Cloud-EGAN: Rethinking CycleGAN From a Feature Enhancement Perspective for Cloud Removal by Combining CNN and Transformer

Xianping Ma, *Student Member, IEEE*, Yiming Huang, Xiaokang Zhang, *Member, IEEE*, Man-On Pun, *Senior Member, IEEE*, and Bo Huang

***Abstract*—Cloud cover presents a major challenge for geoscience research of remote sensing images with thick clouds causing complete obstruction with information loss while thin clouds blurring the ground objects. Deep learning (DL) methods based on convolutional neural networks (CNNs) have recently been introduced to the cloud removal task. However, their performance is hindered by their weak capabilities in contextual information extraction and aggregation. Unfortunately, such capabilities play a vital role in characterizing remote sensing images with complex ground objects. In this work, the conventional cycle-consistent generative adversarial network (CycleGAN) is revitalized from a feature enhancement perspective. More specifically, a saliency enhancement (SE) module is first designed to replace the original CNN module in CycleGAN to re-calibrate channel attention weights to capture detailed information for multi-level feature maps. Furthermore, a high-level feature enhancement (HFE) module is developed to generate contextualized cloud-free features while suppressing cloud components. In particular, HFE is composed of both CNN- and transformer-based modules. The former enhances the local high-level features by employing residual learning and multi-scale strategies, while the latter captures the long-range contextual dependencies with the Swin transformer module to exploit high-level information from a global perspective. Capitalizing on the SE and HFE modules, an effective Cloud-Enhancement GAN, namely Cloud-EGAN, is proposed to accomplish thin and thick cloud removal tasks. Extensive experiments on the RICE and the WHUS2-CR datasets confirm the impressive performance of Cloud-EGAN.**

***Index Terms*—Cloud removal, cycle-consistent generative adversarial network (CycleGAN), feature enhancement, remote sensing images, transformer.**

Manuscript received 4 April 2023; revised 10 May 2023; accepted 23 May 2023. Date of publication 2 June 2023; date of current version 8 June 2023. This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1800800, in part by the Basic Research Project under Grant HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Shenzhen Outstanding Talents Training Fund 202002, Guangdong Research Projects under Grant 2017ZT07X152 and Grant 2019CX01X104, in part by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001, in part by the National Natural Science Foundation of China under Grant 41801323, and in part by the National Key Research and Development Program of China under Grant 2020YFA0714003. *(Xianping Ma and Yiming Huang contributed equally to this work.) (Corresponding authors: Xiaokang Zhang; Man-On Pun.)*

Xianping Ma, Yiming Huang, and Man-On Pun are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: xianpingma@link.cuhk.edu.cn; 222012014@link.cuhk.edu.cn; simonpun@cuhk.edu.cn).

Xiaokang Zhang is with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: natezhangxk@gmail.com).

Bo Huang is with the Department of Geography, The University of Hong Kong, Hong Kong SAR 999077, China (e-mail: bohuang@cuhk.edu.hk).

Digital Object Identifier 10.1109/JSTARS.2023.3280947

I. Introduction

EARTH observation technology has facilitated the acquisition of remote sensing images. These images have been successfully used to extract land surface information in many critical applications, including object detection [1], [2], [3], scene classification [4], [5], [6], and semantic segmentation [7], [8], [9], [10]. However, such optical satellite images are inevitably susceptible to the atmospheric and illumination conditions, which incurs degradation in image quality. In particular, remote sensing images commonly suffer from the contamination of cloud layers, significantly diminishing the signal quality obtained by satellite sensors. Specifically, the cloud layers heavily reduce the visibility and saturation of images, hindering the subsequent image applications [11]. While thin-cloud-covered regions still exhibit limited ground features, the contextual information beneath thick clouds is completely lost. Compared with natural digital images, remote sensing images contain more complex spatial structures and richer spectral information for ground object characterization, making cloud removal more challenging. Therefore, the development of efficient signal processing algorithms is strongly desired to accurately recover the genuine land surface information from remote sensing images distorted by cloud layers. In the literature, existing cloud removal methods can be classified into two approaches, namely conventional methods based on hand-crafted features and deep learning (DL)-based methods [12], [13], [14], [15], [16], [17], [18].

Conventional methods, such as multitemporal dictionary learning (MDL) [19], thin cloud removal using homomorphic filter (TCHF) [20], and signal transmission principles and spectral mixture analysis (ST-SMA) [21], require hand-crafted features to estimate the cloud distribution. In particular, MDL learned dictionaries of cloud-covered and cloud-free regions separately in the spectral domain whereas TCHF utilized a classic homomorphic filter in the frequency domain. Furthermore, ST-SMA was developed based on signal transmission and spectral mixture analysis. Despite their many advantages, these methods were designed for thin cloud removal while overlooking the thick

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

cloud scenarios. Moreover, their feasibility and performance are typically limited by irregular cloud distribution and the choice of hand-crafted features.

