INR-ECGAN: An Enhanced Conditional GAN with Implicit Neural Representation for SAR-to-Optical Image Translation

1st Chenguo Feng†
School of Robotics
Hunan University
Changsha, China
chenguofeng@hnu.edu.cn

2nd Yang Liu†
School of Robotics
Hunan University
Changsha, China
ly758963@hnu.edu.cn

3rd Nan Wang
School of Robotics
Hunan University
Changsha, China
wangnanp@hnu.edu.cn

4th Zhiyang Chen
School of Robotics
Hunan University
Changsha, China
15398097065@163.com

5th Xiaohui Wei
College of Information Science and Engineering
Hunan Normal University
Changsha, China
xh_wei@hnu.edu.cn

6th Haibo Liu*
School of Robotics
Hunan University
Changsha, China
haiboliu@hnu.edu.cn

***Abstract*—The SAR-to-Optical image translation task helps solve the interpretability of SAR images and has attracted much attention from researchers in recent years. Although existing methods based on conditional generative adversarial network (CGAN) have shown appealing potential for this task, the optical images generated by them still suffer from blurred details and color distortion. To this end, a novel enhanced CGAN with implicit neural representation (INR-ECGAN) is proposed in this paper. Specifically, the proposed method utilizes convolutional residual blocks to encode the semantic features of SAR images at multiple scales. Then, a pyramid pooling module is employed to mine the global information of deep semantic features. Subsequently, an implicit decoder is designed to fuse the multi-scale semantic and global features and estimate a continuous function mapping from the fused features to the optical images. Finally, a discriminator that can simultaneously distinguish the generated and real optical images at multiple scales is adopted to improve the translation performance. Extensive experiments on public datasets have demonstrated the superiority of INR-ECGAN. The code will be available at https://github.com/lyltbafw/INR-ECGAN.**

***Index Terms*—SAR-to-Optical image translation, Conditional generative adversarial network, Implicit neural representation.**

I. Introduction

Synthetic aperture radar (SAR) satellites can observe ground targets under all-day and all-weather conditions. Therefore, images obtained by SAR are widely used in various fields, such as object detection [17], semantic segmentation [2], and multi-model image fusion [4]. However, affected by strong speckle noise, it is extremely challenging for people to understand the semantic information of SAR images. SAR-to-Optical image translation (S2OIT), which converts SAR images into optical images, helps improve the interpretability of SAR images and has been widely studied in recent years.

† Authors contributed equally
∗ Corresponding author

Considering the great achievements of conditional generative adversarial network (CGAN) [9] in the S2OIT task, numerous CGAN-based methods have been proposed one after another [24]. They can usually be divided into two categories: unpaired and paired. Unpaired methods allow the SAR and optical images used for training to be mismatched. As the most representative unpaired method, CycleGAN [25] uses a bidirectional loop structure to learn the mapping relationship between the source domain and the target domain, and has been successfully applied in natural image translation. Since then, different variants based on CycleGAN have been designed for the S2OIT task. For example, Wang *et al.* [14] introduced ℓ_1 loss between pixels based on CyclGAN and proposed the supervised CycleGAN for S2OIT. Work [19] combined the dense and residual connections in the generator and employed different normalization layers for image features during the translation process. Lee *et al.* [7] presented a novel coarse-to-fine CGAN, where SAR-to-infrared and SAR-to-visible image translation were jointly trained in the network. Different from unpaired methods, paired methods require that the SAR and optical images are matched during the training process, which enables them to better learn the potential correlations between SAR and optical images. As a universal paired translation framework, Pix2Pix [6] adopted U-Net as a generator and designs a conditional discriminator to distinguish the authenticity of generated images. Given the simple structure and excellent performance of Pix2Pix, many paired approaches for S2OIT have been improved based on it. For example, Zhan *et al.* [21] proposed an improved CGAN for S2OIT by utilizing the style calibration module and multi-scale discrimination strategy. Yang *et al.* [20] applied a CGAN-based framework with parallel branches to combine the low-level and high-level SAR features. By introducing the vision

