

Received March 26, 2020, accepted April 5, 2020, date of publication April 10, 2020, date of current version April 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987105

Feature-Guided SAR-to-Optical Image Translation

JIEXIN ZHANG^{ID}, JIANJIANG ZHOU, (Member, IEEE), AND XIWEN LU

Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Corresponding author: Jixin Zhang (zhangjixin96@163.com)

This work was supported in part by the National Youth Science Foundation of China under Grant 61501228, and in part by the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of education, Nanjing University of Aeronautics and Astronautics, China.

ABSTRACT The powerful performance of Generative Adversarial Networks (GANs) in image-to-image translation has been well demonstrated in recent years. However, most methods are focused on completing an isolated image translation task. With the complex scenes in optical images and high-frequency speckle noise in SAR images, the quality of generated images is often unsatisfactory. In this paper, a feature-guided method for SAR-to-optical image translation is proposed to better take the unique attributes of images into account. Specifically, in view of the diversity of structure features and texture features, VGG-19 network is used as the feature extractor in the task of cross-modal image translation. To ensure the acquisition of multilayer features in the process of image generation, feature matching is carried out on different layers. Loss function based on Discrete Cosine Transform is designed to filter out the high-frequency noise. The generated images show better performance in feature preservation and noise reduction, and achieve higher Image Quality Assessment scores compared with images generated by some famous methods. The superiority of our algorithm is also demonstrated by being applied to different networks.

INDEX TERMS SAR-to-optical image translation, feature extraction, high-frequency noise, generative adversarial networks.

I. INTRODUCTION

With the rapid development of aerial remote sensing technology, researches concerned with earth observation system and remote sensing technology used for monitoring the state of industry, agriculture and forestry have gradually become prevailing in earth system science and information science [1]–[3]. Remote sensing images both near-infrared and obtained with Synthetic Aperture Radar (SAR) technology in microwave band have come to be effective means for earth observation [4], [5]. Optical images are more popular because they are closer to the perception of human eyes. However, they are vulnerable to the effects of the environment [6]. While microwave electromagnetic wave, on which SAR images depend, can make up for this defect, passing through the haze, cloud and cover which visible light electromagnetic wave can't pass through [7]. Unfortunately, there are a lot of speckle noise in SAR images [8], [9], making untrained people unable to get useful information and leading to more than 90% waste of resources. Although images with higher spatial resolution can be obtained by image de-noising, there are still many difficulties in SAR image interpretation. Firstly,

radar signal wavelengths range from mm to cm, while human eyes are more familiar with the perspective projections and signals in the visible part of the electromagnetic spectrum [10]. Secondly, the significant features in SAR images are opportunistic, sparse, and often do not correspond to the physical structures, which is due to signal multiple reflections [11]. Finally, the loss of calculation accuracy in signal processing and the dependence of distance along the range axis lead to geometric distortions [12]. Therefore, additional means are needed to facilitate the interpretation of SAR images. If we can make use of SAR images which can be obtained conveniently and translate them into optical images easy to be recognized, we would be able to make the most of existing resources and greatly improve the efficiency of earth observation.

Similar to automatic language translation, image-to-image translation task is to transform one kind of representation of a scene into another [13]. In recent years, as a new framework of generative model in deep learning [14], Generative Adversarial Networks (GANs) plays an important role in image translation [15]–[17]. In 2016, Isola *et al.* proposed a framework pix2pix [18] for image-to-image translation based on Conditional Generative Adversarial Networks (CGANs) [19]. Wang *et al.* proposed a high-resolution framework

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Ayoub Khan ^{ID}.

pix2pixHD [20] based on pix2pix framework in 2018, which can generate high-resolution images. Pix2pix and pix2pixHD need pairs of images as training data, but in many circumstances, such pairs of images are difficult to find. In order to solve this problem, Zhu *et al.* put forward an unpaired image translation framework using CGANs in a circular consistent way [21], which consists of two separate generators, one from domain A to domain B, while the other one from domain B to domain A. Recently, more and more researchers concentrate on SAR-to-optical image translation based on GANs [22]–[25]. However, a common problem in the existing methods is that, more attention has been paid on completing an isolated image translation task, without thinking about the imaging process involved in image generation.

Generation of images can not be separated from the analysis of original images and target images [26]–[28]. The goal of image analysis is to extract the description parameters which can express the important information of images concisely, and to describe the content of images quantitatively, which is called feature extraction [29]. In the traditional field of computer vision, SIFT [30] and HOG [31] based on gradient histogram are usually used to accomplish this task. With the advent of AlexNet [32], attention has been concentrated back on Convolutional Neural Networks (CNNs) [33], and as a beginning, a variety of networks such as VGG, GoogLeNet and ResNet have sprung up like mushrooms. VGG [34] is a convolutional neural network model proposed by Simonyan *et al.* It achieves 92.7% of the top-5 test accuracy in the ImageNet classification experiment. GoogLeNet [35] designs a module called "Inception" and uses normal dense structure to approximate sparse CNN. Its top-5 test accuracy attains to 93.3% and it is much faster than VGG. ResNet [36] is introduced by Kaiming *et al.* to solve the problem that the neural network is more difficult to train when it is going deeper. Deep neural networks have been widely used in feature extraction and image recognition [37]–[40], but even the same network would show different capabilities in different study. It is of great importance to find the superior network for SAR-to-optical task and apply it to the process of image translation.

Another key point in SAR-to-optical image translation is noise filtering. Spot-shaped speckle noise in SAR images comes mainly from the loss of calculation accuracy in signal processing and the distortion in the process of imaging [41]–[43], different from Gaussian noise in optical images [44]. In traditional methods, people use Fourier Transform (FT) to convert images from spatial domain to frequency domain [45] in order to better characterize and remove high-frequency speckle noise, but it will bring about a problem [46]: since the value of spatial domain only contains the real part of complex numbers, the imaginary part in the results of frequency domain would cause many troubles to the calculation. In order to avoid the appearance of imaginary part in the results, transformation bases without imaginary numbers are needed.

In this paper, we firstly look for the superior network for SAR-to-optical task to extract the features of SAR images and optical remote sensing images, and then take feature matching to guarantee the feature similarity. Meanwhile, we take advantage of Discrete Cosine Transform (DCT) to filter out high-frequency noise and avoid imaginary part in the results of frequency domain. Because most of the energy in images is focused on the low-frequency band [47], DCT can also extract the underlying information of images. Specifically, the major contributions of our study are as follows:

- 1) Unlike most existing methods which are focused on completing an isolated image translation task, this paper proposes a feature-guided SAR-to-optical translation method, which takes unique attributes of multimodal images such as the complex scenes in optical images and high-frequency speckle noise in SAR images into account.
- 2) We use a novel multilayer feature matching module based on VGG's superior ability of cross-modal feature extraction, which can significantly preserve more information than networks without it when applied to the task of SAR-to-optical image translation.
- 3) We improve the loss function by incorporating a novel loss based on DCT. With the low-pass filtering property and characteristics of no imaginary number in transformation bases, it outperforms some famous methods in noise reduction and is easier to be calculated.

The rest of this paper is structured as follows: In Section II, preparations such as datasets used in experiments and the network architectures constructed to translate SAR images to optical images are introduced. In Section III, the process of multimodal feature extraction is displayed, and the feature-guided loss functions are designed. In Section IV, experiment results of SAR-to-optical image translation are presented and analyzed. In order to demonstrate the superiority of the proposed method, comparison with some famous methods and a series of ablation studies are performed and discussed. Finally, the conclusions and outlooks are illustrated in Section V.

