator and discriminator. Group convolution is used in the feature extraction part of the generator for multiscale fusion and utilization of the features. In the discriminator part, pyramid atrous convolution block is added to make the output discriminant matrix correspond to different sizes of the receptive field. The constructed multiscale representation generator (MRG) and multireceptive field discriminator enable the network to extract the features and discriminate the image with multiscale and multireceptive fields, which in turn fuses more information to enhance the ability of scene memory and generate higher quality optical images. In addition, the designed subbands shrinkage denoising module can further suppress the effect of speckle noise in SAR images on the quality of the generated results. Fig. 1 shows the framework of our proposed MS-GAN. A series of experiments over three challenging datasets are conducted to assess the effectiveness and rationality of our MS-GAN for unpaired S2OIT.

Overall, our main contributions can be summarized as follows.

1) We propose an end-to-end scene MS-GAN under the cycle learning framework for unpaired S2OIT, which facilitates scene interpretation of different landforms in SAR images. Extensive experimental evaluations on three challenging datasets SEN1-2 [15], WHU-SEN-City [16], and QXS-SAROPT [17] show the effectiveness of the proposed MS-GAN.
2) We design the MRG to enhance the scene memorization capability by multiscale fusion and utilization of scene features of SAR images.
3) We design the multireceptive field discriminator (MFD) to make the output discriminant matrix correspond to different sizes of the receptive field, thus enabling the network to generate higher quality optical images of different scenes.

The rest of this article is organized as follows. Section II briefly introduces the basic knowledge of GAN, natural I2IT, and S2OIT. The proposed MS-GAN model is presented in detail in Section III. Section IV portrays the experiments conducted over three public challenging datasets. Section V discusses the experimental results and our future work. Finally, the conclusion is provided in Section VI.

II. Related Work

A. Generative Adversarial Network

In 2014, Goodfellow et al. [18] proposed GAN, which is a generative model that consists of a generator for generating images and a discriminator for distinguishing the generated images. The excellent performance of the GAN has led to the emergence of a large number of GAN-based studies. Mirza et al. [19] proposed conditional generative adversarial network (CGAN), which uses the condition information corresponding to the image data on the basis of GAN for training, thus can generate images for a specific data distribution. Radford et al. [20] proposed DCGAN, which combines the ideas of convolutional neural networks (CNN) and GAN, thus improving the performance of GAN. Subsequently, researchers have successively proposed a series of improvement methods to further enhance the capability and stability of GAN in the field of image processing, especially the I2IT [21].

B. Natural Image-to-Image Translation

In 2017, Isola et al. [22] proposed Pix2pix based on CGAN, which serves as a general I2IT framework and provides a broad reference for subsequent I2IT work. Meanwhile, Zhu et al. [7] proposed CycleGAN, which consists of two generators and discriminators, and can use unpaired images for I2IT task. Pix2pix and CycleGAN is currently the baseline in supervised and unsupervised I2IT, respectively. Recently, Li et al. [23] presented the Brownian bridge diffusion model (BBDM) to bridge the gap between distinct domains. Jung et al. [24] introduced contrast learning based on CGAN to improve the effectiveness of translated images. Guo et al. [25] proposed structural consistency constraints to mitigate semantic distortions in unpaired image translation. Hu et al. [26] proposed a new semantic relation consistency regularization, and by introducing decoupled contrast learning, image translation achieved better performance. I2IT approaches oriented to multidomain and object recognition [27], [28], and several unsupervised I2IT methods have also been proposed [51], [30]. However, most of the above I2IT methods only focus on the conversion between natural images or the mutual conversion between natural images and sketches.

C. SAR-to-Optical Image Translation

Based on the development of natural I2IT, the work on S2OIT using deep learning techniques started gradually. In 2018, Merkle et al. [31] achieved image alignment for the first time for generated images of remote sensing images, and also attempted to translate SAR images to optical images, showing the great potential of deep learning in the field of remote sensing I2IT. Later, Wang et al. [16] combined the advantages of CycleGAN and Pix2pix to propose a supervised cycle-consistent adversarial network that preserves both landform and structural information in S2OIT. Zhang et al. [32] proposed a feature-guided S2OIT method that combines the cross-modal feature extraction capability of VGG and uses a multilayer feature matching module to retain more information, thus improving the quality of the generated images. Li et al. [11] presented a wavelet feature learning network combined with the CycleGAN framework, which can learn features more efficiently. Wang et al. [13] designed an image translation network with two subnetworks, which learns richer optical features of the input image through the optical reconstruction subnetwork. At the application level, scholars have further improved the quality of generated images by constructing dilated residual block, DCT residual block, and cross-fusion reasoning structure [12], [33], [34], [35]. All of the above methods are supervised image translation, following a paired training paradigm, which is difficult when dealing with the unpaired S2OIT task.

Currently, the researchers begin to study unpaired S2OIT. Yang et al. [10] designed more complex generator and discriminator combined with normalization groups to enable the network to learn richer features in the image, which improves the effect of S2OIT. Du et al. [14] combined two classical methods Pix2pix and CycleGAN and applied them to SAR images to design a semisupervised image translation framework for image matching. Liu et al. [36] achieved unsupervised translation of SAR images by CycleGAN and guided the subsequent change detection task.

The existing studies are either the direct application of the methods for natural I2IT or the design of the network structure only from the view of natural I2IT. Most of these methods do not take into account the complex scene characterization of SAR images. Moreover, the generator and discriminator in current S2OIT methods have insufficient scene memory for SAR images, resulting in regional landform deformation in the generated images.

III. Proposed Method

We propose MS-GAN under the cycle learning framework to cope with the unpaired S2OIT task, which has a strong scene memorization ability, thus learning more reliable information and generating images closer to the real optical images. The MS-GAN consists of three parts: Subbands shrinkage denoising (SSD) module, MRG, and multi-receptive field discriminator (MFD). The overall framework of the model is shown in Fig. 1. Our MS-GAN framework can be divided into two cyclic processes according to the different categories of input images, i.e., S2OIT and optical to SAR image translation (O2SIT) process. Fig. 1 focuses on the task of this article: the S2OIT, in which the SAR image SAR_raw is first suppressed from its speckle noise by the SSD module, and the denoising result $SAR\_denoise$ is then fed into the translation mapping G_{s2o} containing the MRG module to generate the optical image OPT_fake. Then, the discriminator D_{OPT} and another translation mapping G_{o2s} containing the MRG module process the generated optical image. The D_{OPT} makes the generated optical image closer to the real image OPT_real by discriminating the authenticity of the image, and the translation mapping G_{o2s} converts the

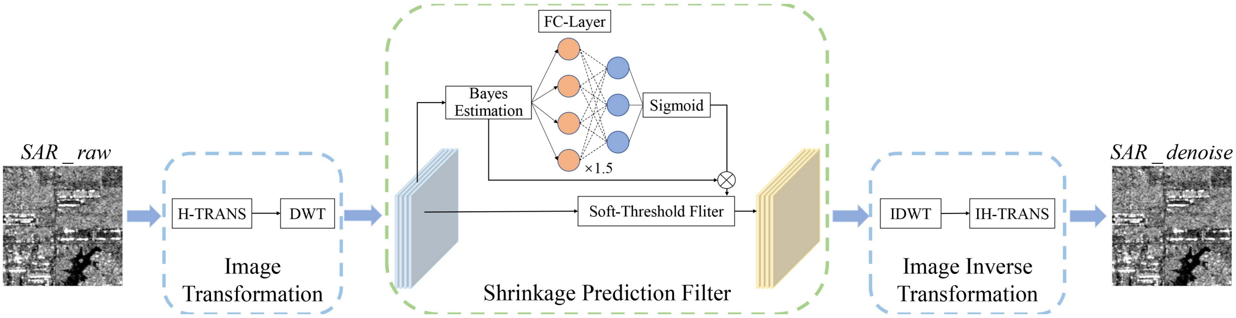

Fig. 2. Structure of SSD module. SAR_raw means the raw SAR image, and $SAR\_denoise$ means the SAR image after denoising.

generated optical image back to the SAR image SAR_rec. The discriminator D_{SAR} makes the generated SAR image SAR_rec closer to the real image SAR_real. The O2SIT process is similar to the S2OIT described above, only the inputs and outputs are different. Cycle-consistency loss L_{cyc} makes the whole network more stable by minimizing the $L1$ loss between input and output images of each cycle process. The redesigned MRG and MFD have the strong ability to memorize scenes in SAR images of different landforms, so as to cope with the complexity and richness of SAR image scenes.

A. Subbands Shrinkage Denoising

The mechanism of SAR coherent imaging leads to the generation of speckle noise, which is an inherent characteristic of SAR images, and this noise will affect the acquisition of valid information in SAR images. To address this problem, we consider the physical model of speckle noise, and design the subbands shrinkage denoising (SSD) module, to suppress the speckle noise in SAR images.

For speckle noise in a SAR imaging system that satisfies the fully developed condition [37], it can be regarded as multiplicative noise. The distribution of the log-transformed speckle can be well-approximated to be additive white Gaussian noise with zero-mean Gaussian distributions [38]. It has been shown in [39] that wavelet shrinkage based on a Bayesian formalism in a homomorphic framework can provide a better reduction of the speckle noise and recover signals from the noisy data more effectively. Based on the above analysis, we design the SSD module to suppress the speckle noise of SAR images. The structure of SSD is shown in Fig. 2.

Let I_{SAR_raw} denote the intensity information of a SAR image, we first perform a homomorphic transform (H-TRANS) on I_{SAR_raw} to transform the multiplicative noise into additive noise, and further perform the discrete wavelet transform (DWT) to obtain four frequency subband images I_{subband}, $subband \in LL, LH, HL, HH$, where LL represent the approximation subband, and LH, HL, HH are three detailed subbands. Then, I_{subband} is input into the subsequent network.

Inspired by SE-Net and deep residual shrinkage network [40], we designed the shrinkage prediction filter (SPF) to obtain the most suitable threshold for each subband image through network learning optimization, and filter each I_{subband} separately based on this threshold to remove the speckle noise in SAR images. Subsequently, the threshold of each subband image can be estimated based on the standard deviations using the Bayesian shrinkage threshold estimation [39], denoted as: $Thr_{Bay}(LL)$, $Thr_{Bay}(LH)$, $Thr_{Bay}(HL)$, and $Thr_{Bay}(HH)$, respectively. Therefore, we get descriptor of each subband image.