2024 China Automation Congress (CAC) | 979-8-3503-6860-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CAC63892.2024.10864577

transformer, Wang *et al.* [16] presented a hybrid CGAN to fuse the category information and semantic features of SAR Images. Wang *et al.* [13] constructed two CGANs to learn the latent space of RGB images and recover RGB images from the corresponding SAR features, respectively. However, the results generated by existing paired methods still suffer from blurred local textures and color distortion. The reason is that they do not fully exploit and utilize the multi-scale semantic information of SAR images. Besides, these methods typically recover the expected optical image from semantic features directly by discrete decoding operations, leading to the loss of detailed information.

Implicit neural representation (INR) is a novel signal representation method. It maps coordinates continuously to the target signal domain through the multi-layer perceptron (MLP). Due to its excellent feature expression ability, INR has been widely used in various visual tasks. For instance, Chen *et al.* [1] explored the potential of implicit function in expressing image features continuously. Shen *et al.* [11] proposed a general continuous refinement model, where semantic segmentation images were upsampled continuously based on the corresponding position encoding. Gao *et al.* [3] designed an implicit diffusion model to continuously denoise and reconstruct high-resolution images. Yang *et al.* [18] first presented an unsupervised adversarial learning model based on INR in low-light image enhancement. Overall, as a continuous image representation idea, INR can more effectively preserve the texture and color features of images.

Motivated by the above works, a novel enhanced CGAN with implicit neural representation, called INR-ECGAN, is proposed for the S2OIT task in this paper. Specifically, its generator first utilizes four simple convolutional residual blocks (CRBs) to encode the semantic features of SAR images at multiple scales. Then, a pyramid pooling module (PPM) is employed to extract the global information of deep semantic features. Subsequently, the implicit decoder presents a multi-scale feature integrator (MSFI) to finely integrate the multi-scale semantic and global features, and then employs an implicit neural representation module (INRM) to learn a continuous function mapping from the fused features to the optical images. Finally, a multi-scale discriminator is adopted to simultaneously distinguish the generated and real optical images at multiple scales.

The major contributions of this paper can be summarized as:

- We propose a novel framework called INR-ECGAN, which can effectively mine and fuse the multi-scale semantic and global features of SAR images to generate realistic optical images.
- To the best of our knowledge, we are the first attempt to establish a continuous mapping directly from SAR features to optical images by exploiting implicit neural representation, thereby compensating for the information lost by discrete operations such as sampling.
- Experimental results on two public datasets demonstrate that INR-ECGAN is superior to the existing methods.

II. Proposed Method

A. Overview

The basic framework of the proposed INR-ECGAN is shown in Fig. 1 (a). Specifically, the generator first applies an encoder to extract the multi-scale semantic features $f_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^{i-1}C}(i = 1,2,3,4)$ from the input SAR image $X \in \mathbb{R}^{1\times H\times W}$, where H and W represent the height and width of the image, respectively, and C denotes the number of feature channels. Then, a PPM is utilized to obtain the global contextual features f_g of deep semantics f_4. Subsequently, an implicit decoder is designed to continuously reconstruct the target optical image $G(X) \in \mathbb{R}^{3\times H\times W}$ from the multi-scale features $\{f_i\}_{i=1}^{3}$ and global features f_g. Finally, a multi-scale discriminator is adopted to distinguish the authenticity of generated $G(X)$ at different scales.

B. Encoder

The encoder includes a shallow feature extractor and four CRBs. The shallow feature extractor uses a 7×7 convolutional layer to capture the detailed features from the input image. CRB is a simple and effective module for feature extraction and denoising [22]. As shown in Fig. 1 (b), each CRB consists of two residual blocks, where the first block is used to downsample the input image, and the second block extracts the semantic features. f_{in} and f_{out} represent the input and output features of CRB, respectively. By cascading four CRBs, the encoder can learn latent semantic features of SAR images at different scales.