Driven by the rapid development of DL techniques, DL-based cloud removal methods have attracted substantial research attention, owing to the superior performance of DL models in mining representative features from remote sensing images [22]. Most existing DL-based cloud removal methods in the literature were built upon convolutional neural networks (CNNs) by exploiting abstract and conceptual representations of remote sensing images. Generally speaking, DL-based networks for cloud removal can be divided into two categories, namely the pure encoder–decoder methods [11], [23], [24] and the generative adversarial networks (GAN)-based networks [12], [25], [26], [27], [28], [29], [30]. For the pure encoder–decoder networks, multiscale features-CNN [23] explored the multiscale high-level features to detect thin-cloud, thick-cloud, and no-cloud pixels simultaneously while residual learning and channel attention mechanism [11] integrated residual connection with a channel attention mechanism to capture details in different convolutional layers. Furthermore, conditional variational autoencoders (CVAE) [24] applied a probabilistic graphical model with CVAE to restore cloud-free images according to the image degradation process. The abovementioned encoder–decoder models employ the encoder to extract enriched features from remote sensing images, while the decoder is exploited to interpret abstract information before recovering the detailed information of cloud-free images. However, these methods are handicapped by their weak feature representation capability of CNNs. As a result, additional efforts are required to enhance the feature representation capability of CNNs to generate high-quality cloud-free images.

Similar to the encoder–decoder methods, the GAN-based models also consist of two parts, i.e., the generator and discriminator [31]. Owing to its remarkable capability of modeling the relationship between input and output data, GAN has gained tremendous popularity in computer vision. For the cloud removal task, conditional GAN (cGAN) [25] employed a simple UNet-based structure as the generator while PatchGAN [32] as the discriminator. Furthermore, a hybrid loss function using the structural similarity (SSIM) loss [33] was designed to improve the SSIM of the generated images with the ground truth. Recently, spatial attention GAN (SpAGAN) [27] was proposed to remove clouds by integrating local-to-global spatial attention to the generator whereas MSDA-CR [29] proposed a grid network based on cloud-distortion-aware representation learning to model the effects of cloud reflection and transmission. In addition, AMGAN-CR [30] generated attention maps through an attentive recurrent network and employed an attentive residual network to remove clouds according to the attention maps. These methods have improved the GAN-based frameworks by enhancing the encoder or loss function design through a single-directional mapping, i.e., from cloudy images to cloud-free images.

Recently, the cycle-consistent GAN (CycleGAN) model [34] has been widely applied to transfer image styles. CycleGAN attempts to learn a bidirectional mapping between domains while incorporating cycle-consistency loss and identity loss to effectively retain the color composition and texture. CloudGAN [12] introduced CycleGAN into cloud removal to learn the mapping of feature representations between cloudy images and their corresponding cloud-free images in a cyclic structure. In the cloud removal task, it is also necessary to learn global color composition and texture outside the cloud area before predicting the objects under the cloud in the forward process. The reverse stage in the cycle process can promote the learning of these global representations in the forward process by restoring the original cloud map. However, it suffers from blurred edges due to its straightforward encoding structure and the lack of modeling channel and spatial relationships. On this basis, SAR-to-optical image translation using SSIM and perceptual loss-based CycleGAN [26] introduced the least squares loss function [35] into the CycleGAN to improve its training stability in image translation. Furthermore, multimodal GAN (MMGAN) [28] was developed to generate multiple most likely cloud-free outputs before selecting the best generated cloud-free images through a perception-based image quality evaluator. Despite their many advantages, these methods suffer from a poor performance in reconstructing detailed features of remote sensing images as they are straightforward extensions from models originally devised for natural images. Compared with natural scene images, remote sensing images exhibit more severe spectral heterogeneity and more complex spatial relationships of ground objects [36], [37]. Typically, undesired cloud layers have various thicknesses, and images are acquired under different lighting conditions [38]. As a result, the performance of those image restoration models developed for natural scene images is usually poor if directly applied to cloud removal. Furthermore, it is challenging for these models to handle large-scale cloud removal tasks due to their prohibitively expensive computational complexity.