These descriptors are then input into a two-layer fully connected (FC) network to continually optimize the threshold estimation, thus accelerating the convergence of the network and obtaining more stable results. The first FC layer sets four nodes, and the activation function is ReLU. The second FC layer sets three nodes, and the sigmoid function is used to normalize the interval [0,1] as the prediction values of the FC layer. As part of the overall network, the parameters of the FC layer are iteratively updated based on the loss values of the generated images. Specifically, in each iteration of the network, the cycle-consistency loss L_{cyc} between SAR_raw and SAR_rec is calculated, which guides the network training by minimizing the $L1$ loss of SAR_raw and SAR_rec, and updating the entire network parameters, including those of the FC layer. In addition, the discriminator D_{OPT} discriminates the generated OPT_fake, calculates the adversarial loss L_{GAN} between OPT_fake and OPT_real, and passes the gradient of this loss value back to the FC layer and updates the parameters accordingly. The input of the FC layer is the descriptors of different subbands, and this operation can combine the information of different frequencies of the image to adjust the filtering threshold of each subband, thus utilizing more information of the image and effectively improving the stability of the network. Then, the descriptors ($Thr_{Bay}(LL)$, $Thr_{Bay}(LH)$, $Thr_{Bay}(HL)$, $Thr_{Bay}(HH)$) are expanded by 1.5 times, respectively [39], and multiplied by the prediction values output from the FC layer to finally obtain the prediction threshold ($Thr_{pre}(LL)$, $Thr_{pre}(LH)$, $Thr_{pre}(HL)$, $Thr_{pre}(HH)$) for the subbands image, specifically for the three detailed subbands [39].

Subsequently, the prediction threshold is input into the filtering unit. Here, we use the soft threshold filtering strategy, where each subband image is filtered according to different thresholds, and this operation can make the images smoother. The soft threshold filtering function is as follows:

$$soft(p_i(x,y)) = \begin{cases} p_i(x,y) + Thr_{pre}() & p_i(x,y) < -Thr_{pre}() \\ 0 & |p_i(x,y)| \leq Thr_{pre}() \\ p_i(x,y) - Thr_{pre}() & p_i(x,y) > Thr_{pre}() \end{cases} \tag{1}$$

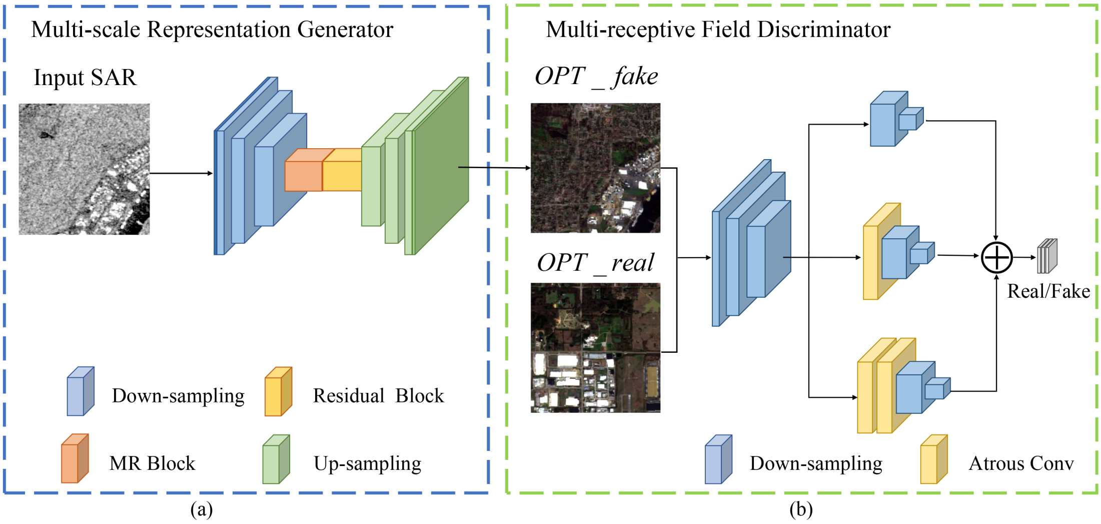

Fig. 3. Structure of the MRG (a) and multireceptive field discriminator (b) in MS-GAN.

where $p_i(x,y)$ describes the value of the ith pixel in I_{subband}, (x,y) denotes the coordinates of the pixel. $Thr_{pre}()$ represents the estimated threshold. After homomorphic and wavelet transforms, negative values are introduced in the resulting subband images, so there are cases where the value of $p_i(x,y)$ is less than zero.

The filtered subband image is first performed an inverse discrete wavelet transform, and then an inverse homomorphic transform, at which point the result is the SAR image after the SSD process.

B. Multiscale Representation Generator

SAR images have a higher viewing angle and a wider field of view than natural optical images. Hence, SAR images contain a greater variety of scene content, and the generator used to generate natural images is difficult to match SAR images. To address this problem, an effective way to enhance the feature extraction capability of the generator is to utilize multiscale information fusion. Currently, many multiscale information extraction methods have been proposed for remote sensing applications, and proved to be effective. Tu et al. [41] proposed a superpixel–pixel–subpixel multilevel network for hyperspectral images (HSIs) classification, which compensates for the insufficiencies of the different levels and decreases the information loss. They designed the global attention module to learn local regular regions based on pixel-level features to extend the global interactive representation capability and reduce the information loss. They further proposed a new multiscale multiangle attention network for HSIs classification that models the internal relationship between image features at local and global scales [42]. Wang et al. [43] presented a multiscale superpixel-guided weighted graph convolutional network for PolSAR images classification. In this method, the multiscale superpixel features are introduced into the weighted graph convolutional network to obtain higher level representation to fully utilize the land cover information in PolSAR images. Wang et al. [44] proposed a multiscale attention superclass CNN to combine multiscale feature fusion with the attention module to improve the integrity of SAR target feature representation. Inspired by the above works, we designed a more complicated generator called MRG, and the MRG is more capable of extracting multiscale features from images and is more suitable for SAR images with complex scenes. The overall structure of the designed MRG is shown in Fig. 3(a).

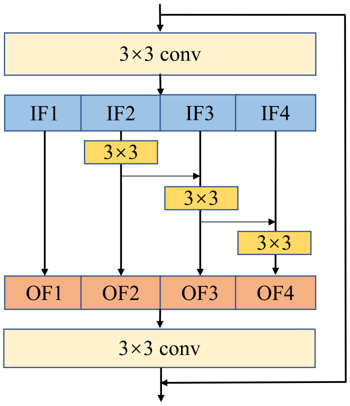

Fig. 4. Structure of the MR Block. (IF: input block features, OF: output block features. The input features are blocked and then convolved for different times to get the output block features, and the obtained output features are spliced and fed into the subsequent network.)

Specifically, we adopt Res2Net [45], but replace the 1 × 1 convolution with the combination of group convolution and regular convolution for extracting the multiscale information of an image. We name this new convolution block the multiscale representation block (MR block), and its structure is shown in Fig. 4. The MR Block extracts the multiscale information of an image by using a combination of group convolution and regular convolution. In the MR Block, a 3 × 3 convolution is first used to extract the features initially to obtain the feature map, and then the feature map is divided into four parts evenly by channel to obtain the block features $IF1 - IF4$ of the input feature map, and different number of convolution operations are

TABLE I
NETWORK ARCHITECTURE OF MRG

Module	Parameters		
	Convolution kernel size	Stride	Channel shift
Conv1	K=(7,7)	1	3→64
Downsampling	K=(3,3)	1	64→128
	K=(3,3)	2	128→256
MR Block×9	K=(3,3)	1	256→256
	-	-	64→64
	K=(3,3)×1	1	64→64
	K=(3,3)×2	1	64→64
	K=(3,3)×3	1	64→64
	K=(3,3)	1	256→256
ResBlock×8	K=(3,3)	1	256→256
Upsampling	K=(3,3)	-	256→128
	K=(3,3)	-	128→64
Conv	K=(7,7)	1	64→3

performed on these block features to obtain the output feature maps $OF1 - OF4$. No convolution operation is performed on the first part of the features, and therefore the output of each part contains information at different scales, i.e., the receptive fields of each part are different. To better fuse the information of different scales, we splice all the obtained results together and go through a 3×3 convolution kernel again. Compared with the simple residual block, MR Block has a larger receptive field and a larger number of parameters.

The new generator is obtained by inserting the MR Block. The input image will first go through the downsampling step, which raises the image dimensions and reduces the size. The downsampling operation performs the initial feature extraction of the image. After this, the image is fed into the 9-layer MR Block to extract the image features from various scales and then fused and output. The MR block ensures that the overall network has a larger field of view and extracts more effective information. The MR block is followed by an 8-layer residual block, which adjusts and filters the image features. Finally, an upsampling layer maps the information of the image to the corresponding image domain and generates a more realistic image. Table I shows the network architecture of our MRG. Compared to the residual block, our network is more complex, a larger number of receptive fields and parameters can ensure the ability of the network to extract features.

C. Multireceptive Field Discriminator

Compared with the traditional generator, the designed MRG has a more complicated network structure, which makes it difficult to perform stable adversarial training with a simple discriminator. To address this issue, we design the MFD combined with atrous convolution. The multireceptive field network enhances the discriminative ability, ensuring stability during adversarial training with the generator, while also guiding the generator to output higher quality images. Several works have shown that atrous convolution avoids reducing the resolution of the feature maps, which helps to mitigate the loss of spatial information [46]. Geometric distortion in SAR images lead to a certain degree of local spatial information loss. The spatial multiscale feature maps obtained by the multireceptive field network with atrous convolution have the potential to reduce the adverse effects of local geometric distortion for SAR image and provide better conditions for generating optical image. Another advantage of atrous convolution is that the number of parameters is smaller while achieving the same receptive field. Atrous convolution requires the input of a convolution dilation rate parameter d, by which the actual size of the atrous convolution kernel is controlled.

TABLE II
NETWORK ARCHITECTURE OF MFD

Module	Parameters		
	Convolution kernel size	Stride	Channel shift
Conv1	K=(4,4)	2	3→64
	K=(4,4)	2	64→128
	K=(3,3)	1	128→256
Atrous Conv	K=(3,3),d=2	1	256→256
	K=(3,3),d=5	1	256→256
Conv×3	K=(4,4)	1	256→256
	K=(4,4)	1	256→1
Contact	-	-	1→3

The structure of our designed MFD is shown in Fig. 3(b). MFD performs pyramid atrous convolution on image features to generate the discriminant matrices, and then these matrices are spliced together. This operation makes it to have three different receptive fields, the size of which are: 70×70, 110×110, and 182×182, which is conducive for the discriminator to combine the details of the generated image with the overall information for discrimination. Table II shows the network architecture of our MFD. The image generated by the generator first goes through the downsampling step to initially get the small receptive field features, then undergoes an atrous convolution layer with d of 2 to get the medium receptive field features, and finally goes through an atrous convolution layer with d of 5 to get the large receptive field features, which will be convolved through three sets of convolutions to get the corresponding discriminant matrices, and finally give the discriminative results.