C. PPM

The PPM [23] is directly employed in the generator to fully mine the global contextual information of SAR. Specifically, the PPM first performs global average pooling and convolution operations on the deep features f_4 at different scales. Subsequently, the obtained multi-scale features are all upsampled to the same size of f_4 by bilinear interpolation. Furthermore, the pooling features are concatenated with f_4 along the channel dimension. Finally, a 3×3 convolutional layer is used to fuse the concatenated features and produce the global semantic features $f_g \in \mathbb{R}^{4C\times\frac{H}{16}\times\frac{W}{16}}$.

D. Implicit Decoder

1) MSFI: To effectively utilize the multi-scale semantic information from the encoder, MSFI is designed to align and fuse the global features f_g and multi-scale features $\{f_i\}_{i=1}^{3}$. As is shown in Fig. 1 (c), it contains three feature fusion branches for different scales. Each branch uses a 1×1 convolutional layer to achieve channel alignment of input features f_i and f_g. The aligned features are sequentially added to the upsampled f_g, and the added features are further fused through a 3×3 convolutional layer. Later, the input f_g and the output features of the three branches are upsampled to the same size and then concatenated along the channel dimension. Finally, a 3×3 convolutional layer is employed to integrate the concatenated features for the refined features $F_I \in \mathbb{R}^{4C\times\frac{H}{2}\times\frac{W}{2}}$.

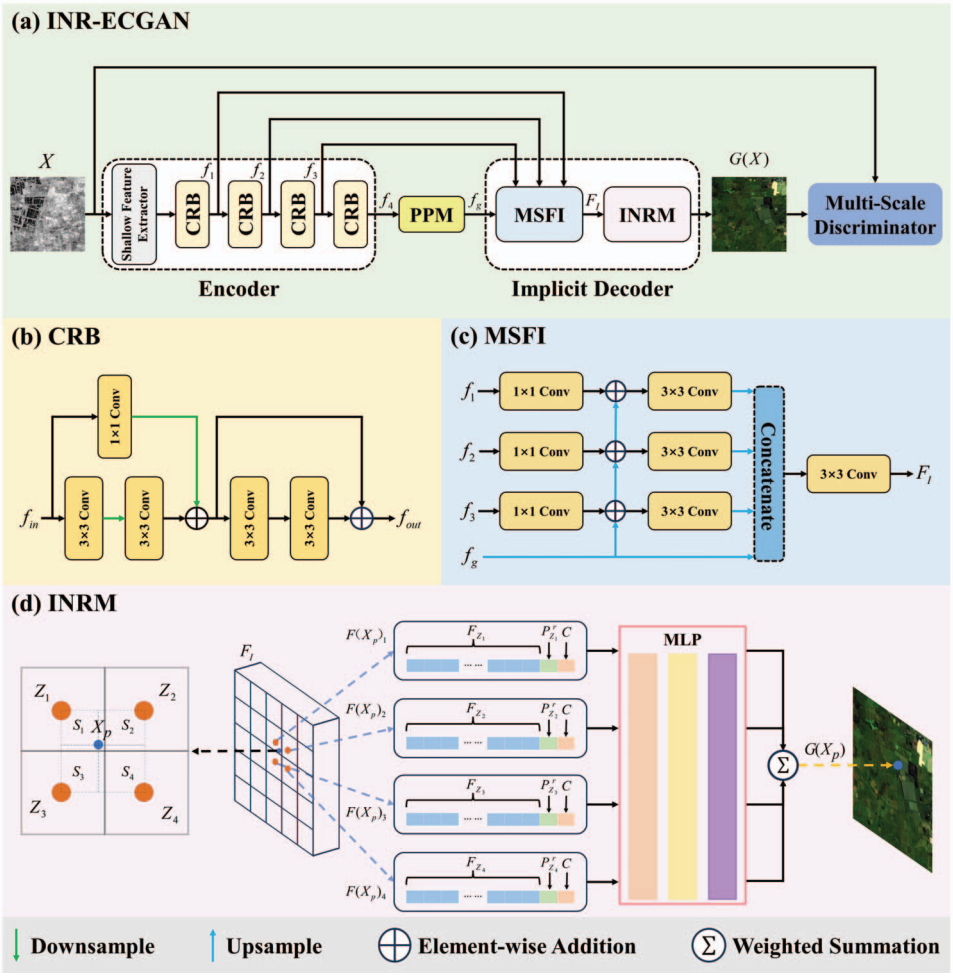


Fig. 1. The overall architecture of proposed INR-ECGAN.