II. PREPARATIONS

A. DATASETS

We take advantage of a paired SAR-Optical remote sensing image dataset provided by Schmitt *et al.* [48]: the SEN1-2 dataset, which has been proved to be feasible in SAR-to-optical image translation [10]. It contains 282,384 pairs of SAR-Optical images which cover various regions and four seasons around the world. In this paper, three kinds of dataset for different tasks are obtained by random sampling from the SEN1-2 dataset: The SEN1-2 Modality Recognition Dataset, The SEN2 Texture Classification Dataset and The SEN1-2 Image Translation Dataset. The SEN1-2 Modality Recognition Dataset used for image modality feature extraction contains the same number of unpaired SAR-optical images. The SEN2 Texture Classification Dataset, where different

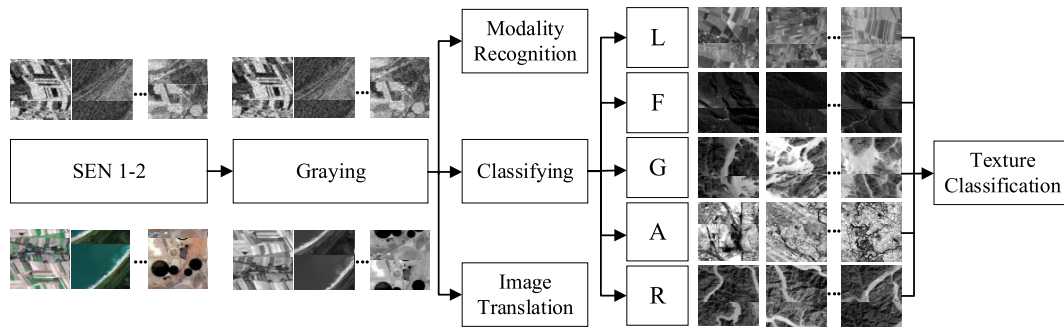


FIGURE 1. The steps of datasets preparation for different experiments. Acronyms in the boxes are as follows: L: FarmLand, F: Forest, G: Gorge, A: Anthropic zone, R: River.

categories of scenes are covered, is used for feature extraction of optical remote sensing images. The SEN1-2 Image Translation Dataset used for SAR-to-optical image translation includes paired SAR-optical images to ensure the transmission of features across modalities. Each dataset contains a training set and a test set. Images in test sets are only used to validate the performance of algorithm, and are not used in training. Scenes and the specific number of images involved in each dataset are shown in Table 1. Our experiments are performed on images with single-channel for the following reasons: firstly, images in SEN-1 acquired from space-borne SAR are single channel. Secondly, transformation from a single channel SAR image to a multi-channel optical image is expected to be an ill-posed problem, as is the colorization of gray-scale images in computer vision [49]. The difficulty of the task is increased by the variability of surface characteristics such as surface standard deviation and correlation length, permittivity, or chemical composition [10]. Finally, transformation from a single channel image to a multi-channel image would lead to the phenomenon that total amount of information can not correspond to each other, causing unnecessary troubles. Thus, we gray the original colorful images in SEN2, using weighted average method [50]. The steps to prepare the datasets are shown in Fig. 1.

B. NETWORK ARCHITECTURES

1) GENERATOR BASED ON RESIDUAL BLOCKS

The network structure of SAR-to-optical generator consists of three parts: encoding module of SAR images, translation module of SAR-to-optical and decoding module of optical images, which is inspired by [51], [52]. Specific architectures are displayed in Fig. 2(a).

In encoding module, the input is passed through a series of layers that progressively down-sample, until a bottleneck layer, at which point the process is reversed [18]. The coding expression and the high-level semantic information of SAR images are obtained in this part. It is composed of five convolutional modules. The input is SAR images with resolution of 256×256 , and the output is SAR image coding tensors with 1024 channels and 8×8 size. The 7×7 convolutional layer is used to expand the receptive field, so that SAR

TABLE 1. The information of SAR images and optical images in SEN1-2 modality recognition dataset, SEN1-2 feature classification dataset, and SEN1-2 image translation dataset.

Dataset	Scene	SAR	Optical
Modality Recognition Dataset	Train	Various	9075
	Test	Various	1134
Texture Classification Dataset	Train	Farmland	445
		Forest	483
		Gorge	-
		Anthropic zone	1024
		River	2486
	Test	Farmland	409
		Forest	18
		Gorge	73
		Anthropic zone	98
		River	55
Image Translation Dataset	Train	Various	1069
	Test	Various	563

image tensors can contain more neighborhood information and avoid interference caused by speckle noise.

SAR-to-optical translation module uses a multilayer residual structure to translate the coding tensors from SAR images to optical images. It is constituted by nine residual blocks, which are used to increase the depth of network and make the feature coding more fully converted. The input is SAR image coding tensors and the output is optical image coding tensors, both of which are 8×8 size.

Decoding module of optical images utilizes the transposition convolution structure to sample encoding tensors of optical images up to optical images of the same size as the input SAR images. The input is optical image coding tensors with channel number of 1024 and size of 8×8 , and the output is optical images with resolution of 256×256 . The 7×7 deconvolutional layer is used to ensure that more information of adjacent pixel units is considered in the generation of optical images.

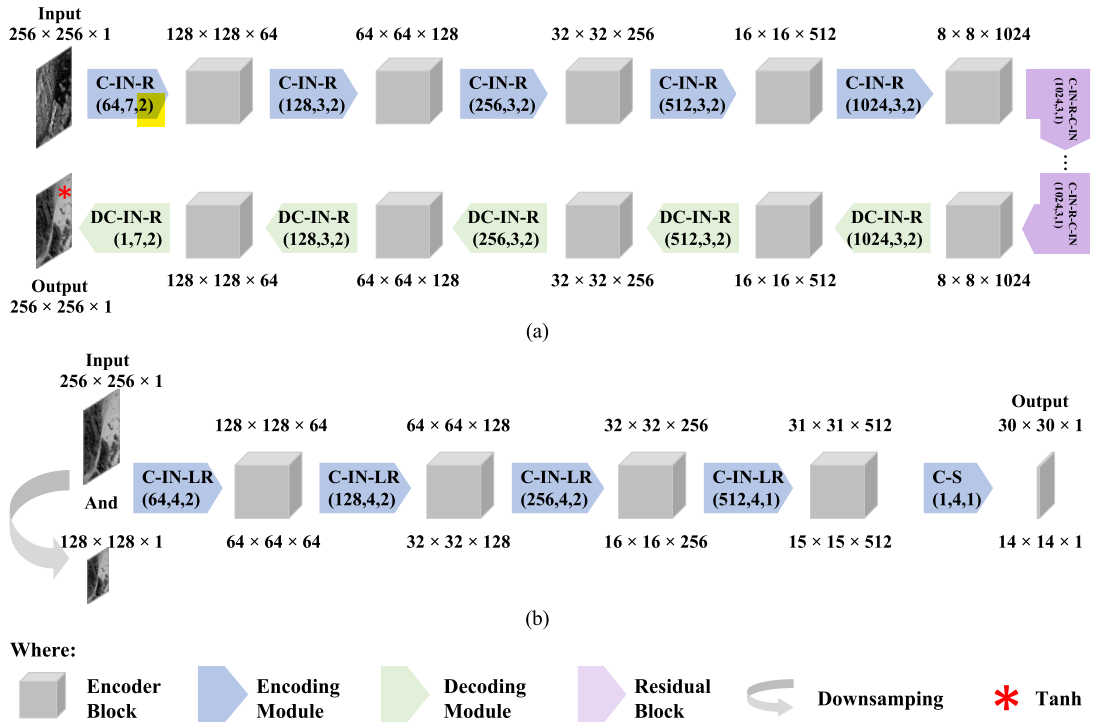


FIGURE 2. Architectures of modules contained in (a) Generator. (b) Discriminator. The numbers in brackets refer to the number of filters, filter size and stride respectively. The numbers above or below the encoder blocks indicate the input and output size of each module. Acronyms shown in modules are as follows: C: Convolution, DC: Deconvolution, IN: Instance Normalization, R: ReLU, LR: Leaky ReLU, S: Sigmoid.