D. Total Loss Function

The total loss function of our proposed MS-GAN consists of four parts, the adversarial loss function in both S2O and O2S translation directions, the overall cycle-consistency loss function, and the identity loss function.

1) Adversarial Loss: The adversarial loss is necessary for GAN training, and we apply the adversarial loss to two translation mappings. For the translation mapping function $G_{o2s} : OPT \rightarrow SAR$ and its discriminator D_{SAR}, the objective function of the mapping can be expressed as follows:

$$\begin{aligned} &L_{GAN}(G_{o2s}, D_{SAR}, OPT, SAR) \\ &\quad = \mathbb{E}_{sar \sim p_{data(sar)}}[log D_{SAR}(sar)] \\ &\qquad + \mathbb{E}_{sar \sim p_{data(opt)}}[log(1 - D_{SAR}(G_{o2s}(opt)))] \end{aligned} \tag{2}$$

where OPT describes the optical image, SAR denotes the SAR image, G_{o2s} represents the generator that produces the SAR image from the optical, D_{SAR} denotes the discriminator of the SAR image, G_{o2s} tries to translate the optical image to the SAR image, and D_{SAR} is used to distinguish the translation result from the real image. We introduce a similar adversarial loss for the mapping function $G_{s2o} : SAR \rightarrow OPT$, i.e., $L_{GAN}(G_{s2o}, D_{OPT}, SAR, OPT)$

2) Cycle-Consistency Loss: Cycle-consistency loss is an important loss in the unpaired I2IT task, which guides the network training by minimizing the $L1$ loss of the image before and after an image undergoing a transformation cycle $(sar \rightarrow G_{s2o}(sar)) \rightarrow G_{o2s}(G_{s2o}(sar)) \approx sar)$, the cycle-consistency is usually denoted as

$$\begin{aligned} & L_{cyc}(G_{s2o}, G_{o2s}) \\ & \quad = \lambda_{s2o}\mathbb{E}_{sar\sim p_{data}(sar)}[||G_{o2s}(G_{s2o}(sar)) - sar||_1] \\ & \qquad + \lambda_{o2s}\mathbb{E}_{opt\sim p_{data}(opt)}[||G_{s2o}(G_{o2s}(opt)) - opt||_1] \end{aligned} \tag{3}$$

where λ_{s2o} and λ_{o2s} denote the loss weights of the two cycle processes.

3) Identity Loss: The identity loss can stabilize the overall training and help the network learn the correct results [47]. The identity loss function is expressed as

$$\begin{aligned} L_{id}(G_{s2o}, G_{o2s}) = \lambda_{s2o}\mathbb{E}_{sar\sim p_{data}(sar)}[||G_{s2o}(sar) - sar||_1] \\ + \lambda_{o2s}\mathbb{E}_{opt\sim p_{data}(opt)}[||G_{o2s}(opt) - opt||_1]. \end{aligned} \tag{4}$$

4) Total Loss: The total loss function of our proposed MS-GAN is

$$\begin{aligned} & L(G_{s2o}, G_{o2s}, D_{SAR}, D_{OPT}) \\ & \quad = L_{GAN}(G_{o2s}, D_{SAR}, OPT, SAR) \\ & \qquad + L_{GAN}(G_{s2o}, D_{OPT}, SAR, OPT) \\ & \qquad + \lambda_{cyc}L_{cyc}(G_{o2s}, G_{s2o}) + \lambda_{id}L_{id}(G_{o2s}, G_{s2o}) \end{aligned} \tag{5}$$

where λ_{cyc}, λ_{id} represent the weights of cycle-consistency loss and identity loss, respectively. We train the network by optimizing this combined total loss function.

IV. Experiments

A. Experimental Setup

1) Datasets: To validate the effectiveness of our proposed method, we conducted comprehensive comparison experiments with state-of-the-art methods on three public challenging datasets SEN1-2 [15], QXS-SAROPT [17], and WHU-SEN-City [16].

SEN1-2 dataset: The SEN1-2 dataset includes 282 384 pairs of corresponding images, SAR images (VV polarization) collected by Sentinel-1 satellite and optical images collected by Sentinel-2 satellite, each with a size of 256 × 256, which originated from all over the world and four seasons. The SEN1-2 dataset is usually used to train SAR image colorization, SAR image matching, and other tasks.

QXS-SAROPT dataset: The QXS-SAROPT dataset consists of 20 000 pairs of SAR and optical images. The SAR images are from Gaofen-3 satellite, and the corresponding optical images are from Google Earth, covering three port cities: San Diego, Shanghai, and Qingdao. The image size is 256 × 256.

TABLE III
Number of Scene Data per Category of the SEN1-2 Dataset

Category	Train	Test
Urban	431	47
Semiurban	480	53
Rural	528	58
Farmland	493	54
Grassland	487	54
Forest	546	60
Mountain	502	52
Others	137	15

WHU-SEN-City dataset: The WHU-SEN-City dataset contains 32 Chinese cities, with a total of 18 542 paired training samples and 4566 paired test samples, in which the SAR images are collected by the Sentinel-1 satellite, containing two polarization channels, VH and VV, with a resolution of 20 m in the range direction and 22 m in the azimuth direction. The optical data are collected from the Sentinel-2 satellite.

Considering the complexity of the SEN1-2 dataset and the influence of training time and GPU memory on the experimental efficiency, we randomly selected 3600 SAR images and 3600 optical images from the SEN1-2 dataset for unpaired training, and 400 SAR images and 400 optical images for testing. For each scene, the training and test sets were selected uniformly with no crossover. Moreover, the distributions of the training and test data are similar but do not overlap. In addition, we divide these training and test images into urban, semiurban, rural, farmland, grassland, forest, mountain, and other scenes as shown in Table III.

For the QXS-SAROPT dataset and the WHU-SEN-City dataset, we also randomly selected the training and test sets with the same number of SEN1-2 datasets. The QXS-SAROPT dataset contains mainly urban scene and some mountain scene, while the WHU-SEN-City dataset contains mostly urban scene. For both datasets, we extracted the training and test sets according to the proportion of scenes in the entire dataset.

SAR data from the SEN1-2 and QXS-SAROPT dataset only contain amplitude information of the observed targets, and are visually represented as grayscale images, while the SAR data from the WHU-SEN-City dataset contain both amplitude and phase information of the observed targets, and are visually represented as color images. We also preprocessed and randomly enhanced the data before inputting into the network. We first resized the image from 256 × 256 to 286 × 286 by double triple interpolation method, and then randomly flipped it left and right, and finally randomly cropped the image to the same 256 × 256 size as the final input image. The random enhancement of the data can prevent the overfitting phenomenon to a certain extent.

2) Implementation Details: In the experiments, we use the PyTorch framework and a single NVIDIA RTX4090 with 24 GB GPU memory for development, and the server system version is Ubuntu 20.04. We set the same training parameters for the three different datasets, and the Adam optimizer is used. In the training stage, the epoch is set to 200, and the initial learning rate is set to 0.0002 for the first 100 epochs, and decays linearly to zero for the subsequent 100 epochs, and the gradient decay rates are set to 0.5 and 0.999. The translation between the two domains under the unpaired task should be symmetric, so the ratio of the weights of cycle-consistency loss function, λ_{o2s} and λ_{s2o} is 1, i.e., λ_{o2s}:λ_{s2o} = 1:1. The values of hyperparameters λ_{id} and λ_{cyc} are set to 2 and 10. The basis for the selection of hyperparameter values will be analyzed in detail in the ablation study section.

3) Evaluation Metrics: We conduct evaluations using four different mainstream evaluation metrics, which are peak signal-to-noise ratio (PSNR) [48], structural similarity index metric (SSIM) [48], Fréchet inception distance (FID) [49], spectral angle mapping (SAM) [50], and learned perceptual image patch similarity (LPIPS) [51]. PSNR, SSIM, FID, and SAM are applied as objective evaluation metrics for image quality, whereas LPIPS as subjective evaluation metric.

Let x denotes the generated image, and y represents the real image, PSNR is defined as follows:

$$\mathrm{PSNR}(x, y) = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{6}$$

where MAX denotes the maximum gray value of x and MSE denotes the mean squared error between x and y. PSNR evaluates the image quality by estimating the ratio of the useful signal to the background noise, with a larger PSNR representing better image quality.

SSIM is defined as follows:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{7}$$

where μ_x, σ_x^2, μ_y, σ_y^2 denote the mean and variance of the features of x and y, respectively. σ_{xy} is the covariance between x and y. $c1$ and $c2$ are two constants used to ensure that the denominator is not zero. SSIM reflects the image structure similarity between the generated and the real image, with a larger SSIM representing higher similarity.

FID is usually defined as follows:

$$\mathrm{FID}(x, y) = ||\mu_x - \mu_y||_2^2 + Tr(C_x + C_y - 2(C_xC_y)^{\frac{1}{2}}) \tag{8}$$

where μ_x, C_x, μ_y, C_y denote the mean and covariance matrix of the features of x and y, respectively. $Tr(\cdot)$ denotes the trace function. The lower the value of FID represents the more similar the distribution of images.

SAM is defined as follows:

$$\mathrm{SAM}(x, y) = \cos^{-1}\left(\frac{(x^*)^T y^*}{(||x^*||)(||y^*||)}\right) \tag{9}$$

where x^* and y^* denotes the vectorization of x and y, respectively. SAM measure the spectral fidelity between the generated and the real image. The lower the value of SAM indicates that the quality of the generated image is prone to be better.

LPIPS calculates the distance between two images after extracting the network through perceptual features, which can better measure the subjective perception distance between x and y. A lower value of LPIPS represents a better quality of the generated image.

B. Experimental Results

We validate the effectiveness of our MS-GAN for unpaired S2OIT task proposed in this article in terms of both subjective and objective evaluation. We chose five unpaired image translation comparison methods, CycleGAN [7], UGATIT [51], DivCo [30], SCC-CycleGAN [25] (abbreviated as SCC in Figs. 5, 6, 7, 8, and 9), and FG-GAN [10]. We use CycleGAN as the baseline. All the above comparison methods use the code and parameter settings publicly available in the original paper.