2) INRM: Fig. 1 (d) shows the structural diagram of INRM. It learns the implicit transformation from SAR features to target image. In particular, each predicted pixel point X_p is continuously described by four surrounding local correlation points $Z_i (i = 1, 2, 3, 4)$. To unify the changes in feature size, INRM first maps the normalized coordinates of Z_i to the target domain. Subsequently, based on nearest neighbor interpolation, $F_{Z_i} \in \mathbb{R}^{4C\times 1}$ is used as the input features for target point X_p, where F_{Z_i} represents the features of Z_i. Then, to ensure continuity in feature expression, INRM introduces corresponding relative position encoding $P^r_{Z_i} \in \mathbb{R}^{2\times 1}$ into all input features. $P^r_{Z_i}$ is obtained by calculating the position offset between Z_i and X_p. In addition, to refine the prediction results, the pixel size of target image is also added into F_{Z_i}. Therefore, the local feature vector of target point can be represented as:

$$F(X_p)_i = \left(F_{Z_i}, P^r_{Z_i}, C\right), i = 1, 2, 3, 4 \tag{1}$$

where $F(X_p)_i \in \mathbb{R}^{(4C+4)\times 1}$ means i^{th} local feature vector of X_p, $C \in \mathbb{R}^{2\times 1}$ indicates the pixel size of target resolution.

After that, $\{F(X_p)_i\}_{i=1}^4$ are processed by the same MLP to obtain four predicted values, respectively. To fully utilize local correlation information, the weighted average of four predictions is computed to achieve a continuous translation. The prediction process of INRM can be represented as:

$$G\left(X_p\right) = \sum_{i\in(1,2,3,4)} \frac{S_{5-i}}{S} MLP\left(F(X_p)_i\right) \tag{2}$$

$$S = \sum_{i\in(1,2,3,4)} S_i \tag{3}$$

where $G(X_p)$ represents the final pixel value of X_p. S_i means the area of rectangular regions between X_p and Z_i, and S_{5-i} denotes the diagonal area of S_i.

E. Multi-scale Discriminator

INR-ECGAN adopts the discriminator strategy of Pix2PixHD [15]. The input SAR and generated optical images are first concatenated in the channel dimension. Then, three pooling layers are used to sequentially downsample concatenated images. Finally, three discriminators $\{D_i\}_{i=1}^3$

with the same structure are employed to distinguish the authenticity of images at different scales.

F. Loss Functions

The loss functions of INR-ECGAN include generative adversarial loss, perceptual loss, and feature matching loss. Different from Pix2Pix [6], the ℓ_1 loss between the generated image and the corresponding ground truth is not used in the proposed method. There are two reasons for this. First, although the ℓ_1 loss helps protect edge information, it is still easy to cause image blurring. Second, due to different imaging resolutions, there may be some spatial offset between paired SAR and optical images. Instead, the perceptual loss and feature matching loss are applied to constrain the generated images in the feature dimension.

The generative adversarial loss L_{GD} of multi-scale discriminator in INR-ECGAN can be expressed as:

$$\mathcal{L}_{\mathrm{GD}} = \min_{G} \max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\mathrm{GAN}}\left(G, D_k\right) \tag{4}$$

$$\begin{aligned} \min_{G} \max_{D_k} \mathcal{L}_{\mathrm{GAN}} = & \mathbb{E}_{Y \sim p_{\mathrm{optical}}(Y)}\left[\log D_k(Y, X)\right] + \\ & \mathbb{E}_{X \sim p_{\mathrm{sar}}(X)}\left[\log\left(1 - D_k\left(G(X), X\right)\right)\right] \end{aligned} \tag{5}$$

where *X*, *G(X)*, and *Y* indicate the input SAR image, the generated optical image, and the corresponding ground truth, respectively.