2) MUTI-SCALE DISCRIMINATOR

In this work, we use two discriminators with different resolutions of images input but the same network architecture shown in Figure 2(b), which is inspired by [20]. They are called high-resolution image discriminator D_1 and low-resolution image discriminator D_2 respectively, which can not only ensure a large receptive field, but also restrain the increase of model parameters, saving the training memory. The detailed work steps are as follows:

- 1) Down-sample the real optical image and the generated optical image with a coefficient of 2 to create an image pyramid with two scales.
- 2) Through training the discriminators D_1 and D_2 , real images and generated images of different resolutions are distinguished. For high-resolution image discriminator, the input size is 256 × 256, and the output size is 30 × 30. For low-resolution image discriminator, the input size is 128 × 128, and the output size is 14 × 14.
- 3) Calculate the output matrices and the label matrices to optimize the two discriminator networks.

Although the discriminators have the same network architecture, they hold different sizes of receptive fields. The low-resolution image discriminator has a larger receptive field, thus owns a more global image field of vision, and can guide to generate an overall consistent image. The high-resolution image discriminator has a smaller receptive field, so it can

better describe the details and lead the generator to focus on smaller points. Multi-scale discriminator not only pays attention to high-level global information of images, making the overall structure view more appropriate, but also ensures that low-level details of images are not ignored.

In order to better model high-frequency features, the structure PatchGAN [18] is used in this paper, which is more attentive to local image blocks and effectively models images as a Markov random field. Instead of inputting the whole image, it only needs to judge true or false of each patch whose size is $N \times N$. Different patches are independent of each other. We cut an image into various patches with the same size of 70 × 70, and then let the discriminators judge each patch, taking the average of all the results as the final output. Leaky ReLU inspired by [53] is also used to ensure the resolution of generated images.

III. FEATURE EXTRACTION AND APPLICATION

A. MULTIMODAL FEATURE EXTRACTION

Because SAR images and optical images belong to different modalities, the task of SAR-to-optical image translation can be regarded as multimodal image-to-image translation [54]. In order to extract modality features from SAR images and optical images, a task of SAR-optical image modality recognition is designed. A binary classification network, which is used as an image modality feature extractor, is trained by inputting unknown modality images and determining whether

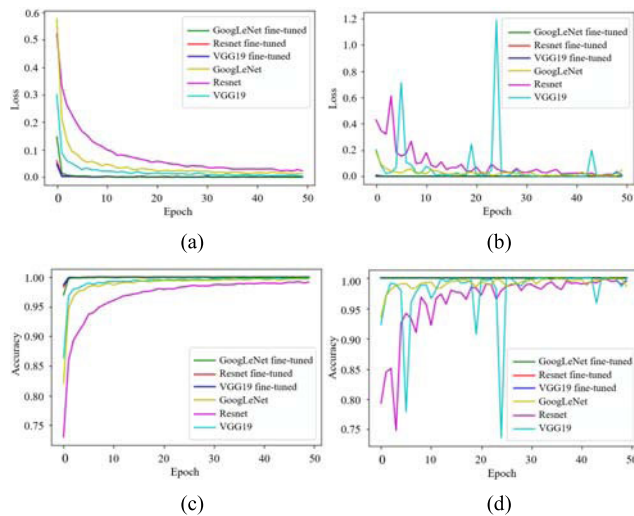


FIGURE 3. Loss function curve and accuracy curve of image modality feature extraction experiments on training sets and test sets. (a) Loss function curve of training sets. (b) Loss function curve of test sets. (c) Accuracy curve of training sets. (d) Accuracy curve of test sets.

they are SAR images or optical images. Similarly, in order to extract texture features of objects in optical images, a multi-classification task for the object types of optical images is designed. A multi-classification network, which is used as an image texture feature extractor, is trained by inputting and classifying optical remote sensing images with unknown type.

In this task, 19-layer VGG, GoogLeNet, and 18-layer ResNet are chosen for comparative experiments. Single-channel images with resolution of 256×256 are input. Two training methods: from scratch and fine-tuned, are used to comprehensively examine the robustness and feature extraction performance of the network. The batch size is set to 64. SGD optimizer is selected to optimize the network, and the momentum is 0.9. Traditional cross entropy loss [55] is chosen as the loss function. The network trains 50 epochs to ensure model convergence. Learning rate is set to a fixed value of 0.001. After the experiments, networks with best capabilities tend to be selected as our feature extractor.

1) MODALITY FEATURE EXTRACTION

Fig. 3 shows the results of three different networks in two training methods. The loss function curve shows the convergence speed of the model in 50 epochs, while the accuracy curve shows the ratio of the number of correct samples to the total number of samples in whole classification experiment. Through the curve of loss function, it can be found that networks fine-tuned converge faster than networks trained from scratch and all three networks converge within one epoch. From the accuracy graph, it is demonstrated that networks fine-tuned achieve higher accuracy in the classification results than networks trained from scratch. Unfortunately, the loss curve and accuracy curve of networks trained from scratch are very unstable in test sets, showing acute fluctu-

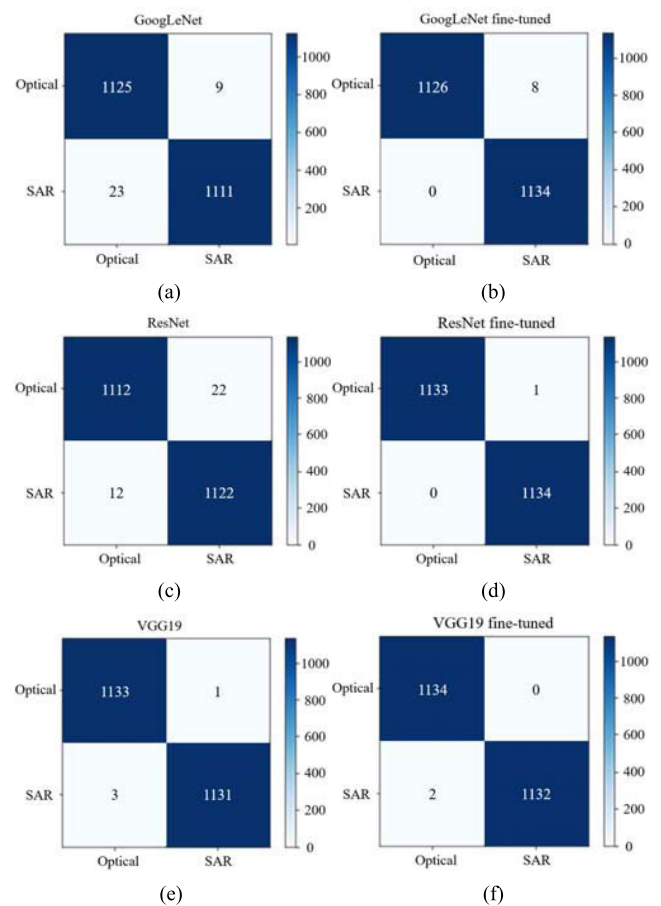


FIGURE 4. Confusion matrices of image modality feature extraction experiments. (a) (c) (e) Confusion matrices of models trained from scratch. (b) (d) (f) Confusion matrices of models fine-tuned.

tuations. Because networks without pre-trained model lack well training on a large amount of image data and computing resources. When there are many parameters, it is difficult to ensure the robustness and feature extraction performance of the model, which leads to worse results than networks fine-tuned.