To improve the representation capability of CNNs and GAN with long-range contextual information, the newly developed transformer has been introduced into the cloud removal tasks. Empowered by its nonlocal attention mechanism, the transformer can establish long-range dependencies with impressive scalability [39], [40]. For instance, SAR-enhanced cloud removal with global-local fusion [15] added Swin transformer layer [41] after each convolutional layer for cross-window feature interaction. CloudTran [42] replaced the CNN-based encoder with an axial transformer [43] to estimate the low-resolution cloud-free images. However, the transformer is only regarded as a feature extractor to exploit global information while lacking the capability to fully extract enriched local features. Compared with the transformer, CNN exploits and aggregates enriched local features using the local receptive fields in the convolutional layers [3], [10]. One trivial approach to take advantage of both transformer and CNN is to directly construct a dual-branch encoder to extract global and local information by transformer and CNN, respectively [44], [45], [46], [47]. More recently, the authors in [48] and [49] proposed to use CNN to extract multiscale features while exploring the ability of transformer to enhance these multiscale features. In contrast, Fang et al. [50] further integrated the Swin Transformer layers and the convolutional layers by exploiting spatial attention after each Swin Transformer layer. However, all methods aforementioned failed to explore the potential enhancement of high-level semantic features provided by exploiting the synergy of CNN

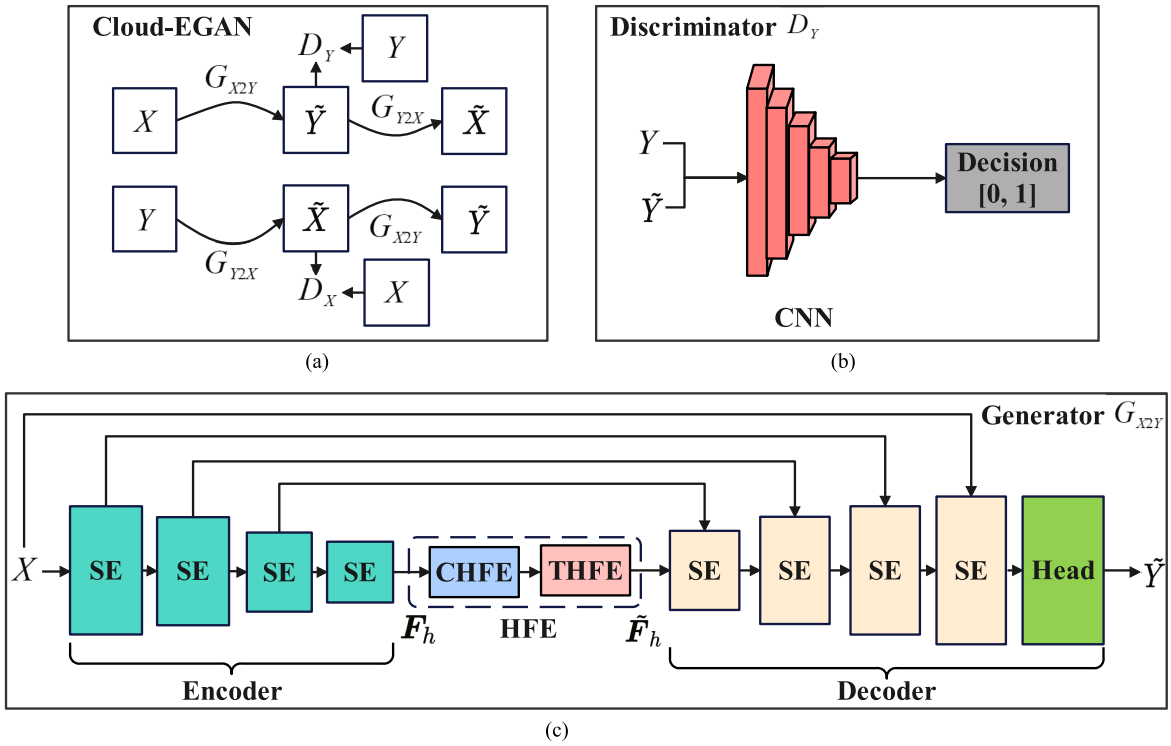

Fig. 1. (a) Overview of Cloud-EGAN framework. G_{X2Y} and G_{Y2X} are the two generators and D_X and D_Y are the two discriminators. $X, \tilde{X}, Y, \tilde{Y}$ represent the authentic cloudy images, the generated cloudy images, the authentic cloud-free images, and the generated cloud-free images, respectively. (b) Discriminator follows the PatchGAN structure. (c) Generator is based on the UNet framework with different output sizes from each SE block. Note that the two generators, G_{X2Y} and G_{Y2X}, are designed with the same structure. Similarly, the two discriminators, D_X and D_Y, have the same structure.

and transformer. Thus, it is of great practical interest to investigate how to fill this gap by combining CNN and transformer in the cloud removal task.