Perceptual loss $\mathcal{L}_{\mathrm{Per}}$ uses a pre-trained VGG-19 model to calculate the similarity between the generated image and the corresponding ground truth. Its computation process is defined as:

$$\mathcal{L}_{Per} = \mathbb{E}_{X \sim p_{\mathrm{sar}}(X)} \sum_{i=1}^{5} \frac{1}{Si} \left\| \phi_i\left(G(X)\right) - \phi_i(Y) \right\|_1 \tag{6}$$

where ϕ_i means the output features of i^{th} layer of VGG-19, and S_i denotes the size of ϕ_i.

Moreover, to reconstruct more realistic results, the feature matching loss $\mathcal{L}_{\mathrm{FM}}$ is introduced into INR-ECGAN. For the multi-scale discriminator, its $\mathcal{L}_{\mathrm{FM}}$ can be represented as:

$$\mathcal{L}_{\mathrm{FM}} = \sum_{k=1}^{3} \mathcal{L}_{\mathrm{FM}}\left(G, D_k\right) \tag{7}$$

$$\begin{aligned} & \mathcal{L}_{\mathrm{FM}}\left(G, D_k\right) = \\ & \mathbb{E}_{(X,Y) \sim p_{\mathrm{data}}(X,Y)} \sum_{i=1}^{T} \frac{1}{N_i} \left[\left\| D_k^{(i)}(Y, X) - D_k^{(i)}(G(X), X) \right\|_1 \right] \end{aligned} \tag{8}$$

where D_k means k^{th} discriminator, and $D_k^{(i)}$ denotes the output of i^{th} layer in D_k. T and N_i represent the total number of layers and the number of pixels in each layer, respectively.

Thus, the total loss of INR-ECGAN can be described as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{GD}} + \omega_1 \mathcal{L}_{\mathrm{Per}} + \omega_2 \mathcal{L}_{\mathrm{FM}} \tag{9}$$

where ω_1 and ω_2 represent the weights of $\mathcal{L}_{\mathrm{Per}}$ and $\mathcal{L}_{\mathrm{FM}}$, respectively.

III. Experiment

A. Experimental Settings

1) Datasets: Adequate experiments are conducted on two public datasets, including SEN1-2 dataset [10] and QXS-SAROPT dataset [5]. SEN1-2 dataset includes 282384 pairs of matched SAR and optical images, which covers different land regions around the world in four seasons. 15902 and 3872 image pairs are randomly selected for training and testing from it, respectively. QXS-SAROPT dataset contains 20000 pairs of matched SAR and optical images, which covers San Diego, Shanghai, and Qingdao. 16000 and 4000 image pairs are randomly chosen for training and testing from it, respectively.

2) Compared Methods: To demonstrate the superiority of proposed method, five existing popular methods in S2OIT are compared with INR-ECGAN, including Pix2Pix [6], CycleGAN [25], Pix2PixHD [15], UVCGAN [12], and ParallelGAN [13].