What is worth mentioning is that, GoogLeNet trained from scratch does not fluctuate greatly, which benefits from Inception module that extracts features on the scale of multiple receptive fields. However, it performs worse than

VGG19 both in convergence speed and accuracy approximation, so is ResNet. The performance of accuracy can be seen more clearly in confusion matrices shown in Fig. 4. In confusion matrix, each column represents the predicted category, of which the total number is the number of images predicted to be that category, while each row stands for the true category, where the total number is the actual number of that category. In the classification of all images, VGG19 trained from scratch only misclassifies four images, while its model fine-tuned categorizes two images incorrectly, showing that VGG model without pre-training

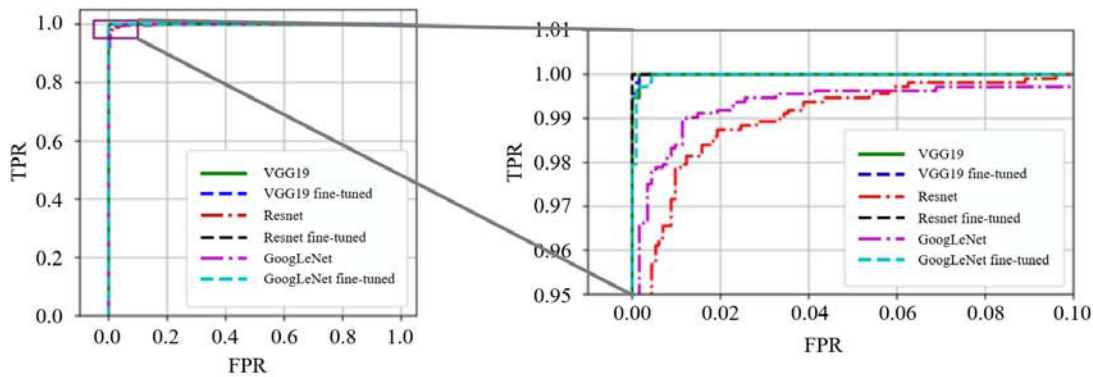


FIGURE 5. ROC diagram of image modality feature extraction experiments.

TABLE 2. Quantitative results of models used in image modality feature extraction experiments.

Model	Accuracy	Precision	Recall	F1-Measure	AUC
GoogLeNet	0.9859	0.9921	0.9800	0.9860	0.9991
GoogLeNet fine-tuned	0.9965	0.9929	1.0000	0.9965	0.9999
ResNet	0.9850	0.9806	0.9893	0.9849	0.9999
ResNet fine-tuned	0.9996	0.9991	1.0000	0.9996	1.0000
VGG19	0.9982	0.9991	0.9974	0.9982	0.9999
VGG19 fine-tuned	0.9991	1.0000	0.9982	0.9991	0.9999

achieves effects similar to the model pre-trained, which is also demonstrated in ROC diagram of Fig. 5.

Table 2 compares the results of six different models through five indexes, namely, Accuracy, Precision, Recall, F1-Measure and AUC. Accuracy denotes the ratio of the number of correct predictions to the total number of samples, representing the accuracy of overall prediction. Precision is the ratio of the number of positive samples correctly predicted to the total number of positive samples predicted, and Recall is the ratio of the number of positive samples correctly predicted to the actual number of positive samples, both of which suggest the accuracy of positive samples prediction. F1 measures the relationship between Precision and Recall, and is a balance point value obtained by judging and weighing the two. AUC which ranges between 0 and 1 is the area under the curve of ROC. The larger the area is, the better the result is. It can be clearly found from Table 2 that after networks being pre-trained, all indexes are significantly improved. Among networks trained from scratch, VGG achieves the highest value in all indexes, and among networks fine-tuned, not only the precision of VGG reaches the highest value of 1.0, but also its other indexes are all above 0.998. It can be concluded that: compared with GoogLeNet and ResNet18, VGG19 model has a stronger feature extraction ability in extracting modality information between SAR images and optical images.

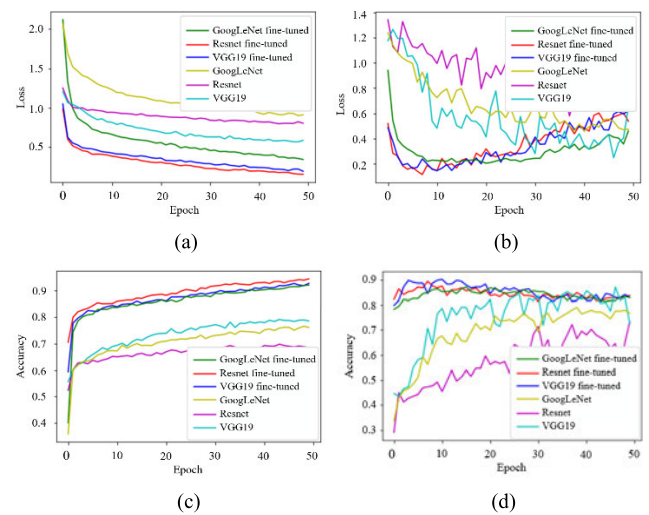


FIGURE 6. Loss function curve and accuracy curve of image texture feature extraction experiments on training sets and test sets. (a) Loss function curve of training sets. (b) Loss function curve of test sets. (c) Accuracy curve of training sets. (d) Accuracy curve of test sets.

2) TEXTURE FEATURE EXTRACTION

Fig. 6 shows the multi-classification experiment results of the networks trained in two different modes. Similar to image modality feature extraction, fine tuning networks, whose accuracy is higher, are also easier to converge than networks trained from scratch. By comparing curves of training sets and test sets, it can be found that after the 5th-8th epoch, loss functions of models in training sets are still decreasing, which start to rise in test sets. Simultaneously, accuracy curves in training sets are still climbing, but begin to decline in test sets. It is more obvious on graphs of fine-tuning models, which reflects that after the 5th-8th epoch, models have been in the state of over-fitting.

By comparing results of different models, it can be demonstrated that within models trained from scratch, whether in training sets or test sets, VGG shows the best performance in convergence and accuracy approximation. In fine-tuning models, although VGG is not as good as ResNet in training

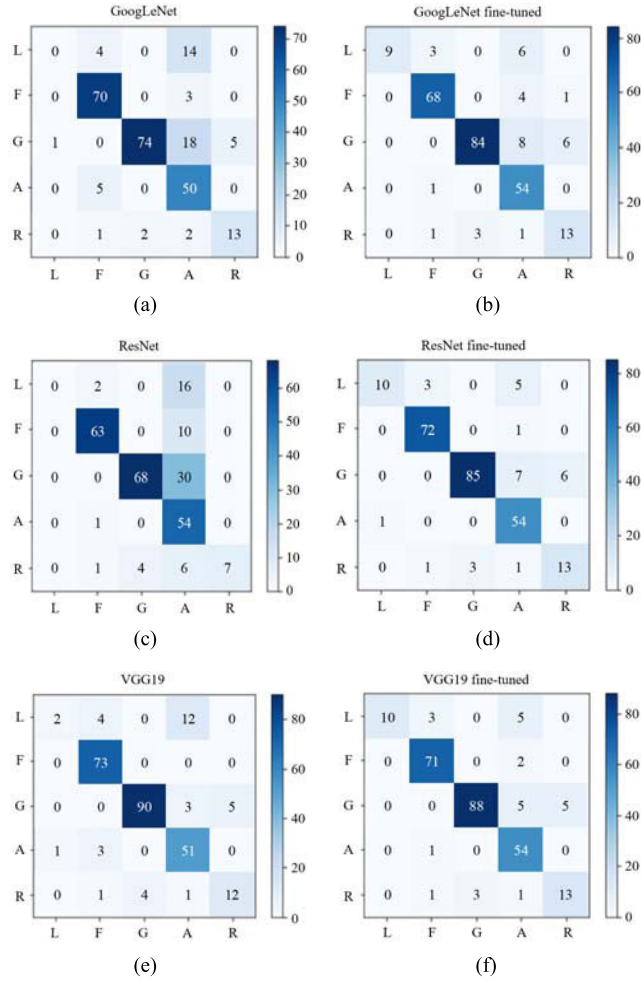


FIGURE 7. Confusion matrices of texture feature extraction experiments. (a)(c)(e) Confusion matrices of models trained from scratch. (b)(d)(f) Confusion matrices of models fine-tuned. Acronyms in the matrices are as follows: L: FarmLand, F: Forest, G: Gorge, A: Anthropic zone, R: River.

set, it still shows the highest competence in test set. The specific performance of accuracy can be seen more clearly in confusion matrices displayed in Fig. 7 and quantitative results in Table 3. After 46 epochs training, VGG trained from scratch achieves the same accuracy rate of 87.02% as GoogLeNet pre-trained, and the fine-tuning VGG achieves the highest accuracy of 90.08% at the 10th epoch. Conclusions can be drawn that: compared with GoogLeNet and ResNet18, VGG19 still has stronger feature extraction capability in extracting texture information of optical images.