1) Metrics Comparison: The comparison results under three datasets, with five evaluation metrics are shown in Table IV, with bold indicating the best. It can be seen that our proposed MS-GAN performs significantly better than the other comparative methods. MS-GAN achieves the best results of the comprehensive evaluation of the metrics under three different datasets, in which the PSNR, SSIM, LPIPS, and SAM metrics are optimal on all three datasets. The above four metrics of our MS-GAN are improved by about 6%, 10%, 6%, and 4.3%, respectively, compared with the other methods. For the FID metric, since it evaluates the images through a network pretrained on the ImageNet dataset, which is mostly images from viewpoints such as cell phone cameras, and quite different from the high-range images of the optical remote sensing images. The two images have different features, so it is not easy to achieve high scores when calculating the distance. In contrast, UGATIT, which focuses on the style transfer task, can achieve better FID results. Combining all the evaluation metrics, our MS-GAN is the best.

2) Visualization Comparison: Figs. 5, 6, 7, 8 show the optical images generated by different methods on three datasets SEN1-2, WHU-SEN-City, and QXS-SAROPT, respectively. In order to intuitively visualize the error distribution of the generated optical images by different methods and real optical images, we also give the residual results of different methods. The blue color indicates the residuals are close to zero, while the yellow color means the residuals are largest, and the color axis from blue to yellow indicates that the residuals vary from small to large. To compare the adaptability of different methods to SAR image scenes, the example images we selected contain a rich variety of scene categories. Figs. 5 and 6 contain images from the SEN1-2 dataset in seven scenes: urban, semiurban, rural, farmland, mountain, grassland, and forest. Since the WHU-SEN-City dataset is basically all urban scene, the images in Fig. 7 are all urbans but contain different landform colors. The QXS-SAROPT dataset is dominated by urban scene and also contains some mountains, so Fig. 8 contains both urban and mountain scenes. By comparison, we find that the DivCo and UGATIT methods lack the ability to memorize the complex and variable SAR scenes. DivCo outputs the same color for different landforms, while UGATIT easily outputs colors that do not match the landform structure. In Figs. 5 and 6, DivCo obtains dark red-based colors for all example images, not getting the correct color information. UGATIT's color information is sometimes right and sometimes wrong, and the results generated by these two methods are much more difficult for interpretation. The residual results of these two methods also demonstrate a large error distribution with respect to real optical images.

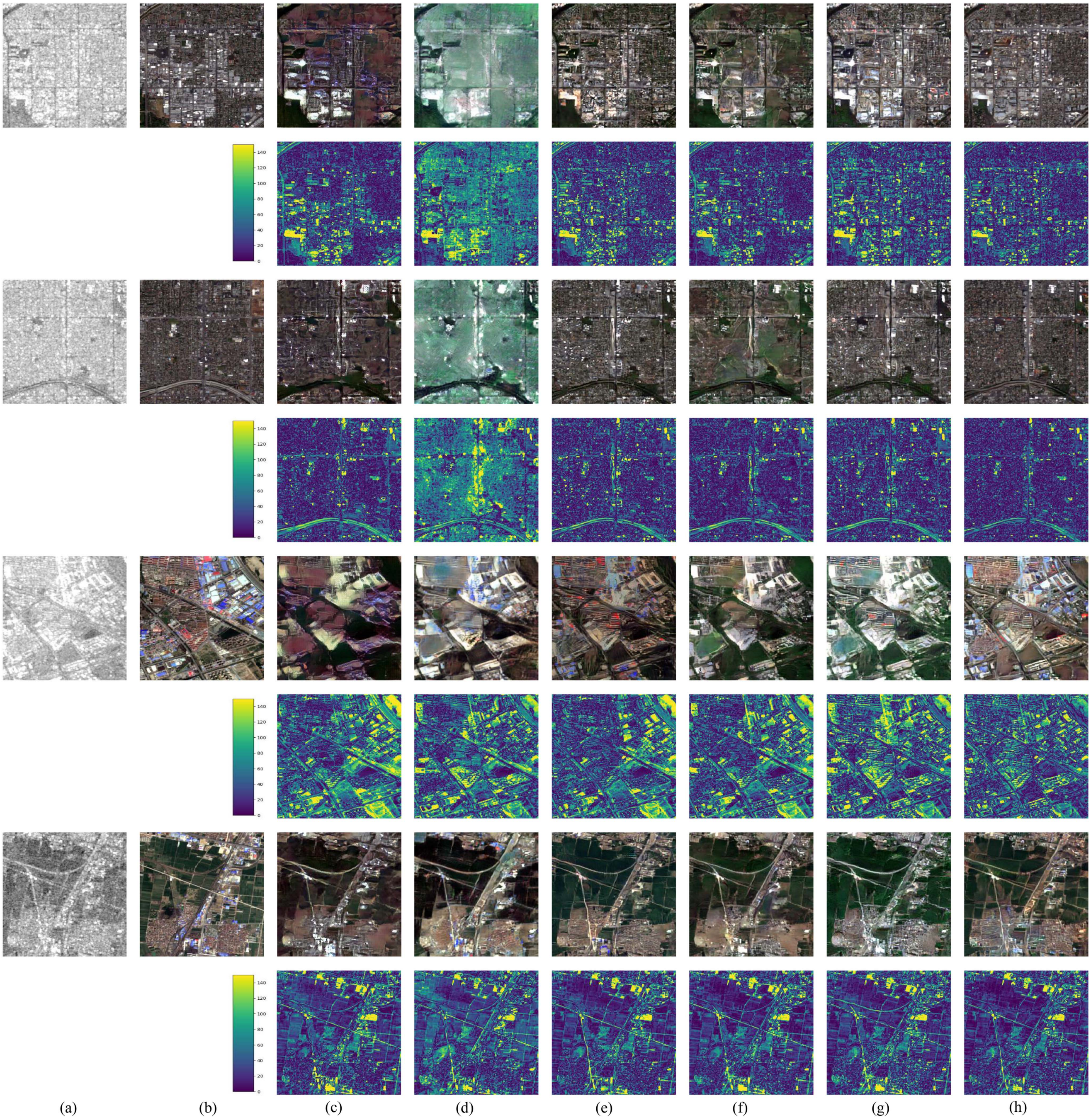

Fig. 5. Visualization results of different methods for S2OIT under different scenes on the SEN1-2 dataset. (a) Original SAR image. (b) Real optical image. (c)–(h) Represent the result of DivCo, UGATIT, CycleGAN, FG-GAN, SCC, and our MS-GAN, respectively. The residual results are under the generated optical images of each method. The scenes are urban, urban, semiurban, rural, from top to bottom.

TABLE IV
EVALUATION INDEX RESULTS ON DIFFERENT DATASETS

	SEN1-2					WHU-SEN-City					QXS-SAROPT				
	PSNR↑	SSIM↑	FID↓	LPIPS↓	SAM↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	SAM↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	SAM↓
CycleGAN *(baseline)* [7]	12.42	0.2032	132.0	0.5018	11.48	9.55	0.1351	126.3	0.4910	10.61	13.79	0.2908	167.0	0.5978	6.28
UGATIT [51] *(ICLR2020)*	12.29	0.2083	135.6	0.5301	11.40	9.77	0.1408	**90.4**	0.4984	10.18	13.42	0.2619	**128.6**	0.6211	6.16
DivCo [30] *(CVPR2021)*	12.16	0.1798	204.8	0.5581	13.99	9.01	0.1344	144.6	0.5034	10.57	13.31	0.2775	220.0	0.6107	6.03
SCC-CycleGAN [25] *(CVPR2022)*	12.20	0.1961	155.3	0.5174	11.95	9.10	0.1224	116.7	0.5110	10.92	13.21	0.2243	132.9	0.6415	6.36
FG-GAN [10] *(TGRS2022)*	12.33	0.1981	163.7	0.5276	10.62	9.84	0.1426	134.4	0.4829	10.14	13.10	0.3000	313.8	0.6470	6.20
MS-GAN *(ours)*	**13.09**	**0.2211**	**128.7**	**0.4951**	**10.46**	**9.91**	**0.1435**	120.1	**0.4821**	**10.09**	**14.62**	**0.3038**	162.2	**0.5727**	**5.44**

↑ means bigger is better, ↓ means smaller is better.

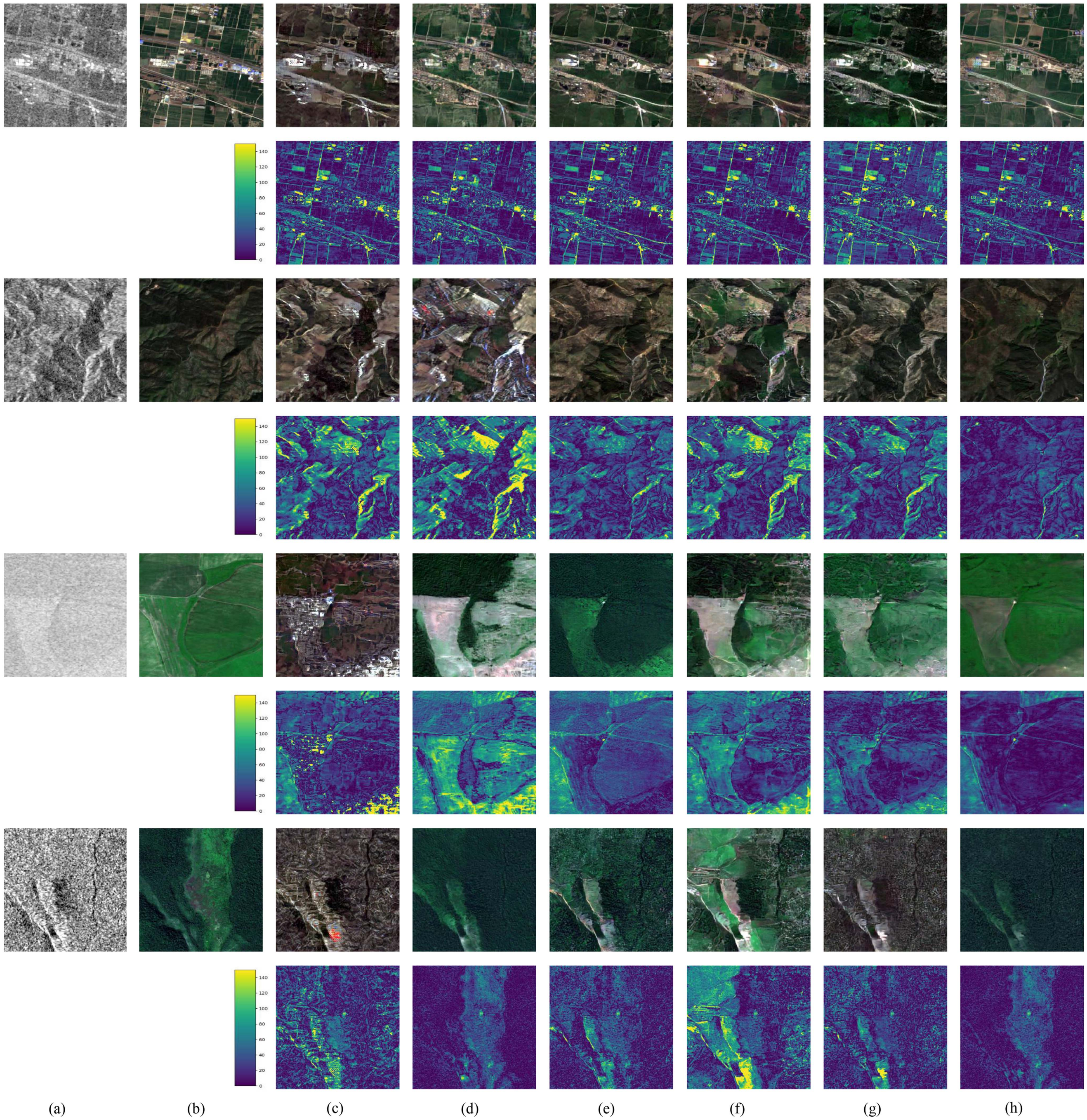

Fig. 6. Visualization results of different methods for S2OIT under different scenes on the SEN1-2 dataset. (continued) (a) Original SAR image. (b) Real optical image. (c)–(h) Represent the result of DivCo, UGATIT, CycleGAN, FG-GAN, SCC, and our MS-GAN, respectively. The residual results are under the generated optical images of each method. The scenes are farmland, mountain, grassland, and forest, from top to bottom.