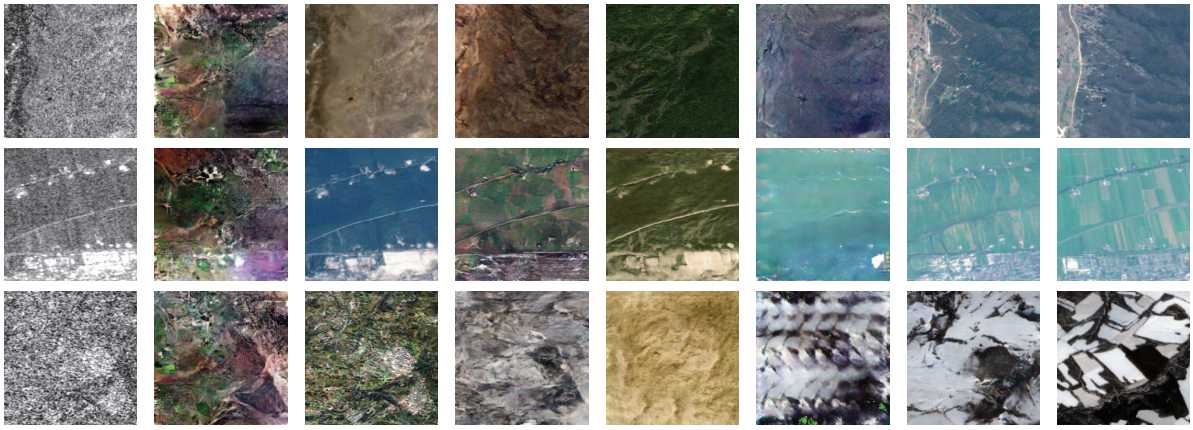
3) Quality Assessment Indicators: To measure the accuracy of generated images, four image quality assessment indicators are used, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), Frchet inception distance (FID), and learned perceptual image patch similarity (LPIPS).

B. Training Details

In our experimental setup, the number of channels C in the proposed generator is set to 128. The adopted PPM uses four pooling scales, which are 1, 2, 3, and 6. For both datasets, the number of training epochs and batch size are set to 200 and 8, respectively. Then, the initial learning rate of the generator and discriminator is set to 0.0002. Particularly, the learning rate remains unchanged in the first 100 epochs and gradually decreases to 0 in the last 100 epochs. For the loss function, ω_1 and ω_2 are both set to 10. Besides, to ensure a fair comparison, the parameters of all compared approaches are set according to the recommendations in the corresponding references. All experiments are performed on a GeForce RTX 3090 GPU.

