Learning SAR-to-Optical Image Translation via Diffusion Models With Color Memory

Zhe Guo, *Member, IEEE*, Jiayi Liu, Qinglin Cai, Zhibo Zhang, 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 the generative adversarial networks paradigm, and encounter the problem of mode collapse, making them difficult to train. SAR and optical images have heterogeneous characteristics and large spectral differences, most of the existing methods do not focus on the color correlation between these two image domains, which leads to spectral distortion and detail errors in translation results. To address these issues, we propose a novel diffusion model capable of memorizing color and directly mapping between the SAR and optical image domains for S2OIT called CM-Diffusion. The color attention Brownian bridge diffusion structure is designed to learn the color correlation and translation between the SAR and optical image domains directly through the bidirectional diffusion process and color attention mechanism, avoiding the conditional information leverage. The color feature extraction module is constructed to provide color and semantic information for the diffusion model. Extensive experiments conducted on three benchmark datasets SEN1-2, QXS-SAROPT, and SEN12MS demonstrate that the proposed CM-Diffusion outperforms the state-of-the-art methods on both subjective and objective evaluation metrics.**

***Index Terms*—Brownian bridge diffusion structure, color memory, diffusion models, synthetic aperture radar (SAR)-to-optical image translation (S2OIT).**

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 [1], [2]. 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 [3]. 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 [4]. 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 [5]. In addition, even for researchers familiar with processing SAR images, SAR-to-optical image translation (S2OIT) is still helpful in providing auxiliary information (translated optical images) to assist researchers in certain SAR processing tasks such as cloud removal [6] and change detection [7]. Due to the rapid development of natural image-to-image translation (I2IT) based on deep learning [8], [9], researchers have begun to pay attention to S2OIT [10]. In recent years, many S2OIT methods based on generative adversarial networks (GANs) [11] have been proposed. As an excellent deep generative model, generative adversarial learning uses generator and discriminator to learn the internal distribution characteristics of data. Some improved methods of GANs, such as conditional generative adversarial nets (CGAN) [12], Pix2pix [13], and Pix2pixHD [14], have been used for I2IT task. However, the above methods are designed by improving the network model and loss function from the perspective of natural images, without considering the color correlation between SAR and optical images, the color information in their S2OIT results is scarce. Moreover, GANs also encounter the problem of mode collapse, making them difficult to train [15].

Diffusion models [16] have recently become a mainstream generative modeling approach that outperforms the current GAN-based generative models for image synthesis [17]. The diffusion model is parameterized by Markov chains and consists of two processes: the forward process, also known as the diffusion process, and the reverse process. The diffusion process corrupts the original distribution by gradually adding Gaussian noise to the data. The reverse process, on the other hand, synthesizes the data from pure noise by iteratively denoising it until clean samples are generated. Diffusion models are trained by optimizing the variational lower bound of negative logarithmic likelihood, thus avoiding the mode collapse often occurs in GANs. Diffusion models have been used for various I2IT tasks, such as superresolution [18], semantic scene synthesis [19], and colorization [20]. Recently, Bai et al. [15] proposed a conditional diffusion model for S2OIT, which incorporates SAR images as conditions into the training and inference process of the

Manuscript received 1 April 2024; revised 28 June 2024; accepted 1 August 2024. Date of publication 6 August 2024; date of current version 26 August 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; liujiayiqqq@mail.nwpu.edu.cn; qlcai@mail.nwpu.edu.cn; billz@mail.nwpu.edu.cn; meish@nwpu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3439516

diffusion models, and generates the optical images with clearer boundaries.

However, SAR and optical images have heterogeneous characteristics and large spectral differences [21], most of the existing S2OIT methods do not focus on the correlation between the color information of optical images and the spatial features of SAR images, which lead to spectral distortion and detail errors in translation results for complex terrain scenes. In addition, as the most competitive I2IT generative approach currently, the potential of diffusion models for S2OIT tasks remains to be explored.

Inspired by the remarkable performance of the diffusion models, in this article, we propose a novel S2OIT framework called CM-Diffusion considering the color correlation and directly mapping between the SAR and optical image domains, which has the promising originality of color memory and distinct domains adaptability. In order to effectively memorize the color information of optical images, we design the color feature extraction (CFE) module, which constructs the correspondence between the spatial features of SAR images and the color features of optical images, providing color and semantic information for the diffusion model. Moreover, inspired by the Brownian bridge diffusion model (BBDM) [20], we propose the color attention Brownian bridge diffusion structure (CA-BBDS), learns the translation between the SAR and optical image domains directly through the bidirectional diffusion process, and further adds a color attention mechanism in the reverse denoising process to better integrate the color correlation between the two image domains. A series of experiments over three challenging datasets are conducted to assess the effectiveness and rationality of our CM-Diffusion for S2OIT.

Overall, our main contributions can be summarized as follows.

1) We propose a color memory diffusion model (CM-Diffusion) for S2OIT to address color memory for different scenes, which considers the color correlation and direct mapping between the SAR and optical image domains, facilitating the generation of optical images with higher color detail and better domain adaptation and thus aiding in the more efficient interpretation of SAR images with complex terrain scenes.
2) We propose the CA-BBDS, which directly learns the translation between the SAR and optical image domains through the bidirectional diffusion process, avoids the conditional information leverage, and further adds a color attention mechanism in the reverse denoising process to better integrate the color correlation between the two image domains.
3) We propose CFE module to construct the correspondence between the spatial features of SAR images and the color features of optical images, providing color and semantic information for the diffusion model.
4) Extensive experimental evaluations on three benchmark datasets SEN1-2 [22], QXS-SAROPT [23], and SEN12MS [24] show that the proposed method outperforms the state-of-the-art on both subjective and objective evaluation metrics, such as peak signal-to-noise ratio (PSNR) [25], structural similarity index metric (SSIM) [26], and learned perceptual image patch similarity (LPIPS) [27].

The rest of this article is organized as follows. Section II briefly introduces the basic knowledge of diffusion models, natural I2IT and S2OIT. The proposed CM-Diffusion 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. Diffusion Models

Diffusion models are recently proposed advanced generative models, which have shown competitive performance in many computer vision tasks compared with GANs. In 2020, Ho et al. [28] proposed denoising diffusion probabilistic model (DDPM) based on the diffusion models, which introduced the diffusion models into the field of image generation. In recent years, diffusion models have developed rapidly as a powerful family of generative models, and methods to accelerate model training have also emerged. Rombach et al. [19] proposed latent diffusion model (LDM), which combines the diffusion models with transformer, and demonstrates the potential of diffusion models in various fields such as text generation image, image editing, image restoration, etc. The impressive stable diffusion is implemented based on LDM. Li et al. [20] presented the BBDM that directly builds the mapping between the input and the output domains through a Brownian bridge stochastic process, rather than a conditional generation process. Compared with GANs, diffusion models have the advantages of diversity, training stability, and scalability, and perform better in high-resolution, large-scale image-to-image translation.

B. Natural Image-to-Image Translation

In 2017, Isola et al. [13] 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. [29] proposed CycleGAN, which consists of two generators and discriminators, and can use unpaired images for I2IT task. Recently, Shaham et al. [30] introduced spatially adaptive pixelwise networks (ASAP-Net) for fast image translation. Jung et al. [31] introduced contrast learning based on CGAN to improve the effectiveness of translated images. Guo et al. [32] proposed structural consistency constraints to mitigate semantic distortions in unpaired image translation. Hu et al. [33] proposed a new semantic relation consistency regularization, and by introducing decoupled contrast learning, image translation achieved better performance. With the advent of the DDPM [28] and LDM [19] methods, diffusion models have been used for various I2IT tasks. Sasaki et al. proposed UNIT-DDPM [34], which uses DDPM without requiring adversarial training. Ruiz et al. [35] presented a method for fine-tuning images using textual cues based on LDM. Zhang et al. [36] proposed the SINE method to try to solve