Motivated by the aforementioned challenges, this work introduces a CycleGAN-based model for thin and thick cloud removal from two different enhancement perspectives. First, the backbone is enhanced by a saliency enhancement (SE) module to extract hierarchical discriminant features with more saliency. Furthermore, in sharp contrast to the existing models that utilize CNN to enhance high-level features [6], [9], [23], this work proposes to explore enriched high-level features by jointly exploiting CNN and transformer. The main contributions of this work can be summarized as follows:

1) An SE module is utilized to generate enhanced hierarchical feature maps derived from each convolutional block by recalibrating the attention weights of feature channels. As a result, cloud-covered components and blurred edges are reduced;
2) A high-level feature enhancement (HFE) module is devised between the encoder and the decoder to effectively explore and aggregate high-level features. Specifically, HFE is composed of a CNN-based HFE (CHFE) module and a transformer-based HFE (THFE) module. CHFE is designed to exploit high-level local features to harvest sufficient detailed information while THFE long-range contextual information. CHFE and THFE are integrated under the cloud-enhancement GAN (Cloud-EGAN) framework to retain the global features of the restored cloud-clear images;
3) Extensive experimental results on the RICE and WHUS2-CR datasets verify the superiority of Cloud-EGAN in segregating clouds and preserving high-quality land surface information.

The rest of this article is organized as follows. Section II elaborates on the proposed model while extensive experimental results are presented and analyzed in Section III. Finally, Section IV concludes this article.

II. Methodology

In this work, a CycleGAN-based architecture with SE and HFE modules in the generator is proposed to extract and aggregate enhanced local and global features from remote sensing images. In the following, an overview of the proposed Cloud-EGAN is presented before each of its key components is elaborated. Finally, hybrid loss functions employed in the proposed model are devised.

A. Framework

As depicted in Fig. 1(a), the proposed Cloud-EGAN is developed based on CycleGAN that consists of two generators G_{X2Y} and G_{Y2X} and two discriminators D_X and D_Y. More specifically, for a supervised cloud removal task, the authentic cloudy image X serves as the input to the generator G_{X2Y} to reconstruct the predicted cloud-free image $\tilde{Y}$ that is then

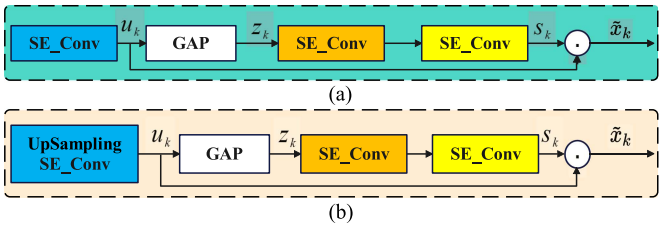

Fig. 2. Architecture of the SE module in (a) encoding and (b) decoding.

discriminated by D_Y with the authentic cloud-free image Y. Meanwhile, according to the cyclic consistency principle, the generator G_{Y2X} is employed to generate the cloudy image $\tilde{X}$ from $\tilde{Y}$. The same operation is performed on the input Y in the Cloud-EGAN.

As illustrated in Fig. 1(b), the discriminators D_X and D_Y adopt a PatchGAN structure with stacked hierarchical convolutional blocks to determine the authenticity of $\tilde{Y}$. Furthermore, the generator is developed based on an UNet architecture [51] by capitalizing on symmetrical concatenations between an encoder and a decoder, as shown in Fig. 1(c). Specifically, the generator combines the SE and HFE modules while SE exploits hierarchical features by reassigning attention weights to feature maps at each level. The resulting high-level feature maps are then fed into the HFE module to further enhance feature representation through the combination of CNN and transformer. After that, a convolutional prediction head is utilized at the end of the generator to recover cloud-clear images. More details about the SE and HFE modules will be elaborated in the following sections.

B. Saliency Enhancement

Following the classical channel attention mechanism [52], the SE module adaptively exploits more salient features from remote sensing images at multiple feature levels by assigning learnable attention weights to feature channels. As a result, SE can enhance information restoration from heavily cloudy regions and generate high-quality cloud-free features.

Fig. 2(a) illustrates the encoding process in which $u_k \in \mathbb{R}^{D_k \times H_k \times W_k}$ denotes the kth level feature map generated from the first convolutional block (SE_Conv), where D_k is the channel dimension, $H_k = H/2^k$ and $W_k = W/2^k$. Furthermore, a global average pooling (GAP) layer as a channel descriptor is applied to exploit enriched features and produce output z_k

$$z_k = G(u_k) = \frac{1}{H_k \times W_k} \sum_{i=1}^{H_k} \sum_{j=1}^{W_k} u_k(i,j) \tag{1}$$

where G stands for the GAP function. After that, two 1×1 SE_Conv blocks are utilized to compute the attention weights through convolution operations with output $s_k \in \mathbb{R}^{D_k \times 1 \times 1}$ being given by

$$s_k = \delta(\mathbf{W}_2(\mathbf{W}_1(z_k))) \tag{2}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are parameters of the two convolutional blocks and $\delta(\cdots)$ is the sigmoid function. Finally, the SE output denoted by $\tilde{x}_k \in \mathbb{R}^{D_k \times H_k \times W_k}$ is derived by multiplying s_k with u_k

$$\tilde{x}_k = s_k \odot u_k \tag{3}$$

where $\odot$ represents the point multiplication operation.