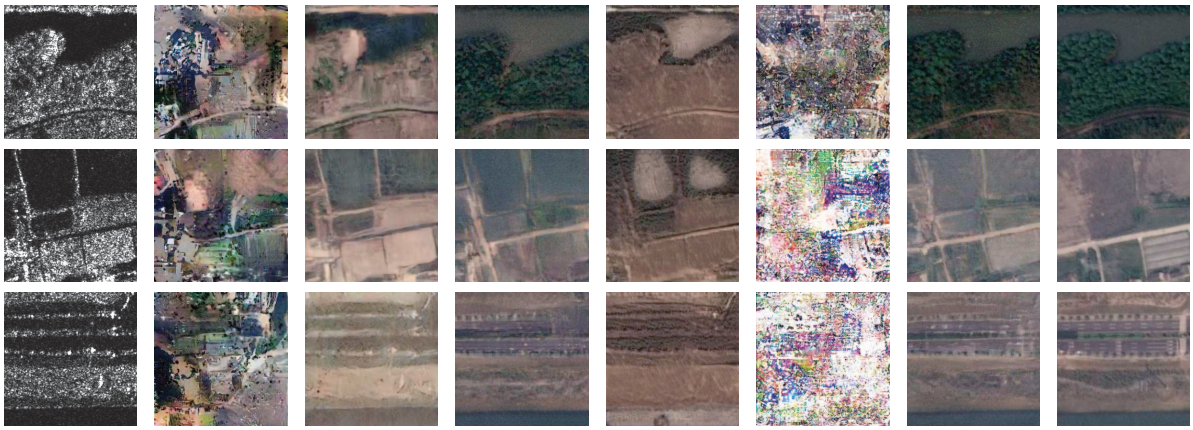
C. Experimental Results

Table I shows the calculation results of evaluation metrics for different compared methods on SEN1-2 and QXS-SAROPT datasets, where the optimal and suboptimal values of each metric have been bolded in red and blue font, respectively. It can be inferred from Table I that the proposed method outperforms other existing methods in almost all metrics. Specifically, for SEN1-2 dataset, INR-ECGAN performs much better than other compared methods in terms of PSNR, SSIM, and LPIPS. Although the FID of CycleGAN is the lowest, INR-ECGAN is far superior to it in other indicators. For QXS-SAROPT dataset, INR-ECGAN is optimal on all evaluation metrics. Besides, to facilitate visual comparison, several RGB images generated by all compared methods on two datasets are represented in Fig. 2 and Fig. 3, respectively. Obviously, in contrast to other methods, the images restored by INR-ECGAN have clearer textures and more accurate color information, which further proves its effectiveness.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Generated results of different compared methods on SEN1-2 dataset. (a) SAR. (b) Pix2Pix [6]. (c) CycleGAN [25]. (d) Pix2PixHD [15]. (e) UVCGAN [12]. (f) Parallel-GAN [13]. (g) INR-ECGAN. (h) Ground Truth.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Generated results of different compared methods on QXS-SAROPT dataset. (a) SAR. (b) Pix2Pix [6]. (c) CycleGAN [25]. (d) Pix2PixHD [15]. (e) UVCGAN [12]. (f) Parallel-GAN [13]. (g) INR-ECGAN. (h) Ground Truth.

TABLE I. Quality evaluation results on SEN1-2 and QXS-SAROPT datasets

Datasets	Method	PSNR↑	FID↓	SSIM↑	LPIPS↓
SEN1-2 [10]	Pix2Pix [6]	12.94	171.75	0.16	0.66
	CycleGAN [25]	11.73	**58.17**	0.16	0.62
	Pix2PixHD [15]	14.35	112.01	**0.23**	**0.56**
	UVCGAN [12]	12.44	148.15	0.16	0.72
	Parallel-GAN [13]	**14.70**	91.45	0.22	0.60
	INR-ECGAN	**16.20**	**70.01**	**0.30**	**0.46**
QXS-SAROPT [5]	Pix2Pix [6]	13.19	161.40	0.18	0.67
	CycleGAN [25]	14.16	105.15	0.29	0.63
	Pix2PixHD [15]	**17.21**	**81.94**	**0.38**	**0.47**
	UVCGAN [12]	15.28	164.29	0.29	0.66
	Parallel-GAN [13]	11.46	203.06	0.10	0.80
	INR-ECGAN	**17.76**	**64.86**	**0.39**	**0.43**

D. Ablation Study

1) INRM: INRM is a key component of INR-ECGAN. To prove its effectiveness, it is replaced by the traditional bilinear interpolation module (BIM) and a classical upsampling module (UM) [8] in deep learning, respectively. Table II shows their calculation results on SEN1-2 and QXS-SAROPT datasets, where the best value of each metric has been bolded in red font. As can be seen in Table II, INRM outperforms BIM and UM in almost all indicators. This result demonstrates that its continuous expression property helps to better recover the details and structural features of optical images.

2) PPM: PPM is used to mine the deep semantic information of images. To emphasize the importance of deep semantic features for image translation tasks, an ablation study is performed on PPM. Table II presents the test results of INR-ECGAN without PPM on two datasets. It can be analyzed that, INR-ECGAN is improved in all evaluation indicators after

TABLE II. Quality evaluation results of ablation study

Datasets	Method	PSNR↑	FID↓	SSIM↑	LPIPS↓
SEN1-2 [10]	BIM	15.80	83.09	0.29	0.51
	UM	15.91	83.09	**0.30**	0.52
	W/o PPM	15.99	72.07	0.29	0.47
	W/ ℓ_1	16.14	82.11	0.29	0.51
	INR-ECGAN	**16.20**	**70.01**	**0.30**	**0.46**
QXS-SAROPT [5]	BIM	17.26	121.04	0.38	0.48
	UM	17.68	71.77	**0.39**	0.45
	W/o PPM	17.46	66.72	0.37	0.44
	W/ ℓ_1	17.47	100.63	0.37	0.47
	INR-ECGAN	**17.76**	**64.86**	**0.39**	**0.43**

inserting PPM, which implies its effectiveness.