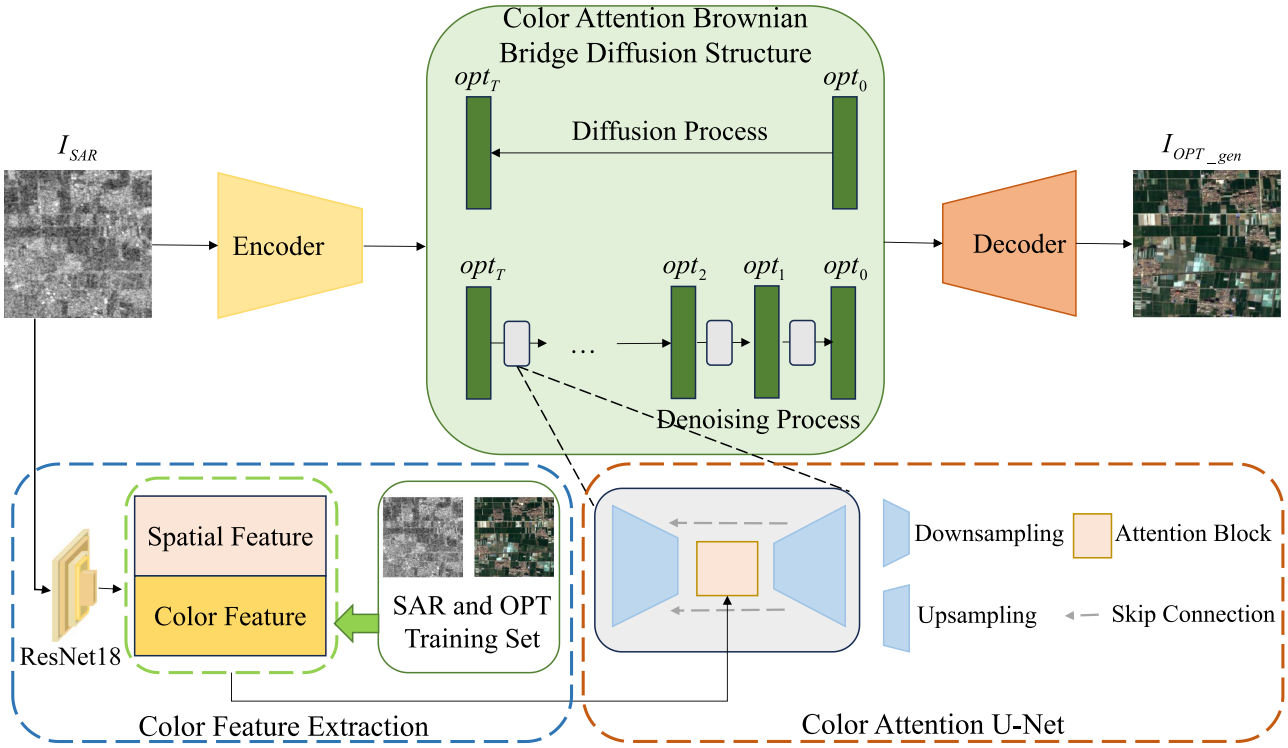

Fig. 1. Overall framework of CM-diffusion model. I_{SAR} means the SAR image, and I_{OPT_gen} means the generated optical image. opt_0 and opt_T represents the start and the end of the diffusion process, respectively.

the editing problem of a single image. Wang et al. [37] replaced the discrete coded latent space with a binary-valued latent space for diffusion, which accomplished the image translation task excellently.

I2IT approaches are oriented to multidomain and object recognition [38]. 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. [7] 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, Tan et al. [39] proposed a feature-preserving heterogeneous remote sensing image transformation model serial GANs. Darbaghshahi et al. [6] removed the clouds using two GANs by translating SAR to optical images and proposed dilated residual inception blocks in the generator. Li et al. [40] presented a wavelet feature learning network combined with the CycleGAN framework, which can learn features more efficiently. Wang et al. [41] designed an image translation network with two subnetworks, which learns richer optical features of the input image through the optical reconstruction subnetwork. Yang et al. [42] 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. [43] combined two classical methods Pix2pix and CycleGAN and applied them to SAR images to design a semisupervised image translation framework for image matching. Recently, Bai et al. [15] applied the diffusion model to the S2OIT task, and proposed a conditional diffusion model that incorporates SAR images as conditions into the training and inference process of the diffusion model.

The existing S2OIT 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. SAR and optical images have heterogeneous characteristics and large spectral differences, which lead to spectral distortion in the generated optical results in urban and rural scenes with complex terrain.

III. Proposed Method

We propose CM-Diffusion model to cope with the S2OIT task, which can memorize the colors of different scenes, and has a strong distinct domains adaptability, thus learning more reliable translation between two image domains and generating high color details and high-quality optical images. The overall framework of the model is shown in Fig. 1. The SAR image $\boldsymbol{I}_{SAR}$ is fed into the vector quantized generative adversarial networks (VQ-GAN) [44] encoder, which maps the $\boldsymbol{I}_{SAR}$ from pixel space to the latent layer space to obtain the latent feature vector $\boldsymbol{sar}$. The $\boldsymbol{sar}$ is further input into the constructed CA-BBDS. The CA-BBDS employs the Brownian bridge diffusion method in the diffusion process, which learns the translation

between the SAR and optical image domains directly through the bidirectional diffusion process. In the backward denoising process of CA-BBDS, the color attention U-Net network is designed, which combines the CFE module to establish the correspondence between the spatial features of the SAR image and the color features of the optical image. This integration of color correlation results in the translation of the latent space vectors of the optical image, denoted as $\boldsymbol{opt}$. Finally, the $\boldsymbol{opt}$ is mapped back to the pixel space by the VQ-GAN decoder to obtain the final generated optical image $\boldsymbol{I}_{OPT\_gen}$.

A. Color Feature Extraction

Since the imaging mechanism of SAR images is different from that of optical images, targets with different colors may have the same grayscale values on SAR images, which will result in targets with different original colors having the same color after translation. To address this issue, we design the CFE module with color memory capability, which can construct the correspondence between spatial features and color features among the SAR and optical image domains, and give the color feature guidance according to the spatial features of the input SAR image, so that the network gives more realistic color generation results.

Specifically, we use paired SAR and optical images dataset to learn the color correlation. We take the pooling layer information of the SAR images after $conv$ 3_2 through the ResNet18 network pretrained on ImageNet. As the pooling operation preserves the structural information of the image, the obtained pooling layer information can be used as the spatial feature S_sar of the SAR image. For the color features, we use colorthief [45] to extract the Top-30 color values (RGB values) of the corresponding optical image in the training set to obtain a color vector. After normalizing this color vector, we get the color feature C_opt with a size of 3×30. The color correlation F_color of each paired SAR and optical image can be described as

$$F\_color = (S\_sar_1, C\_opt_1), (S\_sar_2, C\_opt_2), ..., (S\_sar_m, C\_opt_m) \tag{1}$$

where m represents the capacity of the CFE module, which is the same as the number of the training set.

During the training step, the pooling layer information of the input SAR image after the $conv$ 3_2 of the pretrained ResNet18 network is then output as the spatial feature, which is used to calculate the cosine similarity with all the spatial features saved in the CFE module. The KNN method is used to match the color features that correspond to the spatial features with the closest cosine distance. These color features are then input into the color attention U-Net network of the CA-BBDS through the cross-attention, which guide the network to generate more realistic colors.

B. Color Attention Brownian Bridge Diffusion Structure

Most of the conditional diffusion models suffer from poor model generalization, and can only be adapted to some specific applications where the conditional inputs and outputs are highly similar, while they are not suitable for image translation task between two different domains such as the S2OIT. Furthermore, existing diffusion models based image translation methods do not focus on the correlation between the color information of optical image and the spatial features of SAR image, which leads to spectral distortion and detail errors in translation results for complex terrain scenes.

To solve the above problems, inspired by Li et al. [20], we design CA-BBDS utilizing the Brownian bridge diffusion method in the forward diffusion process, which learns the translation between the SAR and optical image domains directly through the bidirectional diffusion process rather than a conditional generation process. The color attention mechanism is added to the backward denoising process to better incorporate the color correlation extracted by the CFE module.

Different from the existing DDPM methods, our CA-BBDS is not based solely on Gaussian noise as the ending point, but rather on the SAR domain image as the ending point and the optical domain image as the starting point. Let $(\boldsymbol{opt}, \boldsymbol{sar})$ denote the paired latent feature vectors from optical domain and SAR domain by adopting VQ-GAN [44] to map the image to the latent space, the forward diffusion process of our CA-BBDS can be defined as

$$p_{CA}\left(\boldsymbol{opt}_t|\boldsymbol{opt}_0, \boldsymbol{sar}\right) = \mathcal{N}\left(\boldsymbol{opt}_t; \left(1 - \frac{t}{T}\right)\boldsymbol{opt}_0 + \frac{t}{T}\boldsymbol{sar}, \delta_t \boldsymbol{I}\right) \tag{2}$$

$$\boldsymbol{opt}_0 = \boldsymbol{opt} \tag{3}$$

where opt_0 represents the starting point of the diffusion process, opt_t denotes the diffusion result after t steps, T represents the total number of diffusion steps, and δ_t represents the variance. In the diffusion process, the optical domain image is mapped to the SAR domain, while in the denoising process, the SAR domain image is mapped to the optical domain.

The variance δ_t can be calculated as

$$\begin{aligned}\delta_t &= \alpha\left(1 - \left(\left(1 - \left(\frac{t}{T}\right)\right)^2 + \left(\frac{t}{T}\right)^2\right)\right) \\ &= 2\alpha\left(\frac{t}{T} - \left(\frac{t}{T}\right)^2\right)\end{aligned} \tag{4}$$

where α denotes the scaling factor, which is used to adjust the variance in the diffusion process.

The forward diffusion process of our CA-BBDS has the ability to establish a direct mapping relationship between the source and target domains, enabling more effective utilization of paired SAR and optical images for better translation results.

We propose a color attention U-Net that utilizes the attention mechanism to predict images for the denoising process. The structure of the color attention U-Net is illustrated in Fig. 2.

Our color attention U-Net network comprises encode Resblocks, decode Resblocks, downsampling layers, upsampling layers, and attention block. The attention block includes a self-attention network layer and a cross-attention network layer.