The decoding process depicted in Fig. 2(b) is similar to Fig. 2(a) with the convolutional block (SE_Conv) being replaced by an upsampling SE_Conv.

C. High-Level Feature Enhancement

The HFE module is designed to learn enriched high-level local and nonlocal features by combining CHFE and THFE, as shown in Fig. 3. As a result, it is beneficial to further characterize cloud-free representations and propagate contextual information across the feature maps from a global perspective, which can maintain the spatial structure of the restored features identical to the ground truth.

More specifically, a residual learning module [53] and a dilated convolutional module [54] are used in CHFE to process high-level features in parallel. In particular, high-level features $\boldsymbol{F}_h \in \mathbb{R}^{D \times \frac{H}{16} \times \frac{W}{16}}$ are fed into the residual learning module containing three successive residual blocks named HFE_ResConv to extract critical ground information while reducing the feature discrepancy between cloud-covered and cloud-free images. Meanwhile, $\boldsymbol{F}_h$ is passed through a convolutional block with residual structure named HFE_Conv, and three dilated convolutional blocks named HFE_DilatedConv with different dilation rates to exploit multiscale contextual information while alleviating cloud-covered features. After that, the concatenated outputs are further enhanced through an HFE_Conv block to restore the original feature size. Finally, the outputs of the residual learning module and dilated convolutional module are added together to form refined feature maps $\boldsymbol{F} \in \mathbb{R}^{D \times \frac{H}{16} \times \frac{W}{16}}$.

Following the approach of the classical Swin transformer [41], THFE splits $\boldsymbol{F}$ into nonoverlapping patches in the patch partition module before projecting the patches to an arbitrary dimension $\hat{D}$ using a linear embedding layer. The patches are then fed into a successive Swin transformer block and a patch merging layer to generate higher level feature representations. More specifically, as depicted in Fig. 3(b), each successive Swin transformer block consists of the residual architecture, four layer-normalization (LN) layers, a window-based multihead self-attention (WMSA) module, a shifted WMSA (SWMSA) module, and two multilayer perceptron (MLP) layers with GELU function.

The operation of successive Swin transformer blocks is shown in Fig. 3(b). For each head of the WMSA and SWMSA, the input features $\boldsymbol{F}_S$ are fed into the Swin transformer block to calculate the multihead self-attention (MSA) as follows:

$$Q_S = \boldsymbol{F}_S W_Q, K_S = \boldsymbol{F}_S W_K, V_S = \boldsymbol{F}_S W_V \tag{4}$$

and

$$\mathrm{Att}(\boldsymbol{F}_S) = \phi\left(\frac{Q_S {K_S}^T}{\sqrt{d}} + B_S\right) V_S \tag{5}$$

where Q_S, K_S, and V_S denote the projected *query*, *key*, and *value* features, respectively while W_Q, W_K, and W_V the corresponding parameter metrics. Furthermore, B_S is the learnable relative position embedding term in the Swin transformer