B. FEATURE-GUIDED LOSS FUNCTIONS

1) FEATURE EXTRACTION LOSS

According to feature extraction experiments introduced above, VGG19 is finally selected as the feature extractor of SAR-to-optical image translation network, and the loss function of feature extraction is designed. Considering that the deeper layers tend to extract the higher level of spatial structure and semantic information, while shallow network

TABLE 3. Quantitative results of models used in image texture feature extraction experiments.

Model	Precision
GoogLeNet	0.7901
GoogLeNet fine-tuned	0.8702
ResNet	0.7328
ResNet fine-tuned	0.8931
VGG19	0.8702
VGG19 fine-tuned	0.9008

extracts the underlying features such as structure and texture, we take feature matching before each max pooling layer of VGG network in order to encourage generated images to have both high-level and low-level feature representations similar to real ones, as is shown in Fig. 8. Meanwhile, inspired by [55], VGG loss function is designed as follows:

$$L_{VGG}(G, D) = \sum_{i=1}^5 \frac{1}{H_i W_i C_i} \|\phi_i(G(X)) - \phi_i(Y)\|_2^2 \quad (1)$$

where $\phi_i(\cdot)$ is the input characteristic diagram of the i th max pooling layer. $G(\cdot)$ is the generator network. X and Y correspond to the input SAR images and the real optical images separately. H_i , W_i and C_i denote the height, width and channel numbers of input characteristic diagram respectively.

2) DISCRETE COSINE TRANSFORM LOSS

In the light of the high-frequency speckle noise in SAR images, transformations converting images from spatial domain to frequency domain are needed. Traditional methods use Fourier Transform (FT) to accomplish this task, whose imaginary bases shown in Eq. (2) bring imaginary parts in the results of transformation, leading to a lot of calculation troubles. In this paper, we take advantage of Discrete Cosine Transform (DCT) using cosine function as bases to avoid this problem. The differences between FT and DCT are illustrated as below:

$$F(u, v) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) e^{-j2\pi \frac{ux+vy}{N}} \quad (2)$$

$$DCT(u, v) = \frac{2}{N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos \left[\frac{\pi}{N} u \left(x + \frac{1}{2} \right) \right] \times \cos \left[\frac{\pi}{N} v \left(y + \frac{1}{2} \right) \right] \quad (3)$$

where $f(\cdot)$ denotes the input image matrix. N is the width of image matrix. $x, y = 0, 1, \dots, N-1$ are numbers of spatial-domain sequences, and $u, v = 0, 1, \dots, N-1$ are the generalized frequency variables. $F(\cdot)$ and $DCT(\cdot)$ are the result matrices after Fourier Transform and Discrete Cosine Transform respectively.

Because of separability and symmetry in DCT, one-dimensional transforms can be used to construct

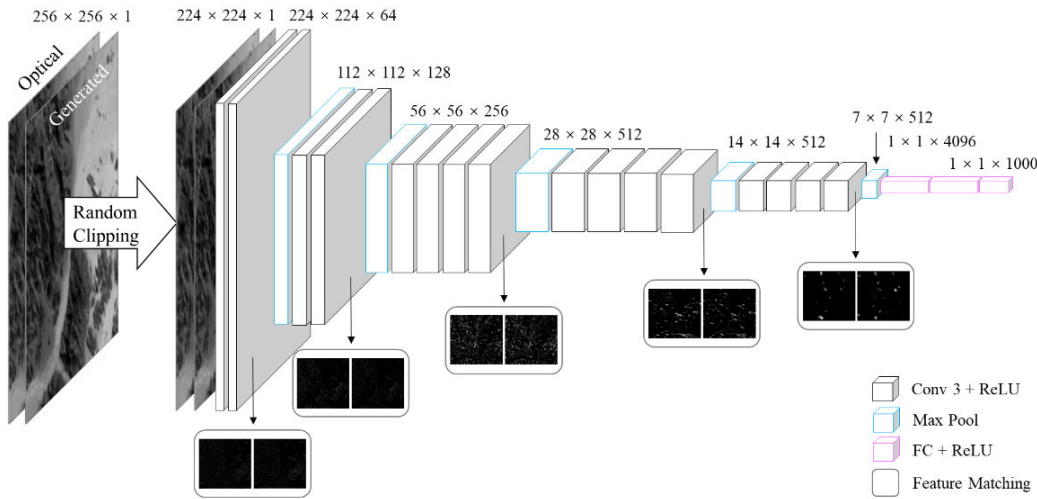


FIGURE 8. The process of feature matching with VGG19 network. Parameters marked above are the height, width and channel number of each layer.

two-dimensional transform when calculating the results. When the input image size is $N \times N$, the transformation can be performed as:

$$DCT(X) = DXT^T \quad (4)$$

$$D(u) = \begin{cases} \frac{1}{\sqrt{N}} \cos \left[\frac{\pi}{N} u(x + \frac{1}{2}) \right], & u = 0, \\ \frac{2}{\sqrt{N}} \cos \left[\frac{\pi}{N} u(x + \frac{1}{2}) \right], & u \neq 0. \end{cases} \quad (5)$$

where D is the weight matrix of one-dimensional DCT, X denotes the input image matrix, and N denotes the width of image matrix.

Through DCT, the imaginary parts in transform results are effectively avoided. At the same time, high-frequency noise is filtered and shallow information is extracted owing to the energy concentration on low frequency band [47]. In the process of designing the DCT loss function, we find that the L2 norm of DCT tends to degenerate into the same gradient as the L2 norm of original images, resulting in image blur mentioned in [18]. As a result, we choose L1 norm of matrix, which is also called Manhattan distance to calculate the distance:

$$L_{DCT}(G, D) = \|DCT(G(X)) - DCT(Y)\|_1 \quad (6)$$

where $G(\cdot)$ is the generator network, X denotes the input SAR images, and Y corresponds to real optical images.

Finally, our loss function is composed of three parts: adversarial loss, VGG loss and DCT loss. Conditional GANs aim to model the conditional distribution of real images given the input labels via the following minimax game [20]:

$$\min_G \max_D L_{GAN}(G, D) \quad (7)$$

where the loss function $L_{GAN}(G, D)$ is given by

$$E_{(X,Y)}[\log D(X, Y)] + E_X[\log(1 - D(X, G(X)))] \quad (8)$$

With the multi-scale discriminators, the learning problem in loss functions becomes a multi-task learning problem:

$$\begin{aligned} \min_G \max_{D_1, D_2} & \left(\sum_{k=1,2} L_{GAN}(G, D_k) \right. \\ & \left. + \lambda_{VGG} \sum_{k=1,2} L_{VGG}(G, D_k) + \lambda_{DCT} \sum_{k=1,2} L_{DCT}(G, D_k) \right) \end{aligned} \quad (9)$$

where G is the generator network, D is the discriminator network, and k refers to the number of the discriminator. L_{VGG} and L_{DCT} denote the VGG loss and DCT loss respectively. λ_{VGG} and λ_{DCT} control the importance of L_{VGG} and L_{DCT} separately.