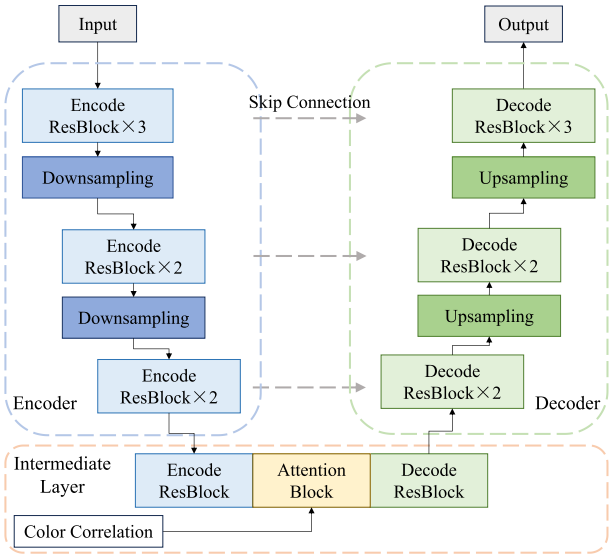

Fig. 2. Color attention U-Net network structure.

The skip connection is able to fuse the shallow detailed features of the encoder with the deeper semantic features of the decoder through splicing operations, allowing the network to better utilize contextual information. For a latent space vector with an input size of 64×64, image features of different sizes are extracted alternately through the encode Resblock and the downsampling layer. Then, the features are extracted through the intermediate layer, and the color features obtained by the CFE module are fused using the attention block. Finally, the image is recovered alternately through the decode Resblock and the upsampling layer, and the prediction results are obtained. The color attention U-Net network fuses the color features of the image during the prediction process. This guides the network to generate optical images with more realistic colors and higher quality.

C. Training Process for the CM-Diffusion Model

1) Forward Diffusion Process: To interfere with and train the proposed CM-Diffusion model, it is necessary to deduce the forward transition probability $p_{CA}(\boldsymbol{opt}_t|\boldsymbol{opt}_{t-1}, \boldsymbol{sar})$. Combined with the marginal distribution at step t of the diffusion process given in (2), the discrete form of $\boldsymbol{opt}_t$ and $\boldsymbol{opt}_{t-1}$ can be computed by giving an initial state $\boldsymbol{opt}_0$ and a target state $\boldsymbol{sar}$ as follows:

$$\boldsymbol{opt}_t = \left(1 - \frac{t}{T}\right)\boldsymbol{opt}_0 + \frac{t}{T}\boldsymbol{sar} + \sqrt{\delta_t}\boldsymbol{\varphi}_t \tag{5}$$

$$\boldsymbol{opt}_{t-1} = \left(1 - \frac{t-1}{T}\right)\boldsymbol{opt}_0 + \frac{t}{T}\boldsymbol{sar} + \sqrt{\delta_{t-1}}\boldsymbol{\varphi}_{t-1} \tag{6}$$

where $\boldsymbol{\varphi}_{t-1}, \boldsymbol{\varphi}_t \in \mathcal{N}(0, \boldsymbol{I})$.

The transition probability $p_{CA}(\boldsymbol{opt}_t|\boldsymbol{opt}_{t-1}, \boldsymbol{sar})$ can be calculated by combining (5) and (6) as follows:

$$p_{CA}\left(\boldsymbol{opt}_t|\boldsymbol{opt}_{t-1}, \boldsymbol{sar}\right) = \mathcal{N}\left(\boldsymbol{opt}_t; \frac{(1-\frac{t}{T})}{(1-\frac{t-1}{T})}\boldsymbol{opt}_{t-1} + \left(\frac{t}{T} - \frac{(1-\frac{t}{T})}{(1-\frac{t-1}{T})}\frac{t-1}{T}\boldsymbol{sar}, \delta_{t|t-1}\boldsymbol{I}\right)\right) \tag{7}$$

where $\delta_{t|t-1}$ can be calculated by δ_t, shown as

$$\delta_{t|t-1} = \delta_t - \delta_{t-1}\frac{(1-\frac{t}{T})^2}{(1-\frac{t-1}{T})^2}. \tag{8}$$

This forward diffusion process allows to fix the mapping between the optical domain and SAR domain.

2) Backward Denoising Process: The backward denoising process aims to predict $\boldsymbol{opt}_{t-1}$ based on $\boldsymbol{opt}_t$. The formula is shown as follows:

$$q_\phi(\boldsymbol{opt}_t|\boldsymbol{opt}_{t-1}, \boldsymbol{sar}) = \mathcal{N}(\boldsymbol{opt}_{t-1}; \mu_\phi(\boldsymbol{opt}_t, t), \tilde{\delta}_t\boldsymbol{I}) \tag{9}$$

where $\mu_\phi(\boldsymbol{opt}_t, t)$ denotes the mean value of the noise obtained by predicting the noise using a neural network with the parameter ϕ, and $\tilde{\delta}_t$ denotes the variance of the noise at each step. In our CM-Diffusion model, we design CA-BBDS by using the color attention U-Net to better incorporate the color correlation extracted by the CFE module. Therefore, the color attention U-Net is used as the denoising neural network here to predict the noise.

3) Training Objective: The training process of our CM-Diffusion model is performed as an optimization problem that combines the evidence lower bound [28] and a reparametrization method [20], which can be formulated as

$$\mathrm{CM} = \mathbb{E}_{\boldsymbol{opt}_0, \boldsymbol{sar}, \varphi}\left[c_{\varphi t}||\frac{t}{T}(\boldsymbol{sar} - \boldsymbol{opt}_0) + \sqrt{\delta_t}\varphi - \varphi_\phi(\boldsymbol{opt}_t, t)||^2\right] \tag{10}$$

$$c_{\varphi t} = \left(1 - \frac{t-1}{T}\right)\frac{\delta_{t|t-1}}{\delta_t} \tag{11}$$

where $\varphi_\phi(\boldsymbol{opt}_t, t)$ represents the predicted noise and $c_{\varphi t}$ denotes a quantity related to the number of steps using the reparameterization technique.

In this article, we use the color attention U-Net as the denoising neural network φ_ϕ to learn the noise, and the loss function used for the color attention U-Net is the $L1$ loss function.

IV. Experiments

A. Experimental Setup

1) Datasets: To validate the effectiveness of our proposed method, we conduct comprehensive comparison experiments with state-of-the-art methods on three public challenging datasets SEN1-2 [22], QXS-SAROPT [23], and SEN12MS [24].

SEN1-2 dataset: The SEN1-2 dataset includes 282 384 pairs of corresponding images, SAR images (VV polarization channel) 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

TABLE I
NUMBER OF SCENE DATA PER CATEGORY OF THE SEN1-2 DATASET

Category	Train	Test
Urban	911	103
Mountain	1044	116
Rural	528	58
Grassland	487	54
Farmland	493	54
Others	137	15

TABLE II
NUMBER OF SCENE DATA PER CATEGORY OF THE SEN12MS DATASET

Category	Train	Test
Urban	3780	415
Mountain	1695	187
Rural	1083	128
Grassland	2542	275
Farmland	2613	285
Plain	438	56
Others	143	20

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.

SEN12MS dataset: The SEN12MS dataset contains 180 662 patch triplets of corresponding Sentinel-1 dual-pol SAR data (both VV and VH polarization channels), Sentinel-2 multispectral images, and MODIS-derived land cover maps. The patches are distributed across the land masses of the Earth and spread over all four meteorological seasons. The image size is 256×256.

For the SEN1-2 dataset, we randomly select 12 300 pairs of SAR and optical images as the training set, and 400 SAR images as the test set. The selection is based on the complexity of the dataset and the impact of data training time on experimental efficiency. For each scene, the training and test sets are 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, mountain, rural, grassland, farmland, and other scenes as shown in Table I.

For the QXS-SAROPT dataset, we also randomly select 3600 SAR images and 3600 optical images for paired training and selected 400 SAR images for testing. The QXS-SAROPT dataset contains mainly urban scene and some mountain scene, thus we extract the training and test sets according to the proportion of scenes in the entire dataset.

For the SEN12MS dataset, we use the SAR images with VV polarization channel and paired optical images with Band 2 to Band 4 (visible blue, green, and red) from the Sentinel-2 multispectral channels. We randomly select 12 294 pairs of SAR and optical images as the training set, and 1366 SAR images as the test set. Following a similar approach to the SEN1-2 dataset, we categorize these training and test images into urban, mountain, rural, grassland, farmland, plain, and other scenes as shown in Table II.

We also preprocess and randomly enhance the data before inputting into the network. We first resize the image from 256×256 to 286×286 by double triple interpolation method, and then randomly flip it left and right, and finally randomly crop 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. The method in this article consists of two parts: the pretrained VQ-GAN and the CM-Diffusion model. The VQ-GAN uses the same pretrained model as the LDM [19]. For the training of the CM-Diffusion model, we set the same training parameters for three different datasets. Specifically, 200 rounds of training are conducted for all data, with a total of 500 diffusion steps and 200 sampling steps. The variance control factor is set to $\alpha = 0.1$, and the optimizer uses a learning rate of 0.0001 with Adam optimization. The basis for the selection of training parameters will be analyzed in detail in the ablation study section.