CycleGAN and the CycleGAN-based methods FG-GAN and SCC-CycleGAN all have preliminary scene memory ability, and can output colors that are relatively consistent with the landform structure, but will output similar colors for the forest and grassland, and the color difference between the generated image and the real optical image is still relatively obvious, as the residual results show.

Meanwhile, the optical images of urban scene generated by the above method will contain many wrong color information, and the results are still not accurate enough. For example, in Fig. 7, the optical translation images of urban and forest scenes from CycleGAN, FG-GAN and SCC-CycleGAN methods all show confusion between grassland and urban scenes. The images in Fig. 8 has a low view angle, which require high scene memory capability of the network. It can be seen that all the other comparison methods except CycleGAN lack scene memory capability and will translate mountain into urban, and the colors of the cities are not accurate. As can be seen from the results in Figs. 5, 6, 7, and 8, the other comparison methods make serious errors when translating SAR images for scenes such as

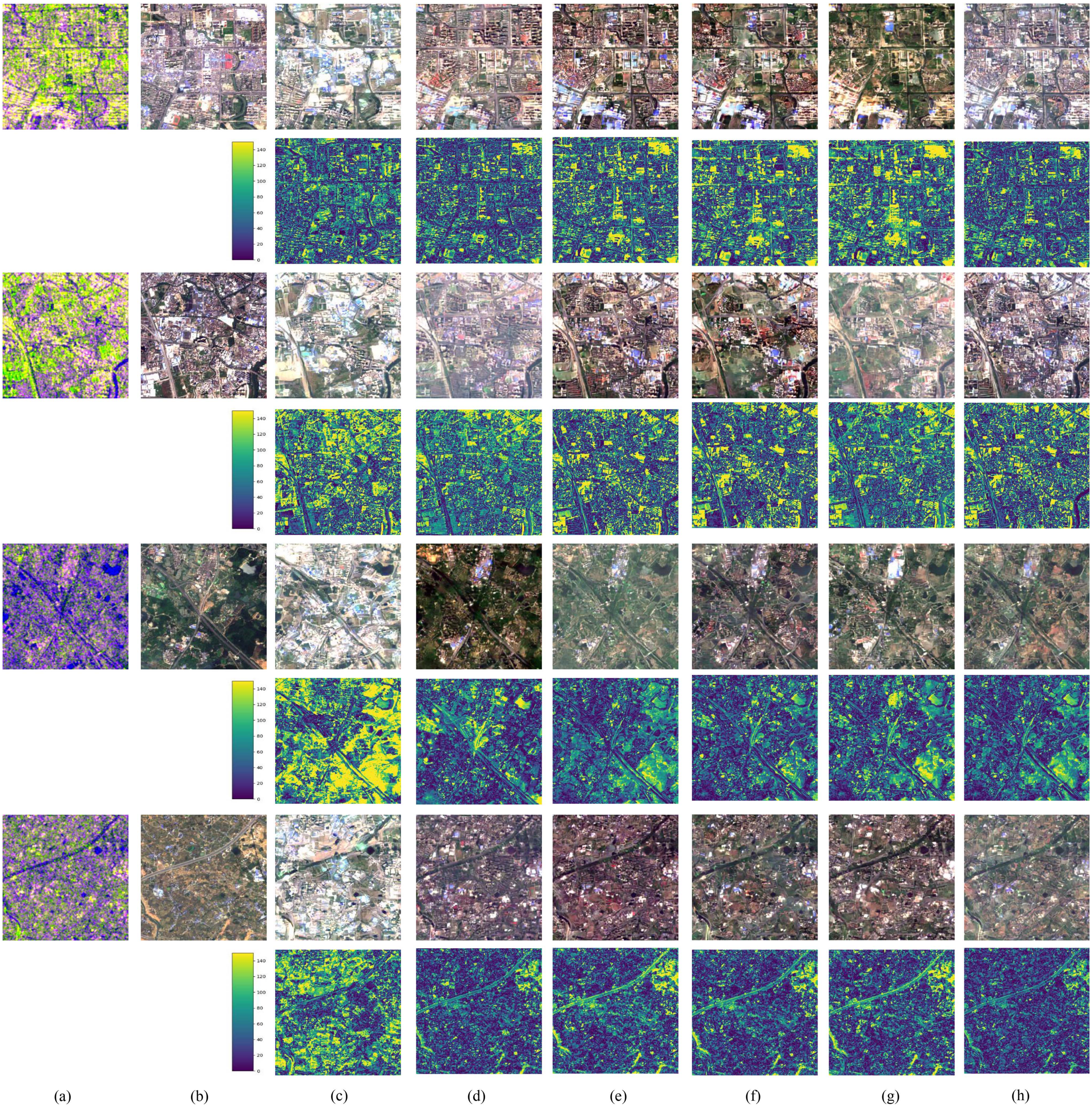

Fig. 7. Visualization results of different methods for S2OIT under different scenes on the WHU-SEN-City dataset. (a) Original SAR image, (b) Real optical image. (c)–(h) Represent the result of DivCo, UGATIT, CycleGAN, FG-GAN, SCC, and our MS-GAN, respectively. The residual results are under the generated optical images of each method. The scenes are all urban.

grassland and forest. The translation results will show various miscellaneous colors (e.g., rows 5 and 7 of Fig. 6). By analyzing the corresponding SAR images, we find that the radar reflections in grassland and forest scenes are relatively uniform, and the amplitude of the actual received radar signals does not differ much. Hence, the difference in the grayscale values of the SAR images is small. In such scenes, the coherence effect becomes obvious, and the speckle noise affects the images more. The noise brings more invalid information to the network, leading to poorer translation results.

To further compare the ability of different methods to memorize the details of the scene, the zoom-in images of the translation results are given in Fig. 9. From the results in Figs. 5, 6, 7, and 8, it can be seen that the overall color of the images generated by the UGATIT and DivCo methods has a large discrepancy with the real image, so the zoom-in images of the above two methods are not shown in Fig. 9. As can be seen from Fig. 9, the other comparison methods lack the ability to extract the detailed features of the image, especially when dealing with the urban scene images, the edges of the buildings and highways

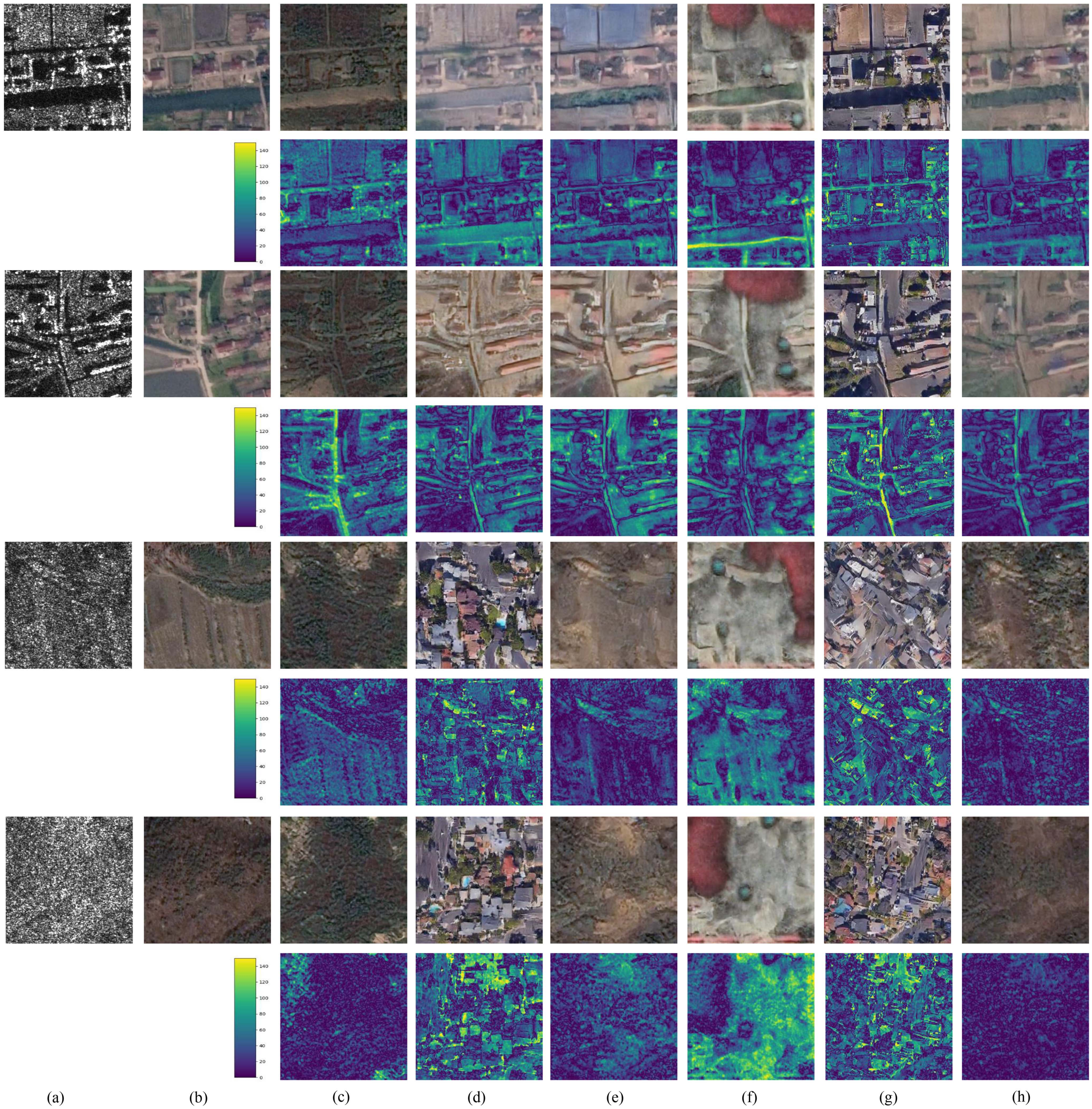

Fig. 8. Visualization results of different methods for S2OIT under different scenes on the QXS-SAROPT dataset. (a) Original SAR image. (b) Real optical image. (c)–(h) Represent the result of DivCo, UGATIT, CycleGAN, FG-GAN, SCC and our MS-GAN, respectively. The residual results are under the generated optical images of each method. The scenes are urban, urban, mountain, and mountain, from top to bottom.

will be seriously missing, and the buildings in the same area will be translated together by the network, which is not able to distinguish the actual scenes.