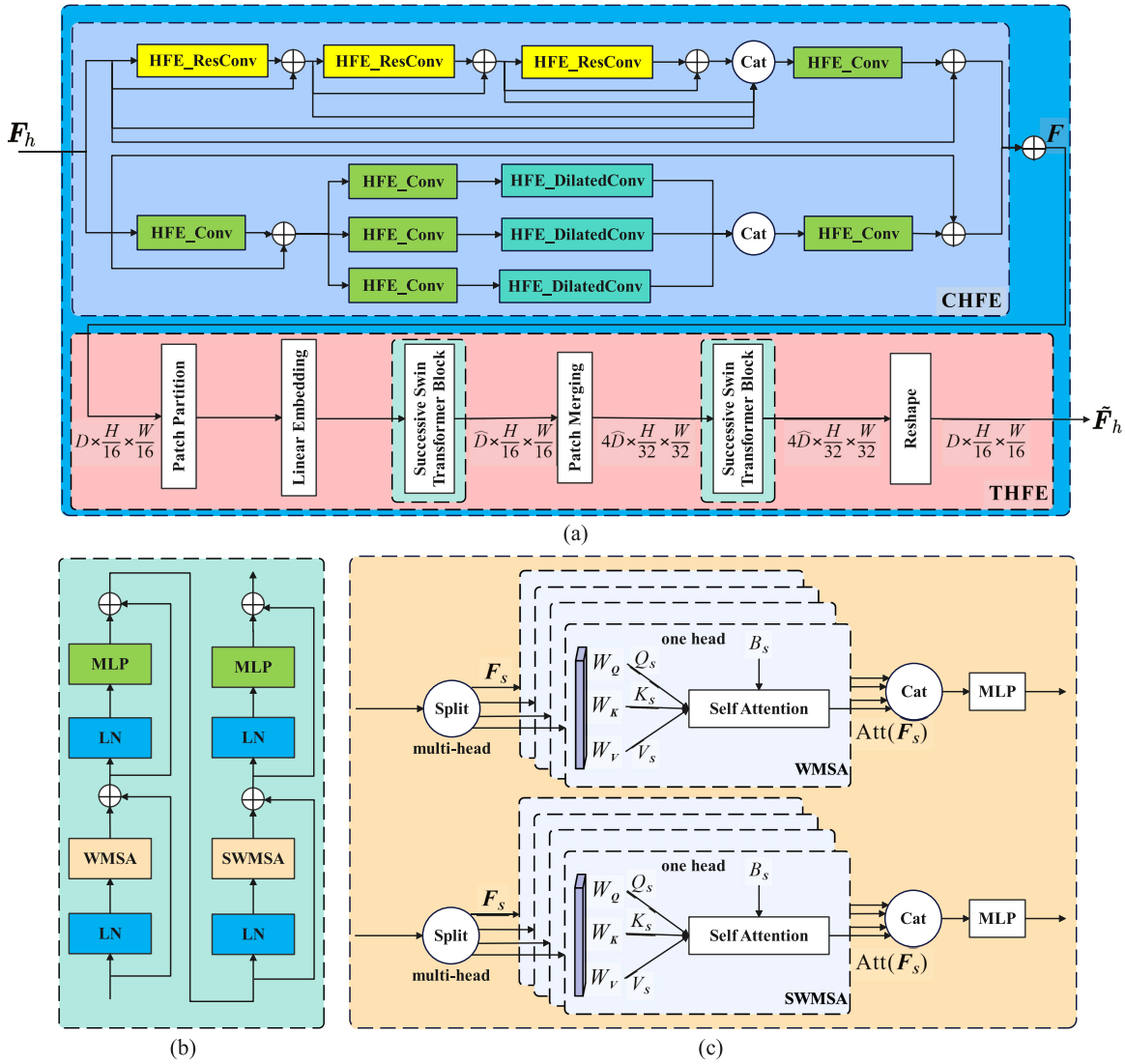

Fig. 3. Illustrations of (a) the HFE module, (b) successive Swin Transformer block, and (c) MSA, LN and MLP represent the LN layer and MLP layer, respectively. WMSA and SWMSA are MSA modules with common and shifted windowing configurations, respectively. (a) HFE. (b) Swin transformer. (c) Multihead self-attention.

whereas $\mathrm{Att}(\boldsymbol{F}_S)$ represents the output of self-attention for each head. In addition, $\phi(\cdot)$ is the softmax function and $d = \hat{D}/4$ is the channel dimension for each head.

After that, the features of each 2×2 neighboring patches generated by the Swin transformer block are concatenated by the patch merging layer. We denote by $H/32$, $W/32$, and $4\hat{D}$ the height, width, and channel after the patch merging layer, respectively. Finally, after two Swin transformer blocks and the reshape operation to maintain the same size as the input $\boldsymbol{F}_h$, the output of the HFE module $\tilde{\boldsymbol{F}}_h \in \mathbb{R}^{D \times \frac{H}{16} \times \frac{W}{16}}$ can be obtained.

D. Loss Functions

In this work, a novel hybrid loss function comprising the adversarial loss L_{adv}, the cycle consistency loss L_{cyc}, the perceptual loss L_{per}, and the identity loss L_{id} is introduced to guide the training of our proposed model. It is notable that L_{adv} is utilized to train both generators and discriminators, while L_{cyc}, L_{per}, and L_{id} are employed for training the generators. The expression of the hybrid loss function L can be formulated as follows:

$$L = L_{\text{adv}} + \lambda_{\text{cyc}} L_{\text{cyc}} + \lambda_{\text{per}} L_{\text{per}} + \lambda_{\text{id}} L_{\text{id}} \tag{6}$$

where λ_{cyc}, λ_{per}, and λ_{id} are adjustable weights of the three loss components. More details about each loss function are provided in the following sections.