3) ℓ_1 Loss: To validate the impact of ℓ_1 loss, INR-ECGAN with ℓ_1 loss is trained and tested on two datasets. Its evaluation results are shown in Table II. As can be seen from Table II, compared to using ℓ_1 loss, INR-ECGAN without ℓ_1 loss performs better on all metrics. Therefore, discarding ℓ_1 loss in the loss functions allows the proposed network to generate more accurate spatial details.

IV. Conclusion

In this paper, we apply a continuous implicit function to the S2OIT task for the first time and propose a novel framework called INR-ECGAN. Its generator uses several CRBs and a PPM to extract the multi-scale semantic and global information from the input SAR image, respectively. Then, an implicit decoder is designed to fully aggregate the multi-scale semantic and global features and continuously map the fused features to the optical image. In addition, a multi-scale discriminator is adopted to distinguish the generated and real images. Experimental results have demonstrated that the proposed method is superior to the existing S2OIT methods. In the future, to extract richer low-frequency and high-frequency features from images, wavelet transform or fourier transform is considered to be combined with the proposed approach.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 12072366) and the Fundamental Research Funds for the Central Universities (No. 531118010744).

References

[1] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8628–8638.

[2] M. Gao, J. Xu, J. Yu, and Q. Dong, "Distilled heterogeneous feature alignment network for sar image semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[3] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 021–10 030.

[4] X. Gong, Z. Hou, Y. Wan, Y. Zhong, M. Zhang, and K. Lv, "Multispectral and sar image fusion for multi-scale decomposition based on least squares optimization rolling guidance filtering," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[5] M. Huang, Y. Xu, L. Qian, W. Shi, Y. Zhang, W. Bao, N. Wang, X. Liu, and X. Xiang, "The qxs-saropt dataset for deep learning in sar-optical data fusion," *arXiv preprint arXiv:2103.08259*, 2021.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[7] J. Lee, H. Cho, D. Seo, H.-h. Kim, J. Jeong, and M. Kim, "Cfca-set: Coarse-to-fine context-aware sar-to-eo translation with auxiliary learning of sar-to-nir translation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[8] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

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

[10] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The sen1-2 dataset for deep learning in sar-optical data fusion," *arXiv preprint arXiv:1807.01569*, 2018.

[11] T. Shen, Y. Zhang, L. Qi, J. Kuen, X. Xie, J. Wu, Z. Lin, and J. Jia, "High quality segmentation for ultra high-resolution images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1310–1319.

[12] D. Torbunov, Y. Huang, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, and Y. Ren, "Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 702–712.

[13] H. Wang, Z. Zhang, Z. Hu, and Q. Dong, "Sar-to-optical image translation with hierarchical latent features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[14] 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. 129 136–129 149, 2019.

[15] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[16] Z. Wang, Y. Ma, and Y. Zhang, "Hybrid cgan: Coupling global and local features for sar-to-optical image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[17] Y. Xu, C. Lin, Y. Zhong, Y. Huang, and X. Ding, "Recognizer embedding diffusion generation for few-shot sar recognization," in *Chinese Conference on Pattern Recognition and Computer Vision*. Springer, 2023, pp. 418–429.

[18] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 918–12 927.

[19] 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[20] X. Yang, J. Zhao, Z. Wei, N. Wang, and X. Gao, "Sar-to-optical image translation based on improved cgan," *Pattern Recognition*, vol. 121, p. 108208, 2022.

[21] T. Zhan, J. Bian, J. Yang, Q. Dang, and E. Zhang, "Improved conditional generative adversarial networks for sar-to-optical image translation," in *Chinese Conference on Pattern Recognition and Computer Vision*. Springer, 2023, pp. 279–291.

[22] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[24] Y. Zhao, T. Celik, N. Liu, and H.-C. Li, "A comparative analysis of gan-based methods for sar-to-optical image translation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.