3) Evaluation Metrics: We conduct evaluations using three different mainstream evaluation metrics, which are PSNR [25], SSIM [26], and LPIPS [27]. PSNR and SSIM 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:

$$\mathrm{PSNR}(x, y) = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{12}$$

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:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{13}$$

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.

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 CM-Diffusion for S2OIT task in terms of both subjective and objective evaluation. We chose seven image translation comparison methods, Pix2pix [13], Pix2pixHD [14], Serial GANs [39], Darbaghshahi's (abbreviated as Dar) [6], ASAP-Net [30], LDM [19], and BBDM [20]. Among the above comparison

TABLE III
EVALUATION INDEX RESULTS ON DIFFERENT DATASETS

	SEN1-2			QXS-SAROPT			SEN12MS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Pix2pix [13]	11.32	0.159	0.604	12.65	0.288	0.608	15.09	0.309	0.574
Pix2pixHD [14]	15.55	0.277	0.386	14.65	0.342	0.544	18.13	0.436	0.426
Serial GANs [39] (*RS2021*)	15.04	0.263	0.396	14.78	0.353	0.521	17.80	0.407	0.491
Dar [6] (*TGRS2022*)	14.99	0.250	0.412	14.32	0.354	0.555	17.66	0.442	0.461
ASAP-Net [30] (*CVPR2021*)	14.64	0.260	0.491	15.23	0.332	0.556	17.61	0.446	0.498
LDM [19] (*CVPR2022*)	15.27	0.293	0.452	15.37	0.341	0.512	18.14	0.448	0.453
BBDM [20] (*CVPR2023*)	15.23	0.289	0.448	15.43	0.347	0.497	17.98	0.450	0.450
CM-Diffusion (*ours*)	**15.86**	**0.351**	**0.378**	**15.66**	**0.355**	**0.450**	**18.46**	**0.464**	**0.404**

↑ means bigger is better, ↓ means smaller is better.
The bold indicates optimal.

methods, Pix2pix and Pix2pixHD are baselines of the I2IT, Serial GANs, Dar and ASAP-Net are all S2OIT methods based on CGANs, while LDM conducts image translation by conditional diffusion models. BBDM is the first work of Brownian bridge diffusion process proposed for I2IT. 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 three evaluation metrics are shown in Table III, with bold indicating the best. It can be seen that our proposed CM-Diffusion performs better than the other comparative methods in PSNR metric, indicating that the translation result of our method is closer to the real optical color image in pixel distance. Our CM-Diffusion is also better than other methods in SSIM metric, indicating that the overall structure of the generated image obtained by our method is more similar to that of the real optical image. On the perceptual distance metric of LPIPS, the result of CM-Diffusion is higher than that of other methods, which reflects that CM-Diffusion has the best effect on human subjective visual perception, and provides convenience for subsequent image information interpretation.

2) Visualization Comparison: Figs. 3–7 show the optical images generated by different methods on three datasets SEN1-2, QXS-SAROPT, and SEN12MS, 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. 3 and 4 contain images from the SEN1-2 dataset in five scenes: urban, mountain, rural, grassland, and farmland. The QXS-SAROPT dataset is dominated by urban scene and also contains some mountains, so Fig. 5 contains both urban and mountain scenes. Figs. 6 and 7 contain images from the SEN12MS dataset in six scenes: urban, mountain, rural, grassland, farmland, and plain.

By comparison, we find that the Pix2pix is not suitable for the S2OIT task due to its relatively simple network structure, the color and terrain details of the translation results are far from reality, and the color difference between the generated image and the real optical image is still relatively obvious, as the residual results show. Pix2pixHD, Dar, and Serial GAN are all based on the CGAN network. However, the overall results of these three methods are not sensitive sufficiently to image details, and the translation results are not realistic enough for urban scenes with dense buildings and roads. In addition, there are some color mottling phenomena and regional color errors in the generated images. As a result, the residual results of these three methods also demonstrate a large error distribution with respect to real optical images. ASAP-Net is also based on CGAN, which uses a lightweight structure to reduce image translation time. However, it produces artifacts in the S2OIT task and the translation results are not accurate enough in terms of both color and details. LDM is based on conditional diffusion models, and suffers from conditional information leverage during the diffusion process. Thus, when there exists a large difference between the spectral distributions of the same terrain shown in the ninth row of Fig. 3, the fifth row of Fig. 4, the third row of Fig. 5, the fifth row of Fig. 6, and the ninth row of Fig. 7, LDM cannot generate satisfactory results due to the mechanism of integrating conditional input into the diffusion model, as the residual results show. BBDM avoids the conditional information leverage, and is more sensitive to details but there is a certain gap with our CM-Diffusion in terms of color accuracy and saturation for the urban residential area and farmland scenes (e.g., the ninth row of Fig. 4, and the first row of Fig. 6). In addition, the residual results between the generated image and the real optical image tend to be more pronounced in the yellow color range compared to our CM-Diffusion.

To further compare the ability of different methods to memorize colors of different scenes, the zoom-in images of the translation results are given in Fig. 8. From the results in Figs. 3–7, it can be seen that the overall color of the images generated by the Pix2pix, Dar, serial GAN, and LDM methods has a large discrepancy with the real image, so the zoom-in images of the above four methods are not shown in Fig. 8. As can be seen from Fig. 8, images produced by other methods lack detail and color accuracy, and often group dense clusters of buildings. Meanwhile, the generated results are affected by blurring, especially for the urban scene that contains more complex semantic information (the fourth row of Fig. 8). In certain cases, incorrect terrain is generated directly, such as mistranslating buildings as land (the first and second rows of Fig. 8). Even in cases of farmland and mountain scenes, other methods also exhibit localized mistranslations in capturing the edge details of terraces

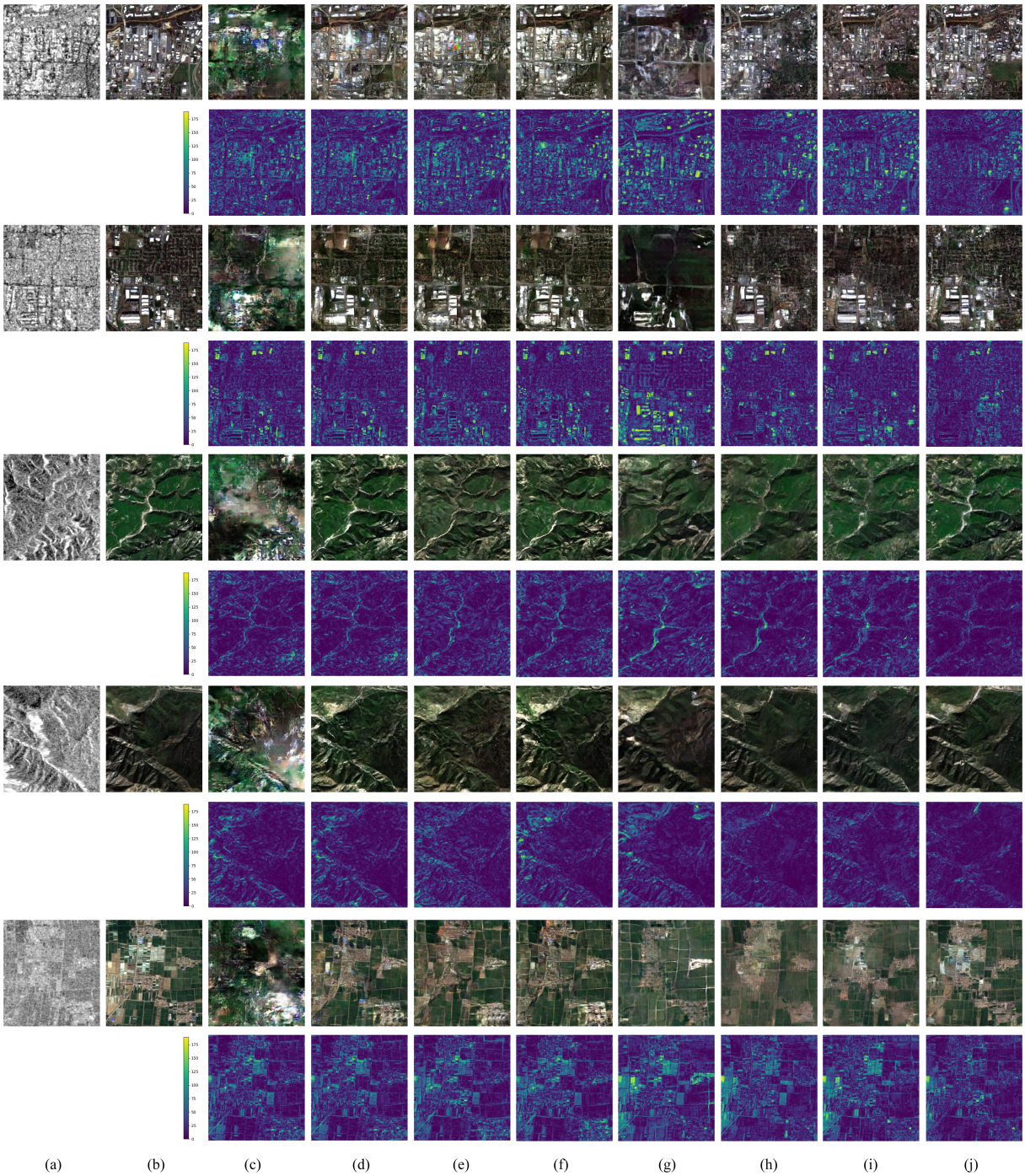

Fig. 3. Visualization results of different methods for S2OIT under different scenes on the SEN1-2 dataset. (a) Original SAR image. (b) Real optical image. (c)–(j) represent the result of Pix2pix, Pix2pixHD, Dar, Serial GAN, ASAP-Net, LDM, BBDM, and our CM-diffusion, respectively. The residual results are under the generated optical images of each method. The scenes are urban, urban, mountain, mountain and rural, from top to bottom.

and mountains (the last two rows of Fig. 8). These issues pose challenges for future image interpretation.