In contrast, our proposed MS-GAN gives the best translation results for SAR images of different scenes in all three datasets and has the most outstanding ability to suppress speckle noise and memorize scene details, as the visualization results and residual images show. Due to the addition of MRG and MFD, MS-GAN has a multiscale receptive field and feature extraction ability, which can not only memorize the scenes with large differences, such as urban and forest, but also distinguish the scenes with small differences, such as forest and grassland (e.g., rows 5 and 7 in Fig. 6), and can relatively accurately memorize more colors in color-rich scenes, and in the building-dense scenes. MS-GAN can generate more image details and improve the accuracy of translated image. Meanwhile, the SSD module in MS-GAN can suppress the speckle noise in SAR images and reduce the effect of noise on image translation, which can output more stable and reliable results for scenes such as grassland and forest that are more affected by noise. Overall,

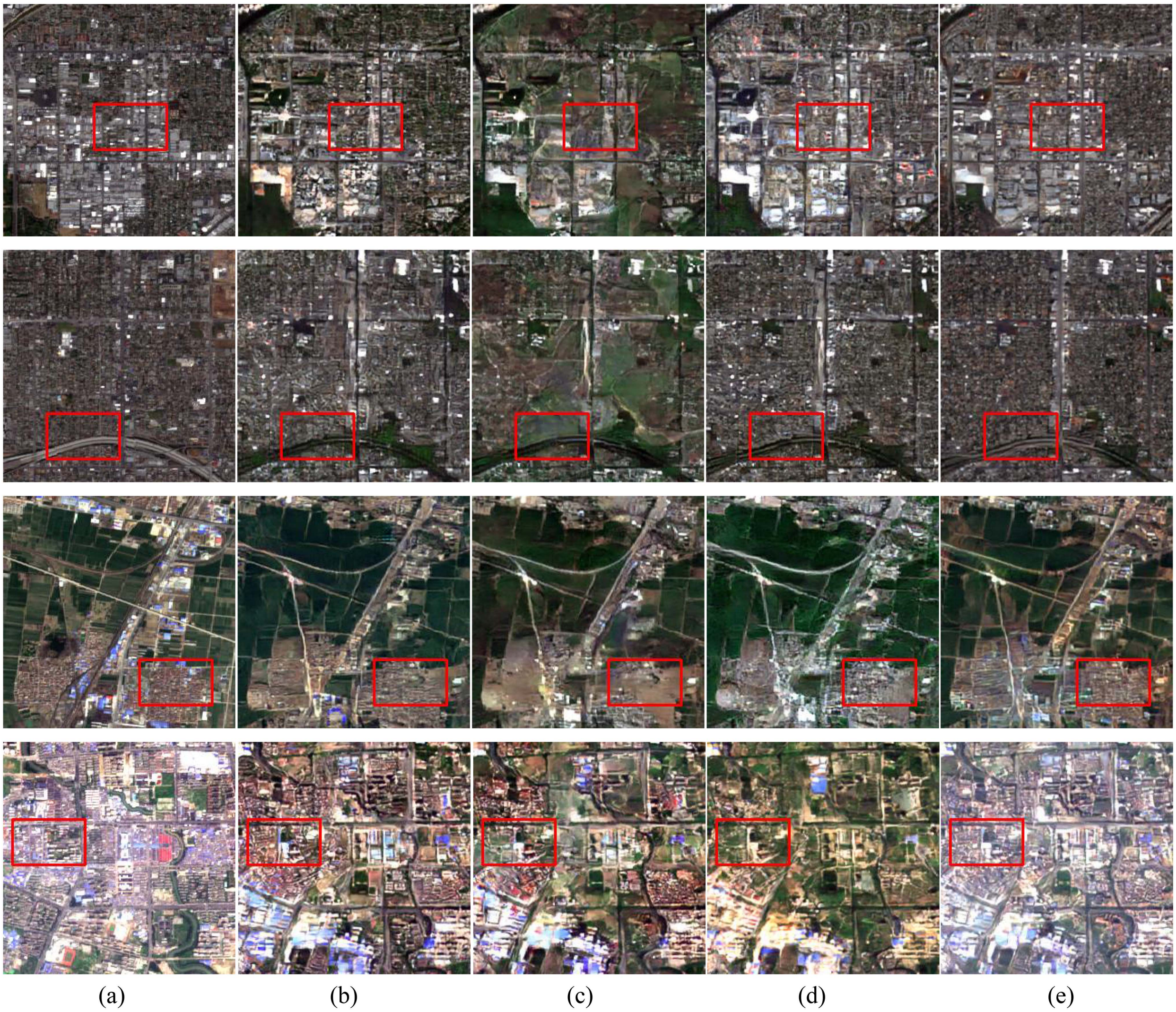

Fig. 9. Zoom-in results of different methods for S2OIT on the SEN1-2 and QXS-SAROPT datasets. (a) Real optical image. (b)–(e) represent the result of CycleGAN, FG-GAN, SCC, and our MS-GAN, respectively. The regions of interest are marked with red boxes, the first three rows of images are from the SEN1-2 dataset, and the fourth row of image is from the QXS-SAROPT dataset.

TABLE V
DIFFERENT COMPONENTS IN THE PROPOSED METHOD ON DIFFERENT DATASETS

	SEN1-2					WHU-SEN-City					QXS-SAROPT				
	PSNR↑	SSIM↑	FID↓	LPIPS↓	SAM↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	SAM↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	SAM↓
Baseline	12.42	0.2032	132.0	0.5018	11.48	9.55	0.1351	126.3	0.4910	10.61	13.79	0.2908	167.0	0.5978	6.28
+SSD	12.75	0.2141	128.8	0.4976	11.20	9.64	0.1380	122.5	0.4868	10.51	14.05	0.3030	167.0	0.6040	6.26
+MRG	12.56	0.2127	130.0	0.4956	11.31	9.61	0.1366	124.7	0.4903	10.56	14.00	0.2894	188.1	0.5886	6.20
+MFD	12.58	0.2056	138.0	0.5063	11.38	9.87	0.1402	125.9	0.4833	10.28	14.30	0.3035	172.1	0.5790	5.68
+all(MS-GAN)	**13.09**	**0.2211**	**128.7**	**0.4951**	**10.46**	**9.91**	**0.1435**	**120.1**	**0.4821**	**10.09**	**14.62**	**0.3038**	**162.2**	**0.5727**	**5.44**

↑ denotes bigger is better, ↓ denotes smaller is better.
The bold values indicate optimal.

the images generated by MS-GAN have more accurate colors, richer landscapes, are less affected by noise, and have better performance in general.

3) Ablation Study:

Validation of module effectiveness. We evaluate each component in the proposed MS-GAN, i.e., SSD, MRG, and MFD. Table V gives the comparison results of the five evaluation metrics on the three datasets, based on baseline (CycleGAN) with the addition of each of the three modules and the combination of all of them. +SSD, +MRG, and +MFD denote adding the three modules in baseline, respectively, and +all represents MS-GAN after all combinations. All the parameter settings are the same in the experiment except for the different modules. As can be seen from Table V, each of our proposed modules combined with the baseline significantly improves the PSNR, SSIM, and SAM metrics of the generated images. This indicates that each of our proposed modules can improve the learning ability of the network for the image, and the result generated by each module in combination with baseline is more consistent with the original image in general. However, we also find that adding