IV. SAR-TO-OPTICAL IMAGE TRANSLATION

A. IMPLEMENT DETAILS

The method proposed in this paper is implemented on PyTorch and the computing platform is a single RTX TITAN GPU with 24GB GPU memory. The input SAR images are single-channel images with resolution of 256×256 . The batch size is set to 1 since instance normalization, which has been demonstrated to be effective at image generation tasks [56], is used to implement the instance-level transformation. Adam optimizer is used to optimize the network, in which parameters β_1 and β_2 are set to 0.5 and 0.999 respectively. λ_{VGG} and λ_{DCT} corresponding to the weight of loss functions mentioned in Section III-B are set to 10 and 100 respectively. 200 epochs are trained to ensure model convergence. Learning rate stays at a fixed value of 0.0002 for the first 100 epochs, and then linearly decreases to 0 in the remaining epochs.

B. RESULTS

Fig. 9 shows results of SAR-to-optical image translation using our algorithm and proves that our method is feasible

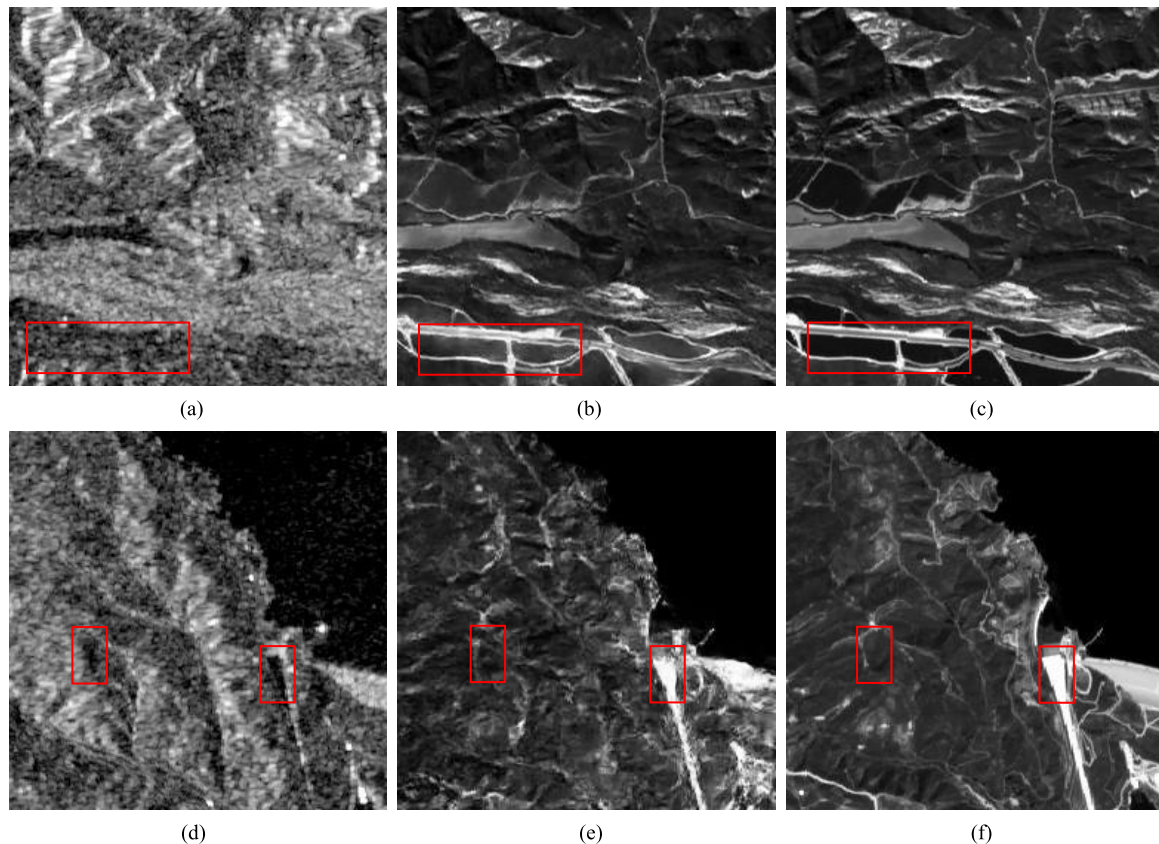


FIGURE 9. Translation results of farmland and bay. (a) (d) SEN-1 SAR images. (b) (e) Generated optical images. (c) (f) SEN-2 optical images.

in SAR-to-optical image translation field. SAR images have been wasted a lot because of difficulties for human beings to identify. But images generated by our method can be recognized easily. For example, in red boxes of the first line, the SAR image corresponds to a fuzzy area, untrained people can not make out the terrain type in it. However, from the generated optical image, it is shown obviously that this is a large area of farmland, with a clear boundary in the middle, which can not be found at all in the corresponding place of SAR image. In two red boxes of the second line, SAR image shows two areas that look very similar in shape and structure. It is hard for untrained people to identify and distinguish them. But in the generated optical image, the red box on the left is a beach, and the one on the right is a crowded street. Through comparison between generated images and real optical images, it can be found that images generated by our method are very similar to real optical images in terms of structural features and texture information. Quantitative analysis will be presented in the next part.

C. COMPARISON WITH DIFFERENT METHODS

Our method is compared with three famous methods in the field of image-to-image translation: pix2pix [18], CycleGAN [21] and pix2pixHD [20]. Training modes are all from scratch. For the sake of fairness, we remove a pair of generator

and discriminator of CycleGAN, and other settings remain unchanged. In the parameter setting of pix2pixHD, the number of label channel is 0.

Firstly, it can be seen from Figure 10 that pix2pix is unable to show correct details, and even has a serious chessboard effect at the edge of images. For example, the vein trend of snow mountain and ridge in red boxes is not shown in pix2pix. Detailed information is very important in the process of earth observation, so it is a fatal disadvantage. CycleGAN also has problems. Because the speckle noise in input SAR images is very serious, there are obvious noises in images generated by CycleGAN, such as mountain area in green boxes in Figure 10. Our experiment results outperform that of these methods including pix2pixHD in both texture feature preservation and high-frequency noise reduction, which benefit from the algorithm we designed.

Secondly, Table 4 shows Image Quality Assessment (IQA) results of four methods. IQA consists of SSIM, FSIM, PSNR and MSE. SSIM calculates the structure similarity between generated images and real optical images, which represents the structural information. Its value is between 0 and 1, and will reach 1 when two images are identical. FSIM [57] proposed based on SSIM is a feature similarity index. It combines phase congruence information and gradient magnitude to characterize the image local quality. PSNR, which denotes