From the overall visual effect, our CM-Diffusion gives the best optical translation results for SAR images of various scenes in all three datasets in terms of color accuracy, color saturation, and texture detail, indicating that our CM-Diffusion has the most outstanding ability to memorize colors. In addition, the CA-BBDS in our method is able to learn the translation between the SAR and optical image domains directly through the bidirectional diffusion process instead of a conditional generation process, avoiding the conditional information leverage and providing better detail sensitivity. The CFE module and color attention U-Net in our CM-Diffusion model can construct the correspondence between the spatial features of SAR images and the color features of optical images, provide color and semantic information for the diffusion model and generate more accurate

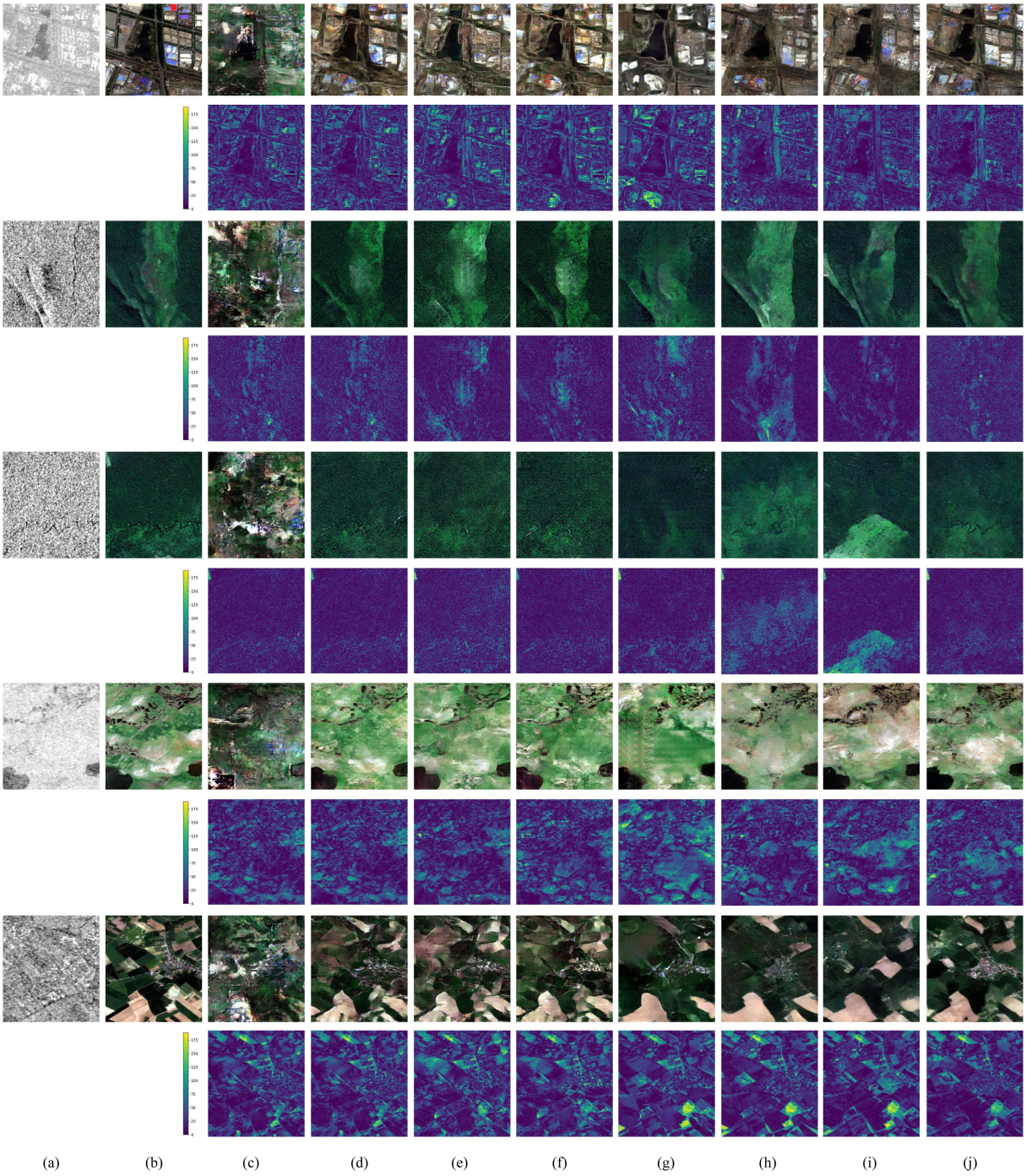

Fig. 4. Visualization results of different methods for S2OIT under different scenes on the SEN1-2 dataset (continued). (a) Original SAR image, (b) the real optical image, (c)–(j) represent the result of Pix2pix, Pix2pixHD, Dar, Serial GAN, ASAP-Net, LDM, BBDM, and our CM-diffusion, respectively. The residual results are under the generated optical images of each method. The scenes are rural, mountain, mountain, grassland, and farmland, from top to bottom.

optical images with higher detail and color accuracy, particularly in building-intensive and color-rich scenes (the urban scene examples in Fig. 8). As shown by the visual comparison results, the optical images generated by our CM-Diffusion have more details, more accurate colors, and have better performance overall.

3) Ablation Study. Validation of Module Effectiveness: We evaluate each component in the proposed CM-Diffusion, i.e., CFE and CA-BBDS. Table IV gives the comparison results of the three evaluation metrics on the three datasets, based on baseline (DDPM) with the addition of each module and the combination of both modules. "+ CFE" and "+ CA-BBDS" denote adding the two modules in baseline, respectively, and "+ all" represents CM-Diffusion after all combinations. Among them, "+ CFE" indicates that directly use the color features as a condition for model generation. All the parameter settings are the same in