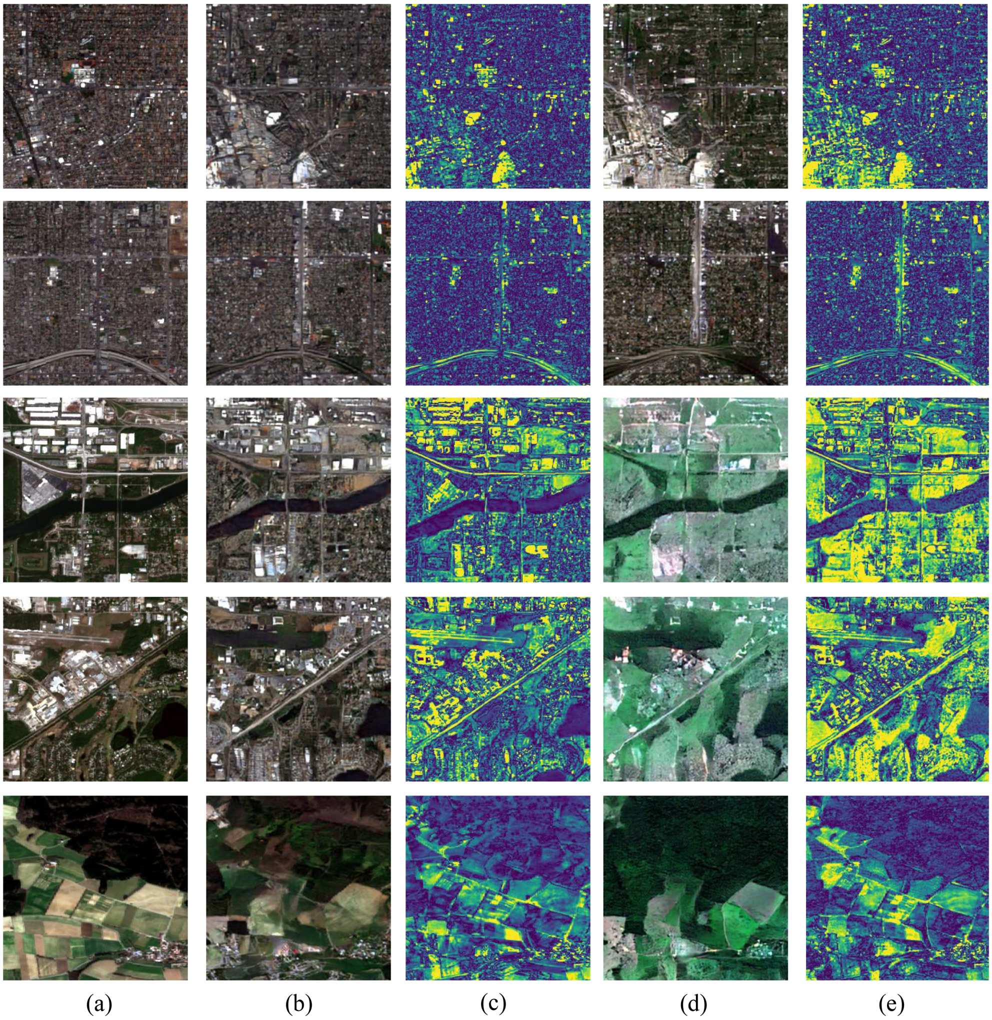

Fig. 10. Visualization results with and without SSD module for S2OIT under different scenes on the SEN1-2 dataset. (a) Real optical image. (b) Result with SSD. (c) Residual of (b), (d) the result without SSD, (e) the residual of (d), respectively. The scenes are urban, urban, semiurban, semiurban, farmland, from top to bottom.

the MRG and MFD module individually sometimes lead to the deterioration of the two metrics, FID and LPIPS. The reasons we analyze are as follows: since FID and LPIPS evaluate the quality of images from the perspective of features through the pretrained network, and the pretrained dataset for these two evaluation metrics is somewhat different from the remote sensing data, which leads to a certain degree of fluctuation in the evaluation metrics. The other reason is that strengthening the discriminator or generator alone will result in one of them being more capable, which cannot guarantee the stability of the GAN adversarial training, thus affecting the results. The image generated by combining all of these three modules (+all) is the best in all the indexes, which shows that our proposed MS-GAN can overcome the above problems.

To further evaluate the effectiveness of the SSD module in MS-GAN, we compare the real optical images in different scenes in the SEN1-2 dataset as well as the generated results with and without the SSD module, and also give the corresponding residual results, as shown in Fig. 10. The color distribution of the residual is the same as in Fig. 5. As can be seen from Fig. 10, the results generated by MS-GAN with SSD are significantly closer to the real optical images and have smaller residuals than the results generated without SSD. Specifically, for the semiurban scene, the color difference between the result without SSD and the real optical image is huge, and almost all the original dark areas become green, and the details of the edges are blurred. For the farmland scene, when SSD is not used, the distinction between different fields is hardly visible, and the color deviation from the real image is also large. It can be seen that the SSD module is able to suppress the effect of speckle noise in SAR images on S2OIT network and improve the quality of the generated optical images.

Hyperparameters Selection: We discuss the values of the hyperparameters in the loss function, and the experiments are all performed on the SEN1-2 dataset, where all other parameters are the same. The results of the discussion are given in Figs. 11 and 12, where the dashed lines represent the optimal results among the other compared methods in Table IV, and the solid lines represent the changes of the evaluation metrics with the weights.

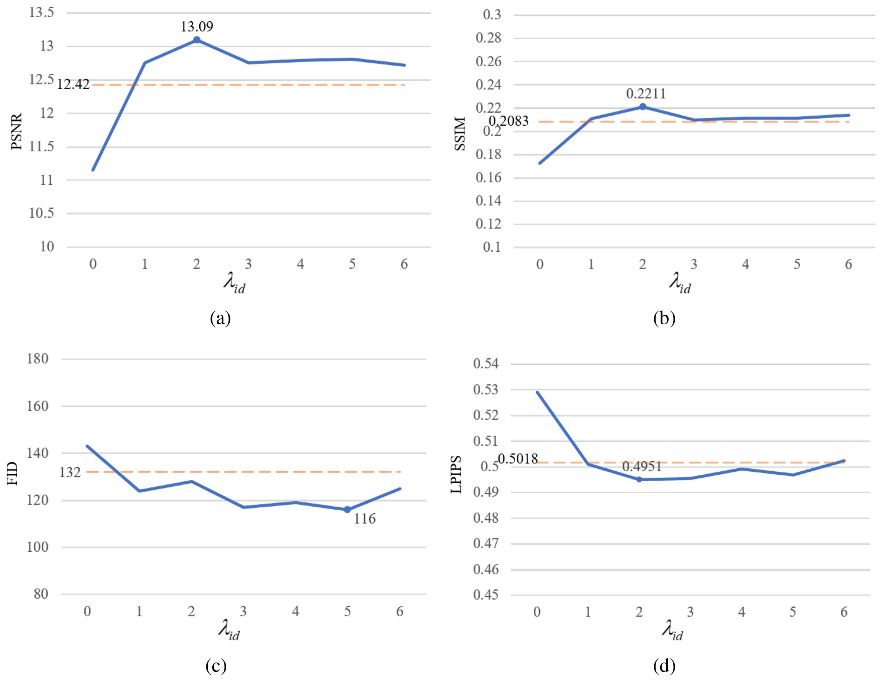

Fig. 11. Comparison results for different λ_{id}. From (a) to (d) denote PSNR, SSIM, FID, and LPIPS metrics, respectively.

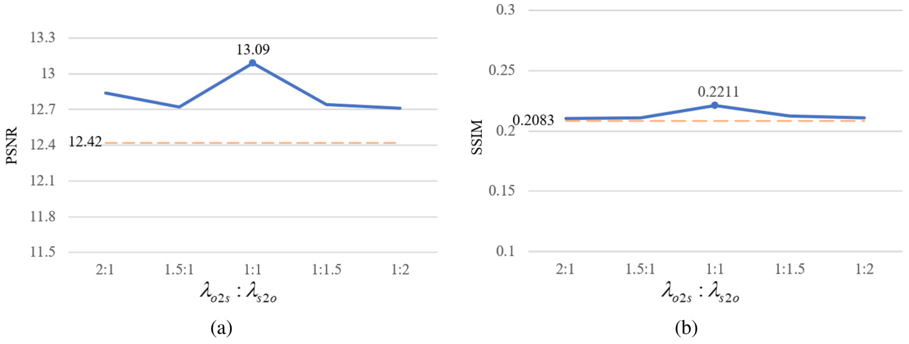

Fig. 12. Comparison results for different $\lambda_{o2s} : \lambda_{s2o}$. (a) and (b) represent PSNR and SSIM, respectively.

Identity loss helps to guide the network to generate stable results and improve the quality of the generated images, while the value of its weight λ_{id} affects the results. Therefore, we selected different values for λ_{id} for the experiments, and the comparison results of the four metrics PSNR, SSIM, FID, and LPIPS, are shown in Fig. 11. The experimental results show that the three metrics reach the optimization when the value of λ_{id} is 2, and the FID index is not much different from the other choices and better than the optimal results in other comparison methods, so we finally chose the value of λ_{id} as 2 in our experiments.

The weight ratio $\lambda_{o2s} : \lambda_{s2o}$ of the cycle-consistency loss in both directions affects the stability of the network, so we select different values for $\lambda_{o2s} : \lambda_{s2o}$ for the experiments, and the results are shown in Fig. 12. Since the values of FID and LPIPS fluctuate in the evaluation process, we determine the weight ratio based on the more stable PSNR and SSIM metrics. The experimental results show that both PSNR and SSIM metrics are the best when the $\lambda_{o2s} : \lambda_{s2o}$ is 1:1, so we finally choose $\lambda_{o2s} : \lambda_{s2o}$=1:1.

MRG network structure parameter selection: The network structure of the MRG will directly affect the effect of image translation. To find the optimal MRG structure, we change the network depth of MR Block and ResBlock on the SEN1-2 dataset for the experiments, and give the comparison results of the evaluation metric PSNR and SSIM, as shown in Table VI. From the comparison results, we can see that PSNR and SSIM do not continue to improve with the deepening of the number of layers of MR Block and ResBlock. When the number of layers of MR Block and ResBlock is 9 and 8, respectively, the results of PSNR and SSIM research the best. When the number of layers of MR Block and ResBlock continues to deepen, the above two metrics have different degrees of decline. Considering the effect of the network and the training cost, we finally chose the structure with the best performance in the experiments, i.e., the number of layers of MR Block and ResBlock is 9 and 8, respectively.

Algorithm implementation efficiency: The execution efficiency of the algorithm directly affects the practical applications.

TABLE VI
COMPARISON RESULTS OF NETWORK STRUCTURE PARAMETER SELECTION FOR MRG

MR Block Layers	ResBlock Layers	PSNR↑	SSIM↑
8	7	12.82	0.2085
	8	12.77	0.2122
	9	12.74	0.2085
9	7	12.45	0.2019
	8	**13.09**	**0.2211**
	9	12.59	0.2108
10	7	12.75	0.2126
	8	12.60	0.2061
	9	12.76	0.2117

↑ means bigger is better.
The bold values indicate optimal.

TABLE VII
ABLATION STUDY OF AVERAGE COMPUTATIONAL INFER TIME ON SEN1-2 DATASET

	Time cost(per image/s)
Baseline	0.221
+SSD	0.348
+MRG	0.279
+MFD	0.291
+all(MS-GAN)	0.371

TABLE VIII
TRAINING AND TESTING TIMES AND MODEL SIZE OF OUR MS-GAN

Model	Infer time per image(s)	Training time (hours)	Generator Param.		Discriminator Param.	
			FLOPs(G)	Param(M)	FLOPs(G)	Param(M)
MS-GAN	0.371	26.03	41.23	12.37	5.37	5.00

We compare the average inference time of different structural networks for a single image, and the experiments are conducted on the SEN1-2 dataset. The results are shown in Table VII. The methods of adding the three modules SSD, MRG, MFD, and the overall combination respectively in Table VII are the same as in Table V. From the comparison results, it can be seen that there is a small increase in computational infer time after combining baseline with SSD, MRG, and MFD modules, and the infer time after combining all the modules, i.e., MS-GAN, is only increased by 0.15 s compared to baseline. However, the two metrics PSNR and SSIM are improved by 5.3% and 8.8%, respectively, as shown in Table V. Therefore, the MS-GAN proposed in this article effectively improves the quality of the generated images with only a small increase in the infer time. We also give the training time of the MS-GAN model on the SEN1-2 dataset and the parameters of our proposed generator and discriminator, as shown in Table VIII. From Tables V, VII, and VIII, it can be seen that our MS-GAN has a small increase in the number of parameters compared to the baseline CycleGAN, and the model training speed and infer time are comparable, but the performance metrics for generated optical images are improved more substantially. The experiments prove that MS-GAN has high implementation efficiency and practical application value.