1) Adversarial Loss: The adversarial loss aims to make the reconstructed cloud-free images close to the corresponding ground truth. Adopting a structure similar to the classical CycleGAN, we define the adversarial loss as

$$\begin{aligned} L_{\text{adv}}^{X2Y} &= E_{y \sim P_{\text{data}}(y)}[\log D_Y(y)] \\ &\quad + E_{x \sim P_{\text{data}}(x)}[\log(1 - D_Y(G_{X2Y}(x))] \end{aligned} \tag{7}$$

$$L_{\text{adv}}^{Y2X} = E_{x\sim P_{\text{data}}(x)}[\log D_X(x)] + E_{y\sim P_{\text{data}}(y)}[\log(1 - D_X(G_{Y2X}(y))] \quad (8)$$

where x and y are the input cloudy and cloud-free image samples, respectively. Furthermore, $P_{\text{data}}(x)$ and $P_{\text{data}}(y)$ represent the distributions of cloudy and cloud-free images. The total adversarial objective L_{adv} is comprised of L_{adv}^{X2Y} and L_{adv}^{Y2X} used to train forward process and reverse process, respectively.

2) Cycle Consistency Loss: The cycle consistency loss measures the pixel-wise difference between the generated images and their corresponding ground truth. It is adopted to reduce blurring regions and keep the reconstructed images closer to the ground truth. The cycle consistency loss takes the following form:

$$L_{\text{cyc}} = E_{x\sim P_{\text{data}}(x)}\left[||G_{Y2X}(G_{X2Y}(x)) - x||_1\right] + E_{y\sim P_{\text{data}}(y)}\left[||G_{X2Y}(G_{Y2X}(y)) - y||_1\right] \quad (9)$$

where G_{X2Y} and G_{Y2X} are the two generators in the Cloud-EGAN and $||\cdot||_1$ stands for the $L1$-norm of the enclosed quantity.

3) Perceptual Loss: Based on the computation of losses in pixel colors and edges, the perceptual loss [55] is introduced to measure the consistency between convolutional outputs of the ground truth and the restored images obtained by a pretrained network, e.g., VGG19 pretrained on the ImageNet [56]. Moreover, the capability of extracting perceptual semantic features via the convolutional layers can be evaluated. Mathematically, the expression of the perceptual loss can be defined as

$$L_{\text{per}} = \sum_k \frac{1}{C_k H_k W_k}\left[E_{x\sim P_{\text{data}}(x)}||\phi_k(x') - \phi_k(x)||_1 + E_{y\sim P_{\text{data}}(y)}||\phi_k(y') - \phi_k(y)||_1\right] \quad (10)$$

where ϕ_k denotes the feature map extracted from the kth layer in the pretrained VGG19 network, and C_k, H_k, W_k denote the number of channels, height, and width of the kth feature map, respectively. Moreover, x' and x represent the pixel intensities in the original cloudy images and the generated cloudy images by Cloud-EGAN, respectively. Meanwhile, y' and y stand for the pixel intensities in the ground truth cloud-free images and the generated cloud-free images by Cloud-EGAN, respectively.

4) Identity Loss: The identity loss aims to retain the color consistency between the input and the output. For the cloud removal task, the clouds are expected to be eliminated in the generated cloud-free images and the cloud-free regions are expected to remain unchanged in texture details and color compositions. The proposed model can avoid color distortion in cloud-free regions by applying identity loss. It can be formulated as follows:

$$L_{\text{id}} = E_{x\sim P_{\text{data}}(x)}[||G_{X2Y}(x)) - x||_1] + E_{y\sim P_{\text{data}}(y)}[||G_{Y2X}(y) - y||_1]. \quad (11)$$

III. Experimental Results

In this section, experimental datasets will be first described. After that, the parameter settings and evaluation metrics are introduced before the comparisons with other DL-based models are reported and analyzed.

A. Datasets

In this section, the proposed model is evaluated on the RICE dataset [57] and the WHUS2-CR dataset [58]. Specifically, the RICE dataset comprises two subdatasets named RICE1 and RICE2. In particular, the RICE1 contains 500 pairs of cloud-covered and cloud-free images from Google Earth, with the ground resolution being 5 m/pixel. Most of the samples in RICE1 are thin clouds where the ground objects are mostly identifiable. In sharp contrast, the RICE2 dataset includes Landsat-8 images of 736 groups of the ground resolution 30 m/pixel. The images in this dataset contain abundant thick clouds, where the ground objects are hardly identifiable. Taking into account the large discrepancy in terms of cloud thickness and image resolution, we perform our evaluation on these two subdatasets separately. Furthermore, in sharp contrast to MSDA-CR [29] and CR-MSS [58] that utilize multispectral data as input, we mainly focus on visible (RGB) bands in our evaluation. This is because RGB images are more commonly available [11], [12], [28], [59]. However, we also perform supplement experiments to demonstrate that the proposed model can work well with multispectral data by exploiting both RGB and near-infrared (NIR) data.