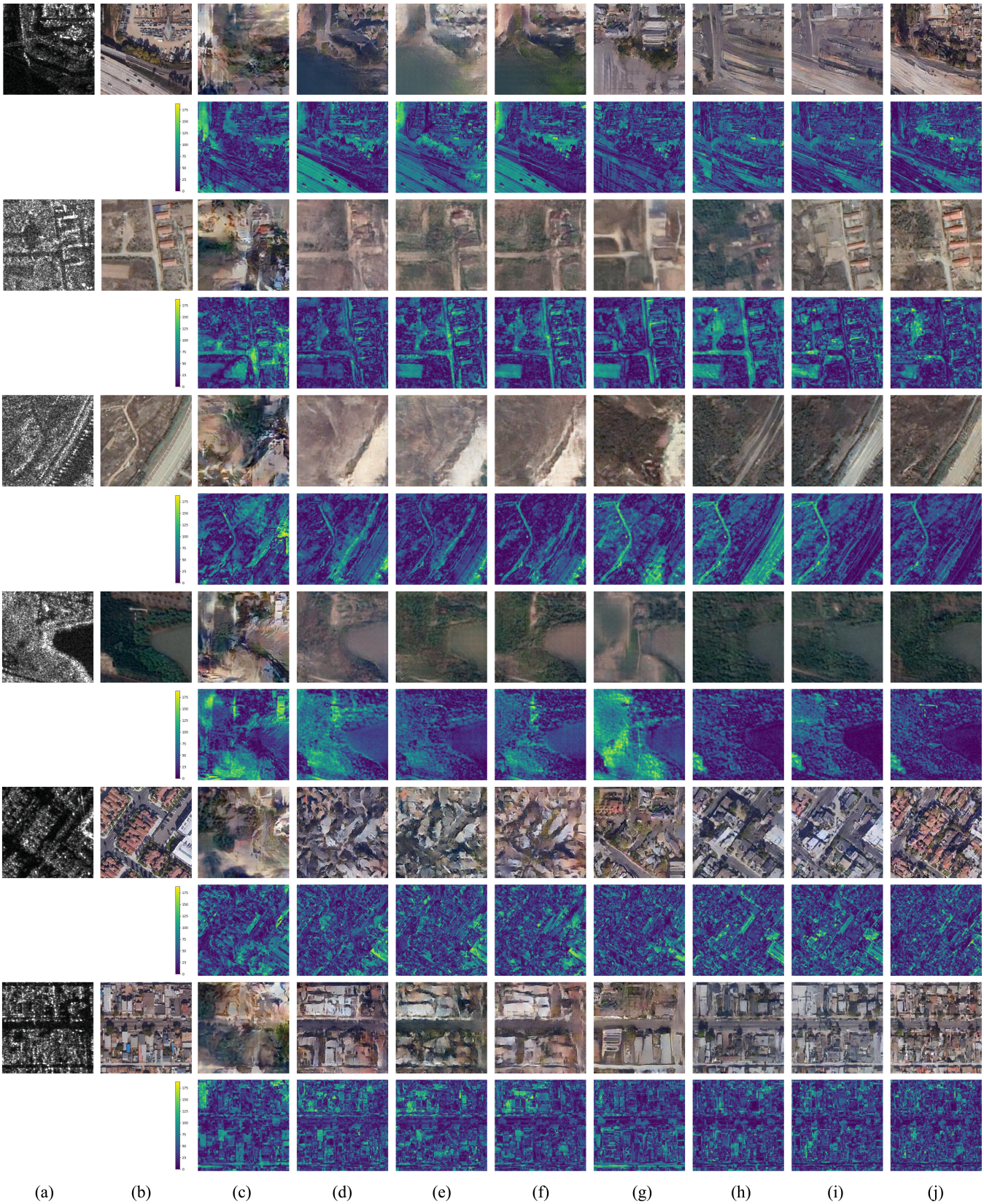

Fig. 5. Visualization results of different methods for S2OIT under different scenes on the QXS-SAROPT dataset. (a) Original SAR image, (b) the real optical image, (c)–(j) represent the result of Pix2pix, Pix2pixHD, Dar, Serial GAN, ASAP-Net, LDM, BBDM, and our CM-diffusion, respectively. The residual results are under the generated optical images of each method. The scenes are urban, urban, urban, mountain, urban and urban, from top to bottom.

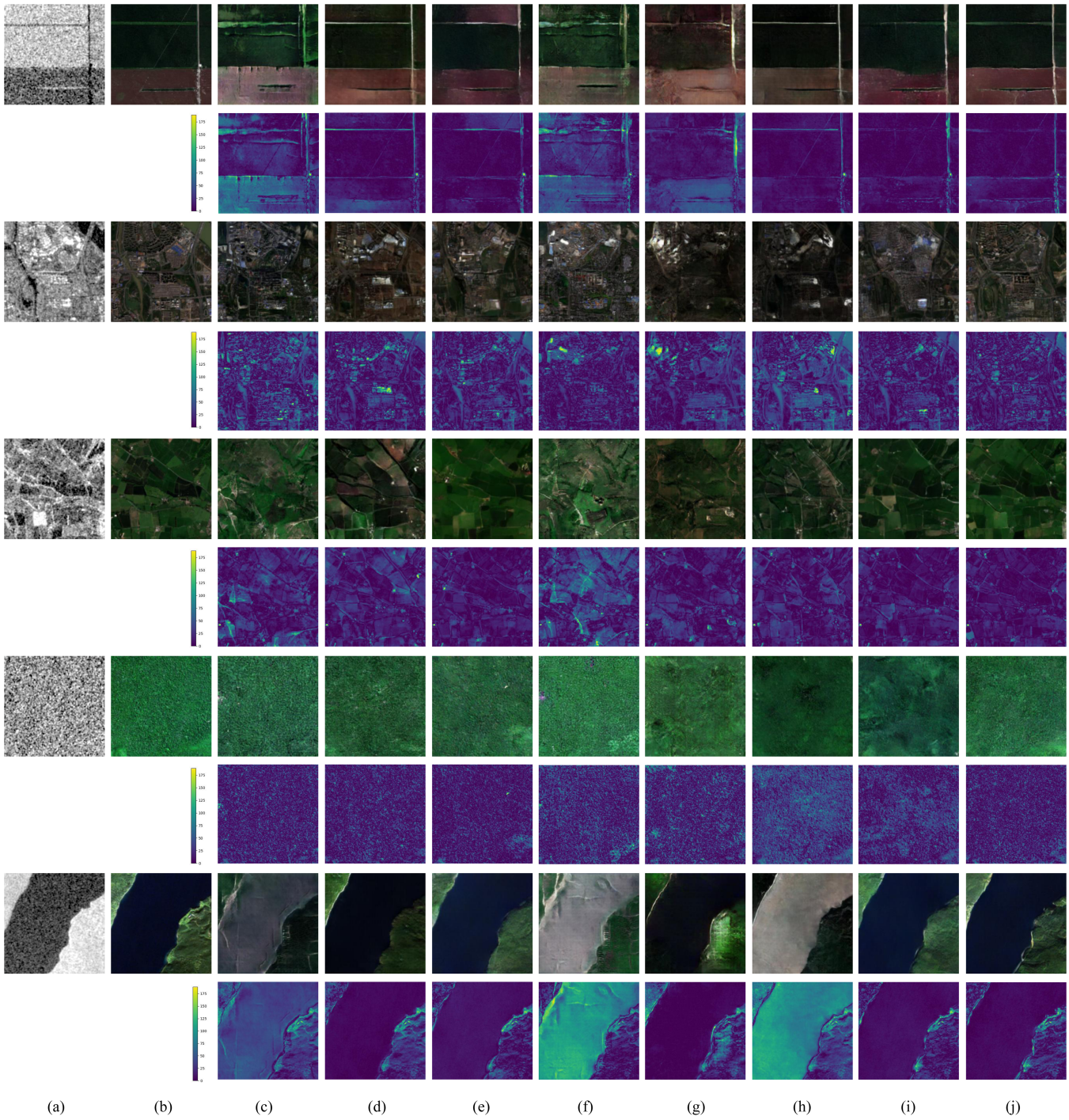

Fig. 6. Visualization results of different methods for S2OIT under different scenes on the SEN12MS dataset. (a) Original SAR image, (b) the real optical image, (c)–(j) represent the result of Pix2pix, Pix2pixHD, Dar, Serial GAN, ASAP-Net, LDM, BBDM, and our CM-diffusion, respectively. The residual results are under the generated optical images of each method. The scenes are farmland, urban, farmland, grassland, and mountain, from top to bottom.

the experiment except for the different modules. As can be seen from Table IV, each of our proposed modules combined with the baseline significantly improves the PSNR, SSIM, and LPIPS of the generated images. This suggests that each of our proposed modules can improve the model's learning ability, and the optical image generated by each module combined with the baseline is more consistent with the real optical color image overall. The optical images generated by combining both of these two modules ("+ all") are optimal in all metrics.

Training Parameters Selection: We discuss two important training parameters in the diffusion model: the number of sampling steps and the variance control factor α. The number of sampling steps affects the quality of the generated image, while the variance control factor α affects the stability