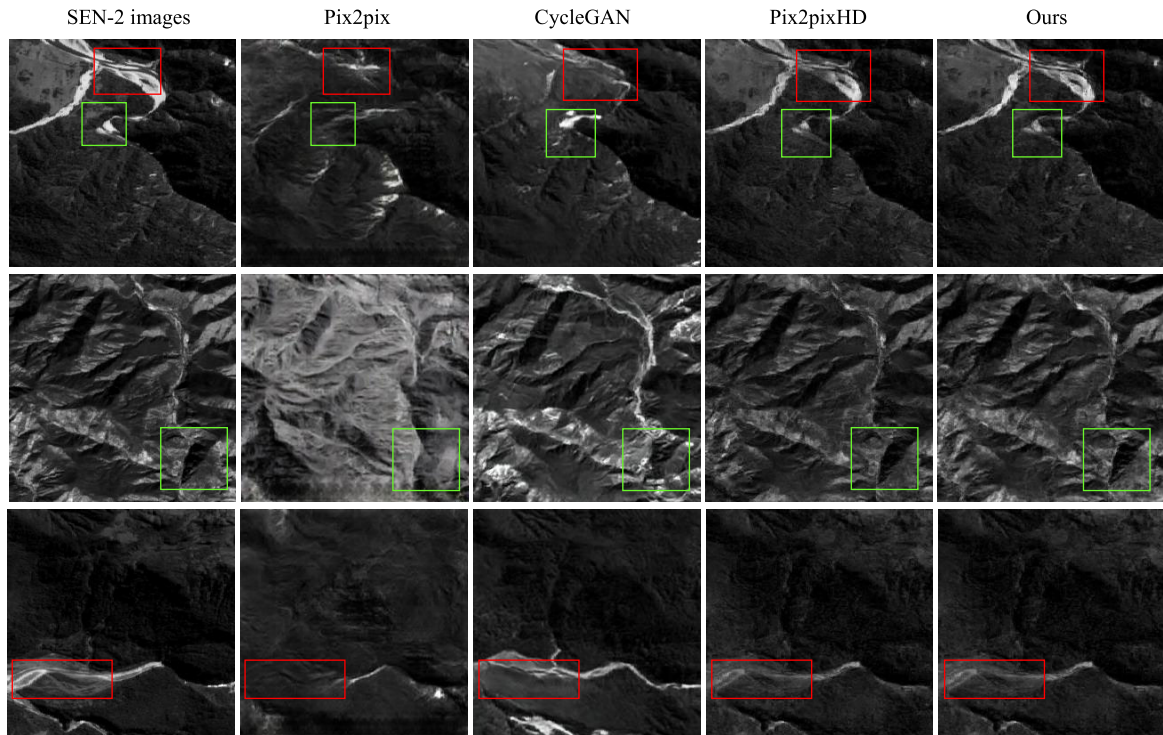


FIGURE 10. Translation results of different methods. The size of all images is 256×256 . The images from left to right are SEN-2 optical remote sensing images, Pix2pix, CycleGAN, Pix2pixHD, and Our method.

TABLE 4. IQA results of different methods.

Method	SSIM	FSIM	PSNR	MSE
Pix2pix	0.33	0.71	18.36	0.02
CycleGAN	0.37	0.72	18.24	0.02
Pix2pix HD	0.54	0.74	20.42	0.01
Ours	0.61	0.79	21.57	0.01

the peak signal-to-noise ratio of generated images, and MSE, which is the mean square error closely related to PSNR, both suggest the image pixel statistical information. These indexes are commonly used in the field of image quality evaluation. From Table 4, it is demonstrated that our experiment obtains the optimal value in all indexes. Meanwhile, the scores of feature similarity are greatly improved compared with structure similarity, for the reason that the weight of low-level features in FSIM is higher than SSIM, while differences between generated images and real optical images are mainly concentrated on the field of high-level.

Finally, it can be concluded that in SAR-to-optical image translation task, pix2pix needs to be improved in acquisition and generation of detailed information, while CycleGAN based on ResNet is too vulnerable to high-frequency noise although it can produce rich texture. Pix2pixHD is better than the two, but not as good as ours. Our method not only pays attention to detailed texture, but also effectively suppresses

high-frequency noise, making it possible to generate optical images very close to the real ones.

D. SUPERIORITY OF ALGORITHM

In our method, more than one loss function is used, each of which plays a unique role. In order to prove the correctness of our choice, we did ablation study to show effects of different loss functions.

As is described in Section III-B, loss functions correspond to different properties. For example, due to the large amount of high-frequency speckle noise in SAR images, DCT loss aims at suppressing the noise components of SAR images so as to facilitate subsequent processing. Furthermore, VGG is selected as the feature extractor of image translation task based on the results in Section III-A. The combination of multiple loss functions is very effective. When a loss is removed, its influence can be obviously displayed in the translation process. Fig. 11 shows two sets of experimental results. The first column is the original generated images, the second column removes DCT loss, and the third column removes VGG loss. It can be found from red boxes that, because the multi-scale features in images are extracted by VGG, after VGG loss being removed, the texture features are no longer clear, which shows that VGG loss function contributes the most to the texture preservation of images. Meanwhile, it can be seen from green boxes that, after the loss based on DCT being removed, the speckle noise is not well disposed, and there are serious artifacts in the generated optical image,

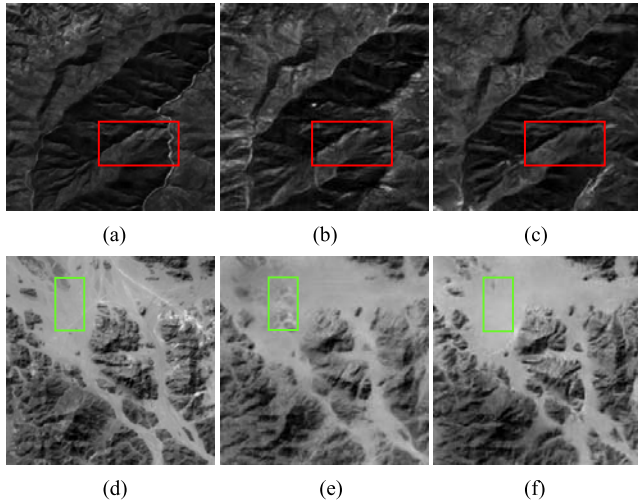


FIGURE 11. Ablation study to show effect of different losses. (a)(d) Generated optical images. (b)(e) Generated optical images using method without DCT loss. (c)(f) Generated optical images using method without VGG loss.

TABLE 5. IQA results of our method with or without different losses.

Method	SSIM	FSIM	PSNR	MSE
Ours	0.61	0.79	21.57	0.01
Ours (w/o DCT loss)	0.53	0.73	20.07	0.01
Ours (w/o VGG loss)	0.46	0.72	20.22	0.01

TABLE 6. IQA results of different models with or without our loss function.

Method	SSIM	FSIM	PSNR	MSE
Pix2pix	0.33	0.71	18.36	0.02
Pix2pix (w/ our loss)	0.53	0.74	20.60	0.01
CycleGAN	0.37	0.72	18.24	0.02
CycleGAN (w/ our loss)	0.38	0.72	18.56	0.02

which reflects the effects of DCT loss function on high-frequency noise reduction. Table 5 also indicates that without VGG loss, the structure similarity and the feature similarity decrease most significantly, and without DCT loss, the peak signal-to-noise ratio decreases most obviously, which shows that VGG loss plays the most important role on image feature preservation, while DCT loss makes a huge contribution to noise filtering. The rationality and effectiveness of our loss function can be proved by this phenomenon.

In order to better demonstrate the superiority of our algorithm, we add loss functions to different models for comparison. Table 6 shows IQA indexes of pix2pix [18] and CycleGAN [21] before and after adding our loss functions. It is demonstrated that IQA results of two models are both improved in the experiment, especially pix2pix, whose SSIM increases from 0.33 to 0.53, FSIM grows from 0.71 to 0.74, and PSNR rises from 18.36 to 20.60. Because there are as

many as 54M parameters in pix2pix generator model, its improvement is more obvious than that of CycleGAN with only 11M parameters.

V. CONCLUSIONS AND OUTLOOKS

In this paper, a novel feature-guided method is proposed to translate SAR images to optical images. The generated images are very close to the real optical images and can be easily recognized by human eyes. Looked upon the results of the experiments, the images generated dependent on multi-scale discriminator and multi-level feature matching better preserve the original features, and the measurement of loss function based on Discrete Cosine Transform guarantees the suppression of high-frequency noise.