V. DISCUSSION

Due to the redesigned generator MRG and discriminator MFD, our proposed MS-GAN has a multiscale receptive field and feature extraction ability, which can not only memorize the scenes with large differences, such as urban and forest, but also distinguish the scenes with small differences, such as forest and grassland, and can relatively accurately memorize more colors in color-rich scenes, and in the building-dense scenes. Five evaluation metrics also show that our proposed MS-GAN performs significantly better than the other comparative methods, in which the PSNR, SSIM, LPIPS, and SAM metrics of our MS-GAN are improved by about 6%, 10%, 6%, and 4.3%, respectively. However, due to the unavailability of real optical images paired with SAR images in the training set, the generated optical images generated by our MS-GAN still lose much structural information and small details information, and the image contrast is also lower. Although the above phenomenon is quite common for unpaired image translation methods, a possible attempt can be made to enhance the structural and detail information for the generated optical images by adding some semisupervised auxiliary information, such as scene category information, so that SAR and optical images of the same scene are fed into the network for model training. Furthermore, since the backbone of our method is CycleGAN, which contains image translation networks in both directions, which complicates the process of network training. Therefore, an attempt can be made in subsequent studies to address the lack of learning efficiency by trying other framework such as diffusion model for directly learning translation between two image domains. In addition, future research can utilize more frequency band information in the remote sensing data, as well as using the multispectral data to increase the amount of information, which will be beneficial to the network to extract richer features and further suppression of SAR image noise.

VI. CONCLUSION

In this article, we propose an end-to-end unpaired S2OIT model called MS-GAN, which has a strong scene memorization ability. Our MS-GAN has innovatively designed MRG and MFD modules for the complex scene characterization of SAR images. The MRG and MFD are capable of extracting multiscale features of complex scene of SAR images and giving multireceptive field discriminative results, which leads to strong memorization of complex and rich SAR image scenes. The SSD filters further suppress the effect of speckle noise in SAR images on the quality of the generated results. Comprehensive experiments demonstrate the effectiveness of our proposed MS-GAN, thus our MS-GAN can provide a new general solution framework for the unpaired S2OIT task.

REFERENCES

[1] J. Oh, G. Y. Youm, and M. Kim, "SPAM-Net: A CNN-based SAR Target recognition network with pose angle marginalization learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 701–714, Feb. 2021.

[2] M. Santangelo, M. Cardinali, F. Bucci, F. Fiorucci, and A. C. Mondini, "Exploring event landslide mapping using sentinel-1 SAR backscatter products," *Geomorphology*, vol. 397, 2022, Art. no. 108021.

[3] Y. Chen and L. Bruzzone, "Self-supervised SAR-optical data fusion of sentinel-1/-2 images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[4] Y. Zhao, T. Celik, N. Liu, and H. C. Li, "A comparative analysis of GAN-based methods for SAR-to-optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[5] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimedia*, vol. 24, pp. 3859–3881, 2022.

[6] S. Li, J. van de Weijer, Y. Wang, F. S. Khan, M. Liu, and J. Yang, "3D-aware multi-class image-to-image translation with NERFs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 652–662.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[8] A. Sebastianelli et al., "PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[9] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 14–34, 2021.

[10] X. Yang, Z. Wang, J. Zhao, and D. Yang, "FG-GAN: A fine-grained generative adversarial network for unsupervised SAR-to-optical image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[11] H. Li, C. Gu, D. Wu, G. Cheng, L. Guo, and H. Liu, "Multiscale generative adversarial network based on wavelet feature learning for SAR-to-optical image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[12] Z. Guo, H. Guo, X. Liu, W. Zhou, Y. Wang, and Y. Fan, "SAR2color:learning imaging characteristics of SAR images for SAR-to-optical transformation," *Remote Sens.*, vol. 14, no. 15, 2022, Art. no. 3740.

[13] H. Wang, Z. Zhang, Z. Hu, and Q. Dong, "SAR-to-optical image translation with hierarchical latent features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[14] W. L. Du, Y. Zhou, H. Zhu, J. Zhao, Z. Shao, and X. Tian, "A semi-supervised image-to-image translation framework for SAR–optical image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] M. Schmitt, L. Hughes, and X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 141–146, 2018.

[16] L. Wang et al., "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 136–149, 2019.

[17] M. Huang et al., "The QXS-saropt dataset for deeplearning in SAR-optical data fusion," 2021, *arXiv:2103.08259*.

[18] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Comput. ence*, 2015, doi: 10.48550/arXiv:1511.06434.

[21] J. Birrell, M. Katsoulakis, L. Rey Bellet, and W. Zhu, "Structure-preserving GANs," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1982–2020.

[22] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[23] B. Li, K. Xue, B. Liu, and Y. K. Lai, "BBDM: Image-to-image translation with Brownian bridge diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1952–1961.

[24] C. Jung, G. Kwon, and J. C. Ye, "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 260–269.

[25] J. Guo, J. Li, H. Fu, M. Gong, K. Zhang, and D. Tao, "Alleviating semantics distortion in unsupervised low-level image-to-image translation via structure consistency constraint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 249–259.

[26] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, and Q. Li, "QS-ATTN: Query-selected attention for contrastive learning in I2I translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 291–300.

[27] D. Tan, Y. Lin, and K. Hua, "Incremental learning of multi-domain image-to-image translations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1526–1539, Apr. 2021.

[28] L. Zhang, P. Ratsamee, Z. Luo, Y. Uranishi, M. Higashida, and H. Takemura, "Panoptic-level image-to-image translation for object recognition and visual odometry enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 938–954, Feb. 2024, doi: 10.1109/TCSVT.2023.3288547.

[29] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–18.

[30] R. Liu, Y. Ge, C. L. Choi, X. Wang, and H. Li, "DIVCO: Diverse conditional image synthesis via contrastive generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 377–386.

[31] N. Merkle, S. Auer, R. Mueller, and P. Reinartz, "Exploring the potential of conditional adversarial networks for optical and SAR image matching," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 6, pp. 1811–1820, Jun. 2018.

[32] J. Zhang, J. Zhou, and X. Lu, "Feature-guided SAR-to-optical image translation," *IEEE Access*, vol. 8, pp. 925–937, 2020.

[33] F. N. Darbaghshahi, M. R. Mohammadi, and M. Soryani, "Cloud removal in remote sensing images using generative adversarial networks and SAR-to-optical image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–9, 2022.

[34] J. Wei et al., "CFRWD-GAN for SAR-to-optical image translation," *Remote Sens.*, vol. 15, no. 10, 2023, Art. no. 2547.

[35] H. Liao, J. Xia, Z. Yang, F. Pan, Z. Liu, and Y. Liu, "Meta-learning based domain prior with application to optical-ISAR image translation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, 2023, doi: 10.1109/TCSVT.2023.3318401.

[36] Z. G. Liu, Z. W. Zhang, Q. Pan, and L. B. Ning, "Unsupervised change detection from heterogeneous data based on image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[37] J. W. Goodman, "Some fundamental properties of speckle," *J. Opt. Soc. Amer.*, vol. 66, pp. 1145–1150, 1976.

[38] M. I. H. Bhuiyan, M. O. Ahmad, and M. N. S. Swamy, "Spatially adaptive wavelet-based method using the Cauchy prior for denoising the SAR images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 500–507, Apr. 2007.

[39] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[41] B. Tu, Q. Ren, Q. Li, W. He, and W. He, "Hyperspectral image classification using a superpixel–pixel–subpixel multilevel network," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.

[42] J. Hu, B. Tu, Q. Ren, X. Liao, Z. Cao, and A. Plaza, "Hyperspectral image classification via multiscale multiangle attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024.

[43] R. Wang, Y. Nie, and J. Geng, "Multiscale superpixel-guided weighted graph convolutional network for polarimetric SAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3727–3741, 2024.

[44] D. Wang, Y. Song, J. Huang, D. An, and L. Chen, "SAR target classification based on multiscale attention super-class network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9004–9019, 2022.

[45] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[46] J. N. Turnes, J. D. B. Castro, D. L. Torres, P. J. S. Vega, R. Q. Feitosa, and P. N. Happ, "Atrous cGAN for SAR to optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[47] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–14.

[48] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 2366–2369.

[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.

[50] F. A. Kruse, A. B. Lefkoff, and J. W. Boardman, "The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2–3, pp. 145—163, 1993.

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

Zhe Guo received the B.S., M.S., and Ph.D. degrees in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2005, 2008, and 2012, respectively.

She is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University. Her research interests include image processing, computer vision, and pattern recognition.

Zhibo Zhang received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China in 2021. He is currently working toward the M.S. degree in information and communication engineering with the School of Electronics and Information, Northwestern Polytechnical University.

His research interest is image processing.

Qinglin Cai received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2019. He is currently toward the M.S. degree in information and communication engineering with the School of Electronics and Information, Northwestern Polytechnical University.

His research interest is image processing.

Jiayi Liu received the B.S. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2023. She is currently working toward the a M.S. degree in information and communication engineering with the School of Electronics and Information, Northwestern Polytechnical University.

Her research interest is image processing.

Yangyu Fan received the M.S. degree in electromechanical engineering from Shaanxi University of Science and Technology, Xi'an, China, in 1992, and the Ph.D. degree in acoustics signal processing from Northwestern Polytechnical University, Xi'an, China, in 1999.

He is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. His research interests include signal processing, image processing, virtual reality, pattern recognition, and optical communications.

Shaohui Mei (Senior Member, IEEE) received the B.S. degree in electronics and information engineering and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively.

He was a Visiting Student with the University of Sydney, Camperdown, NSW, Australia, from 2007 to 2008. He is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. His research interests include hyperspectral remotesensing image processing and applications, intelligent signal and information acquisition and processing, video processing, and pattern recognition.

Dr. Mei received the First prize of Natural Science Award of Shaanxi Province in 2022, the Excellent Doctoral Dissertation Award of Shaanxi Province in 2014, the Best Paper Award of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems in 2017, the Best Reviewer of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing in 2019, and IEEE Transactions on Geoscience and Remote Sensing in 2022. He is currently an Associate Editor for IEEE TGRS and IEEE JSTARS, a Guest Editor for Remote Sensing, and the Reviewer for more than 30 international famous academic journals.