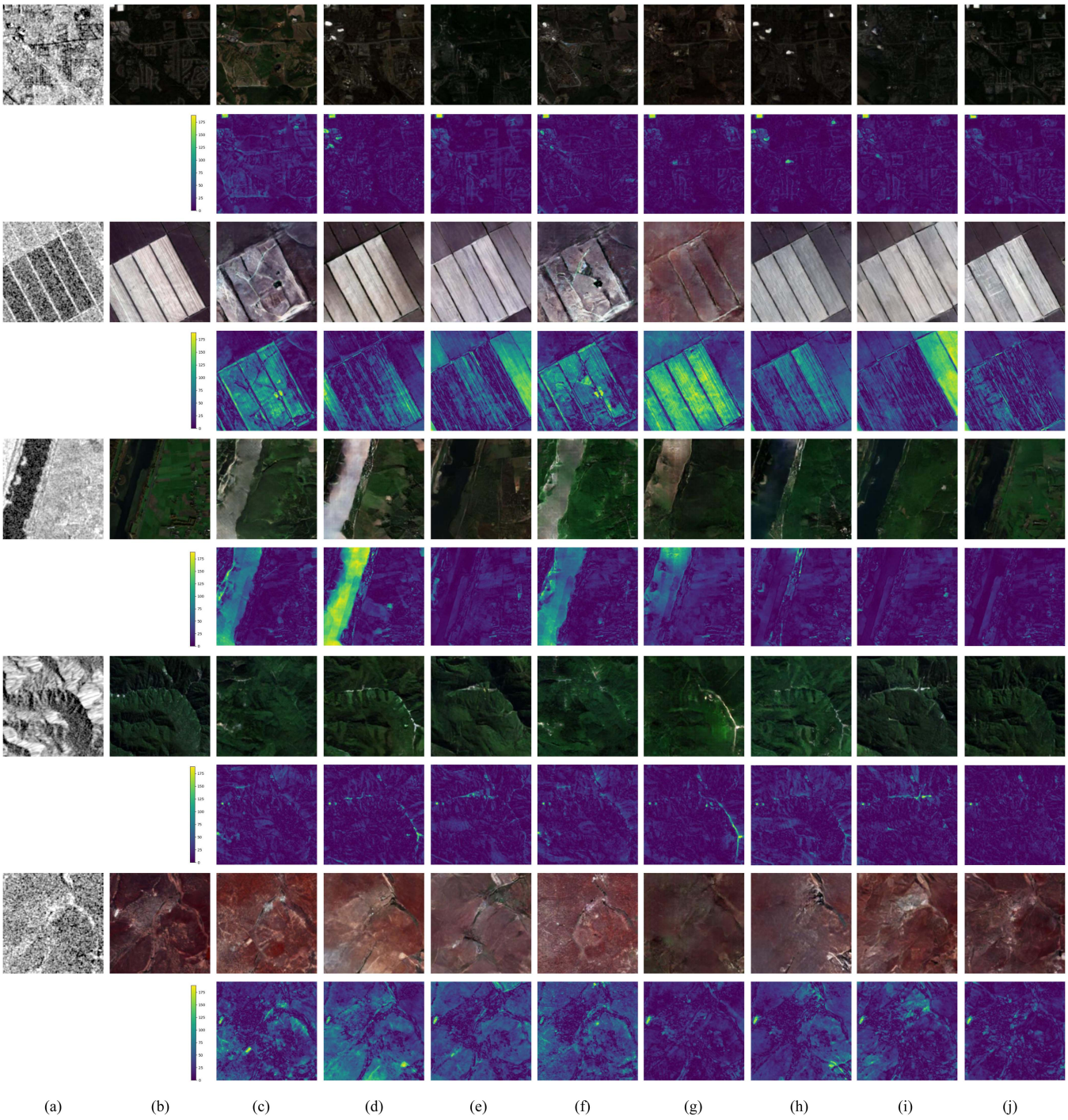

Fig. 7. Visualization results of different methods for S2OIT under different scenes on the SEN12MS dataset (continued). (a) Original SAR image, (b) the real optical image, (c)–(j) represent the result of Pix2pix, Pix2pixHD, Dar, Serial GAN, ASAP-Net, LDM, BBDM, and our CM-Diffusion, respectively. The residual results are under the generated optical images of each method. The scenes are rural, farmland, farmland, mountain, and plain, from top to bottom.

of the conversion between the two domains. Experiments are conducted on the SEN1-2 dataset, with all other parameters held constant. Corresponding results are presented in Tables V and VI. Bolding indicates optimal, underlining indicates sub-optimal. The quality of the generated image is affected by the number of sampling steps. The results show that when the number of sampling steps is less than 200, the quality of the optical image generated by CM-Diffusion improves with an increase in the number of sampling steps. This is because more sampling steps refine the image details, resulting in more stable network outputs. However, once the number of sampling steps exceeds 200, the evaluation metrics, except for the PSNR, which rises slightly, both SSIM and LPIPS decrease. Therefore, increasing the number of sampling steps does not improve the

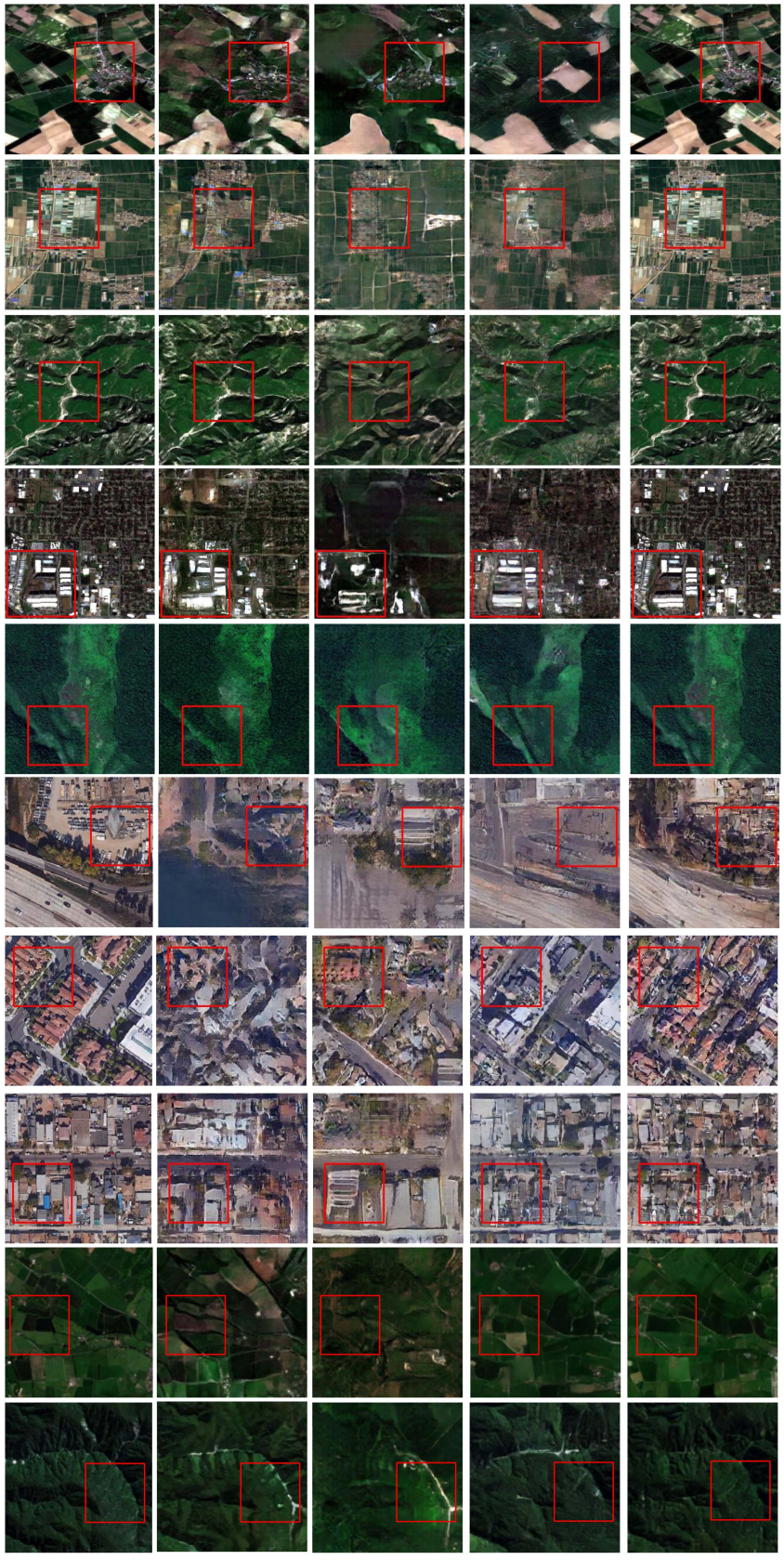

Fig. 8. Zoom-in results of different methods for S2OIT on the SEN1-2, QXS-SAROPT, and SEN12MS datasets. (a) Real optical image, (b)–(e) represent the result of Pix2pixHD, ASAP-Net, BBDM, and our CM-diffusion, respectively. The regions of interest are marked with red boxes, the first five rows of images are from the SEN1-2 dataset, images in the sixth to eighth rows are from the QXS-SAROPT dataset and the others from the SEN12MS dataset.