It is worth mentioning that the feature guidance of our work performs better on natural scenes such as farmland, mountains and rivers, but worse on anthropic zone with complex features. Because the edge and line features in natural scenes are more obvious, while the buildings and walls in large cities may be blurred or merged into larger elements [10]. For the same reason, we only focus on the larger element of forest, lack of analysis of individual trees. However, the generated images can be interpreted from the overall categories and more obvious details through cross-modal transformation. Not only can they be used as alternative data when optical images are deficient, but also avoid the waste of SAR image resources.

Future work will be concentrated on finding the appropriate optimization strategies to recover larger elements, and exploring new network framework and loss functions to achieve better translation quality. Meanwhile, the benefits of generated images to feature extraction and object detection will be further explored.

REFERENCES

- [1] M. J. Delgado-Aguilar, F. E. Fassnacht, M. Peralvo, C. P. Gross, and C. B. Schmitt, "Potential of TerraSAR-X and sentinel 1 imagery to map deforested areas and derive degradation status in complex rain forests of Ecuador," *Int. Forestry Rev.*, vol. 19, no. 1, pp. 102–118, Mar. 2017.
- [2] T. Kraus, M. Schmidt, S. W. Dech, and C. Samimi, "The potential of optical high resolution data for the assessment of leaf area index in east african rainforest ecosystems," *Int. J. Remote Sens.*, vol. 30, no. 19, pp. 5039–5059, Sep. 2009.
- [3] J. Santos, "Airborne P-band SAR applied to the aboveground biomass studies in the Brazilian tropical rainforest," *Remote Sens. Environ.*, vol. 87, no. 4, pp. 482–493, Nov. 2003.
- [4] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
- [5] C. O. Dumitru, S. Cui, D. Faur, and M. Datcu, "Data analytics for rapid mapping: Case study of a flooding event in Germany and the tsunami in Japan using very high resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 1, pp. 114–129, Jan. 2015.
- [6] L. Wang, X. Xu, Y. Yu, R. Yang, R. Gui, Z. Xu, and F. Pu, "SAR-to-Optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129136–129149, 2019.
- [7] J. C. Ni and Y. L. Xu, "SAR automatic target recognition based on a visual cortical system," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, vol. 2, Dec. 2013, pp. 778–782.
- [8] J.-S. Lee, "Speckle suppression and analysis for synthetic aperture radar images," *Opt. Eng.*, vol. 25, no. 5, 1986, Art. no. 255636.

- [9] M. Simard, G. DeGrandi, K. P. B. Thomson, and G. B. Benie, "Analysis of speckle noise contribution on wavelet decomposition of SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 6, pp. 1953–1962, 1998.
- [10] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, 2019.
- [11] S. Auer, S. Hinz, and R. Bamler, "Ray-tracing simulation techniques for understanding high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1445–1456, Mar. 2010.
- [12] F. Argenti, A. Lapini, T. Bianchi, and L. Alparone, "A tutorial on speckle reduction in synthetic aperture radar images," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 3, pp. 6–35, Sep. 2013.
- [13] Z. Shen, M. Huang, J. Shi, X. Xue, and T. S. Huang, "Towards instance-level image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3683–3692.
- [14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [15] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.
- [16] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for Image-to-Image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.
- [17] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.
- [18] 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.
- [19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: <http://arxiv.org/abs/1411.1784>
- [20] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
- [21] 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. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [22] S. Fu, F. Xu, and Y.-Q. Jin, "Translating SAR to optical images for assisted interpretation," 2019, *arXiv:1901.03749*. [Online]. Available: <http://arxiv.org/abs/1901.03749>
- [23] H. Toriya, A. Dewan, and I. Kitahara, "SAR2OPT: Image alignment between multi-modal images using generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 923–926.
- [24] K. Enomoto, K. Sakurada, W. Wang, N. Kawaguchi, M. Matsuoka, and R. Nakamura, "Image translation between SAR and optical imagery with generative adversarial nets," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1752–1755.
- [25] A. Ley, O. Dhondt, S. Valade, R. Haensch, and O. Hellwich, "Exploiting GAN-based SAR to optical image transcoding for improved classification via deep learning," in *Proc. 12th Eur. Conf. Synth. Aperture Radar (EUSAR)*. Frankfurt, Germany: VDE, Jun. 2018, pp. 1–6.
- [26] J. A. Richards and J. A. Richards, *Remote Sensing Digital Image Analysis*, vol. 3. Berlin, Germany: Springer, 1999, pp. 10–38.
- [27] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH image to ImageJ: 25 years of image analysis," *Nature Methods*, vol. 9, no. 7, pp. 671–675, 2012.
- [28] P. Soille, *Morphological Image Analysis: Principles and Applications*. Springer, 2013.
- [29] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, "Feature extraction, foundations and applications," *Stud. Fuzziness Soft Comput.*, vol. 205, no. 12, pp. 68–84, 2006.
- [30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [33] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
- [34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [37] B. Praveen and V. Menon, "Novel deep-learning-based spatial-spectral feature extraction for hyperspectral remote sensing applications," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 5444–5452.
- [38] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
- [39] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-resolution SAR image classification via deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2351–2355, Nov. 2015.
- [40] S. Hijazi, R. Kumar, and C. Rowen, *Using Convolutional Neural Networks for Image Recognition*. San Jose, CA, USA: Cadence Design Systems, 2015, pp. 1–12.
- [41] R. P. Perry, R. C. Dipietro, and R. L. Fante, "SAR imaging of moving targets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 35, no. 1, pp. 188–200, Jan. 1999.
- [42] Y. Ding and D. C. J. Munson, "A fast back-projection algorithm for bistatic SAR imaging," in *Proc. Int. Conf. Image Process.*, vol. 2, Sep. 2002, p. 2.
- [43] S. J. Wei, X. L. Zhang, J. Shi, and G. Xiang, "Sparse reconstruction for SAR imaging based on compressed sensing," *Prog. Electromagn. Res.*, vol. 109, pp. 63–81, 2010.
- [44] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [45] R. N. Bracewell and R. N. Bracewell, *The Fourier Transform and its Applications*, vol. 31999. New York, NY, USA: McGraw-Hill, 1986.
- [46] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. 100, no. 1, pp. 90–93, Jan. 1974.
- [47] A. B. Watson, "Image compression using the discrete cosine transform," *Math. J.*, vol. 4, no. 1, p. 81, 1994.
- [48] M. Schmitt, L. Haydn Hughes, and X. Xiang Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," 2018, *arXiv:1807.01569*. [Online]. Available: <http://arxiv.org/abs/1807.01569>
- [49] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 649–666.
- [50] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing: Analysis and Machine Vision*. Boston, MA, USA: Cengage, 2014.
- [51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.
- [52] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
- [53] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: <http://arxiv.org/abs/1511.06434>
- [54] J. Y. Zhu, R. Zhang, D. Pathak, D. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.
- [55] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 694–711.
- [56] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: <http://arxiv.org/abs/1607.08022>
- [57] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.



JIEXIN ZHANG received the B.Eng. degree in electronics engineering from Tianjin University, Tianjin, China, in 2018. She is currently pursuing the M.S. degree with the Nanjing University of Aeronautics and Astronautics.

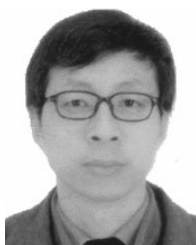
Her research interests include radar imaging, deep learning, and remote sensing imagery processing.



XIWEN LU received the B.Eng. degree in electronics engineering from the South China University of Technology, Guangzhou, China, in 2017.

He is currently a Graduate Researcher with the Nanjing University of Aeronautics and Astronautics. His research interests include machine learning and computer vision.

...



JIANJIANG ZHOU (Member, IEEE) received the M.S. and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics (NUAA), in 1988 and 2001, respectively.

He is currently a Professor and the Director of the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, NUAA. His main research interests include aircraft radio frequency stealth, radar signal processing, and array signal processing.