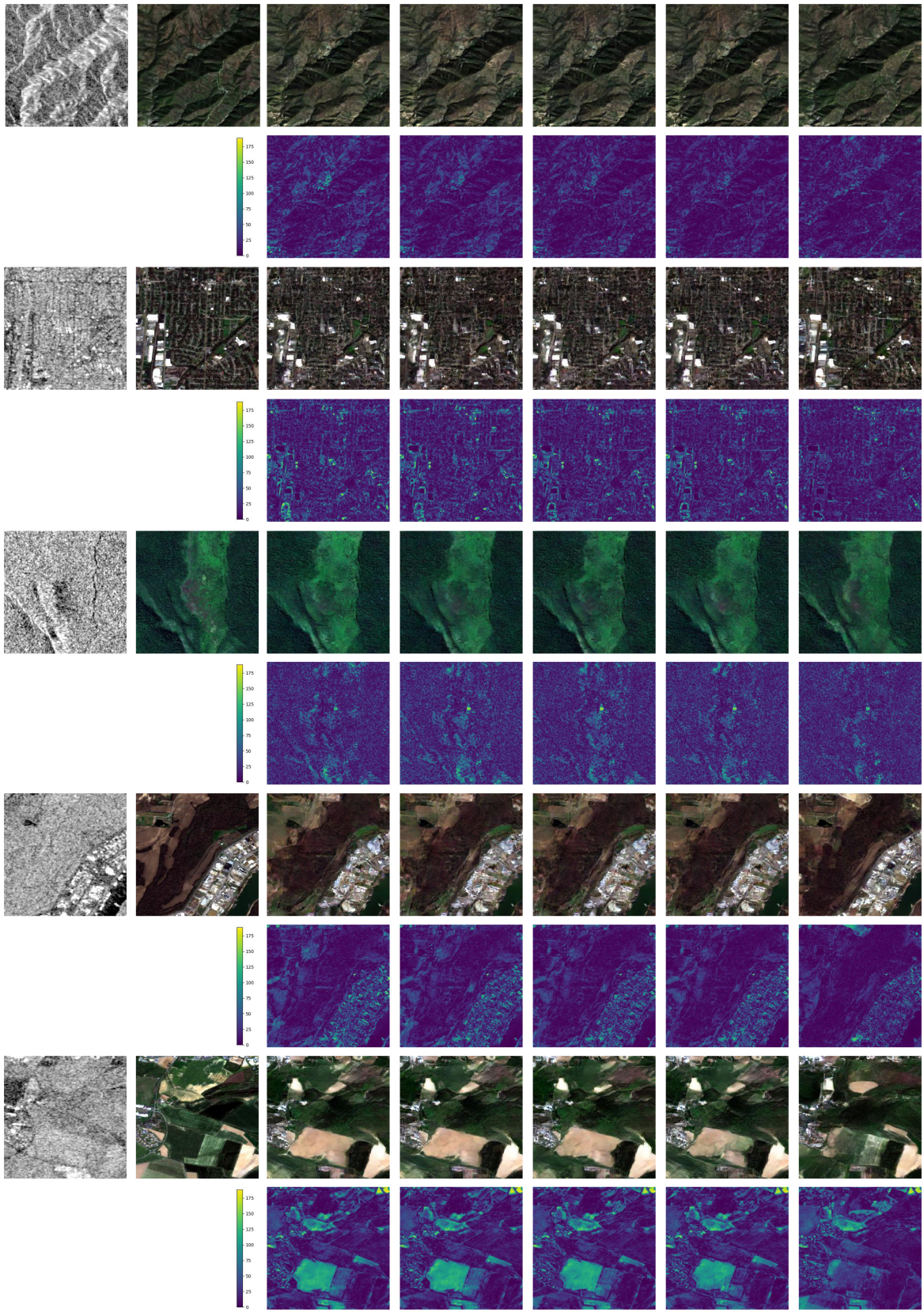

Fig. 9. Visualization results of different α for S2OIT on the SEN1-2 dataset. (a) Original SAR image, (b) the real optical image, (c)–(g) represent the results when the values of α are 2, 1, 0.5, 0.05, and 0.1, respectively. The residual results are under the corresponding generated optical images.

TABLE IV
DIFFERENT COMPONENTS IN THE PROPOSED METHOD ON DIFFERENT DATASETS

	SEN1-2			QXS-SAROPT			SEN12MS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline (*DDPM*)	14.92	0.273	0.453	15.06	0.301	0.492	17.63	0.439	0.463
+ CFE	15.36	0.302	0.445	15.13	0.315	0.475	17.99	0.451	0.450
+ CA-BBDS	15.54	0.312	0.419	15.30	0.332	0.470	18.27	0.458	0.409
+ all (*CM-Diffusion*)	**15.86**	**0.351**	**0.378**	**15.66**	**0.355**	**0.450**	**18.46**	**0.464**	**0.404**

↑ denotes bigger and better, ↓ denotes smaller and better.
The bold indicates optimal.

TABLE V
COMPARISON RESULTS OF DIFFERENT SAMPLING STEPS

Sampling Steps	PSNR↑	SSIM↑	LPIPS↓
20	15.22	0.302	0.413
50	15.45	0.318	0.392
100	15.54	0.324	0.389
200	15.86	**0.351**	**0.378**
300	15.86	0.350	0.379
500	15.87	0.349	0.381
1000	**15.88**	0.348	0.382

↑ denotes bigger and better, ↓ denotes smaller and better.
The bold indicates optimal.

TABLE VI
COMPARISON RESULTS OF DIFFERENT VARIANCE CONTROL FACTOR α

α	PSNR↑	SSIM↑	LPIPS↓
0.05	15.78	0.349	**0.371**
0.1	**15.86**	**0.351**	0.378
0.5	15.81	0.326	0.420
1	15.54	0.312	0.419
2	15.36	0.302	0.445

↑ denotes bigger and better, ↓ denotes smaller and better.
The bold indicates optimal.

composite metrics much, but rather has a negative impact. As a result, we select 200 sampling steps based on the results of three metrics.

The variance control factor's values determine the maximum variance difference in the image diffusion process. In the generation task, controlling the maximum variance of the diffusion model can control the image's diversity. However, in the image translation task, the output must be as similar as possible to the target image. Therefore, this article aims to reduce the diversity of the generated image to ensure a stable result from the network. Table VI shows the evaluation results for each image evaluation metric at various variance control factors α, and Fig. 9 shows the visualization results of different α for S2OIT on the SEN1-2 dataset. The results show that the evaluation metrics increase as α decreases when $\alpha > 0.1$. However, after $\alpha < 0.1$, the evaluation metrics no longer consistently increase and instead fluctuate across different metrics. This indicates that a variance control factor that is too small can lead to network instability. The visualization results are consistent with metric results, the residual is minimized when $\alpha = 0.1$, indicating that the generated optical image when $\alpha = 0.1$ is closest to the real optical image. Based on the results of various evaluation metrics, we finally select $\alpha = 0.1$.

V. DISCUSSION

Due to the designed CFE and CA-BBDS modules, our proposed CM-Diffusion has the ability to learn color correlation and direct mapping between the SAR and optical image domains, which cannot only memorize the color information of complex scenarios, but also distinguish the color details with small differences, and thus aiding in the more efficient interpretation of SAR images with complex terrain scenes, particularly in building-intensive and color-rich scenes. Three evaluation metrics also show that our proposed CM-Diffusion performs significantly better than the other comparative methods. However, since diffusion models require significant computational resources and time for training and sampling, although accelerated sampling methods are also used, they are still not fast enough compared to one-step generation models such as GANs. Therefore, an attempt can be made to further increase the sampling rate while maintaining the quality of images generated by the diffusion models to make the network more efficient. In addition, future research can also 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 assist in SAR image interpretation.

VI. CONCLUSION

In this article, we propose a S2OIT framework based on the diffusion models called CM-Diffusion, which has the promising originality of color memory and distinct domain adaptability. Our CM-Diffusion has innovatively designed CA-BBDS and CFE modules for color correlation and direct mapping between SAR and optical image domains. The CA-BBDS learns the translation between the SAR and optical image domains directly through the bidirectional diffusion process. The color attention mechanism fused with CFE module is added to the backward denoising process to improve the fusion of color correlation between the two domains. Comprehensive experiments demonstrate the effectiveness of our proposed CM-Diffusion, thus our CM-Diffusion can provide a new general solution framework based on diffusion models for the S2OIT task.

REFERENCES

[1] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617515.
[2] J. Lian, S. Mei, S. Zhang, and M. Ma, "Benchmarking adversarial patch against aerial detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634616.

[3] 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.
[4] 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.
[5] Y. Chen and L. Bruzzone, "Self-supervised SAR-optical data fusion of Sentinel-1/-2 images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5406011.
[6] 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, 2022, Art. no. 4105309.
[7] 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.
[8] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimedia*, vol. 24, pp. 3859–3881, 2021.
[9] 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.
[10] 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, 2022, Art. no. 3512605.
[11] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, vol. 2, pp. 2672–2680.
[12] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[13] 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.
[14] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.
[15] X. Bai, X. Pu, and F. Xu, "Conditional diffusion for SAR to optical image translation," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 4000605.
[16] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 2256–2265.
[17] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 8780–8794.
[18] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.
[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.
[20] 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.
[21] S. Mei, R. Jiang, M. Ma, and C. Song, "Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600713.
[22] 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.
[23] M. Huang et al., "The QXS-SAROPT dataset for deeplearning in SAR-optical data fusion," 2021, *arXiv:2103.08259*.
[24] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.
[25] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, 2008.
[26] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 2366–2369.
[27] 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.
[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[29] 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.
[30] T. R. Shaham, M. Gharbi, R. Zhang, E. Shechtman, and T. Michaeli, "Spatially-adaptive pixelwise networks for fast image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14882–14891.
[31] 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.
[32] 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.
[33] 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.
[34] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "Unit-DDPM: Unpaired image translation with denoising diffusion probabilistic models," 2021, *arXiv:2104.05358*.
[35] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22500–22510.
[36] Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6027–6037.
[37] Z. Wang, J. Wang, Z. Liu, and Q. Qiu, "Binary latent diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22576–22585.
[38] S. Mei, Y. Geng, J. Hou, and Q. Du, "Learning hyperspectral images from RGB images via a coarse-to-fine CNN," *Sci. China Inf. Sci.*, vol. 65, 2022, Art. no. 152102.
[39] D. Tan et al., "Serial GANs: A feature-preserving heterogeneous remote sensing image transformation model," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3968.
[40] 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, 2022, Art. no. 5236115.
[41] 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, 2022, Art. no. 5233812.
[42] 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, 2022, Art. no. 5621211.
[43] 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, 2022, Art. no. 4516305.
[44] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12873–12883.
[45] A. F. Peters, P. Peters, and K. Littlewood, "The color thief: A family's story of depression," Albert Whitman & Company, pp. 1–24, Sep. 2015.

Zhe Guo (Member, IEEE) 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.

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 M.S. degree in information and communication engineering with the School of Electronics and Information, Northwestern Polytechnical University. Her 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 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.

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.

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 October 2007 to October 2008. He is currently a Professor with the School of Electronics and Information, Northwestern Polytechnical University. His research interests include hyperspectral remote sensing 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 (JSTARS) in 2019, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2022. He is an Associate Editor for IEEE TGRS and IEEE JSTARS, the Guest Editor for Remote Sensing, and the reviewer for more than 30 international famous academic journals.