The images in the RICE dataset are of size 512×512 pixels each. Moreover, the WHUS2-CR dataset involves 848 pairs of Sentinel-2 image patches of size 256×256 pixels. The acquisition time lag of the cloud-covered images and their corresponding cloud-free images is less than 10 days. Furthermore, 400 and 100 image pairs were chosen for training and testing in the RICE1 dataset, respectively. In addition, 589 and 147 pairs were adopted as the training and testing set in the RICE2 dataset, respectively. For the WHUS2-CR dataset, 679 pairs were obtained as training data, and the remaining 169 pairs were reserved for testing. Some typical samples in the RICE1, RICE2, and WHUS2-CR datasets are displayed in Fig. 4.

B. Implementation Details

In the generator of the Cloud-EGAN, four convolutional layers with a kernel size of 4×4 and stride of 2 are utilized in the encoder and decoder, with $\{32, 64, 128, 256\}$ channels for the former and $\{256, 128, 64, 32\}$ for the latter. After that, a convolutional layer with a kernel size of 4×4, a stride of 1 and 3 channels is utilized to restore the cloud-free images with the same size as the input. In the discriminator, four convolutional layers with a kernel size of 4×4 and a stride of 2 are exploited with $\{64, 128, 256, 512\}$ channels. Meanwhile, a convolutional layer with a kernel size of 4×4, a stride of 1, and a channel number of 1 is used to discriminate whether the generated cloud-free images are authentic or not. Notably, these convolutional layers are followed by the instance normalization [34] and the Leaky ReLU function [60] parameterized by 0.2, except for the classifier in the decoder and the discriminator.

In Cloud-EGAN, the learning rate α was initially set to 0.0001 before being decayed by half after every 20 epochs. Furthermore,

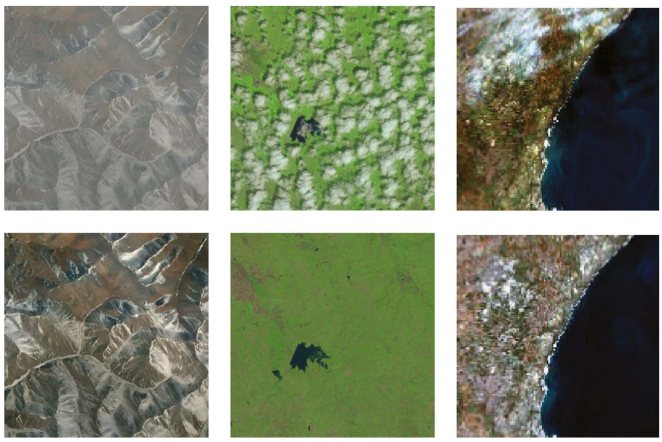

Fig. 4. Typical image samples in the (a) RICE1, (b) RICE2 and (c) WHUS2-CR datasets. The first and the second rows are cloud-covered images and cloud-free images, respectively.

The batch size was set to 4. In addition, the Adam optimizer [61] with default momentum parameters, i.e., $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was adopted. Finally, λ_{cyc}, λ_{per}, and λ_{id} in the loss function were set to 10, 1, and 9, respectively. All experiments were implemented on a single NVIDIA GeForce RTX 3090 GPU with 24-GB RAM.

The proposed Cloud-EGAN is compared against six state-of-the-art DL-based cloud removal methods, namely cGAN [25], CloudGAN [12], SpAGAN [27], CVAE [24], MSDA-CR [29], and MMGAN [28].

C. Metrics

Two widely used metrics, SSIM [62] and peak signal-to-noise ratio (PSNR) [63], were utilized for quantitative evaluation. Specifically, SSIM is expressed as

$$\text{SSIM} = \frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1} \frac{2\sigma_{xy} + C_2}{{\sigma_x}^2 + {\sigma_y}^2 + C_2} \tag{12}$$

where μ_x, σ_x, and σ_{xy} represent the average, variance, and covariance, respectively. C_1 and C_2 are constants for stabilizing the division with a weak denominator. A larger SSIM value stands for the greater similarity between the generated cloud-free images and ground truth, which indicates a higher quality of the generated cloud-free images. Moreover, PSNR is defined as

$$\text{PSNR} = 20 \log_{10} \frac{\text{MAX}_I}{\sqrt{\text{MSE}}} \tag{13}$$

where

$$\text{MSE} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \|I(i,j) - J(i,j)\|^2 \tag{14}$$

and MAX_I represents the possible maximum pixel value in the generated cloud-free images I. Moreover, the generated cloud-free image I and the corresponding ground-truth J are of size $M \times N \times 3$, and (i, j) represents the pixel index in I and J. A larger PSNR value represents less image distortion in the reconstructed cloud-free images.

Finally, we evaluated the computational complexity of the proposed method using the following metrics, namely the floating point operation count (FLOPs), the number of parameters (M), and the frames per second (FPS). More specifically, FLOP is used to evaluate the model complexity whereas M measures the memory requirement. In addition, FPS is used to evaluate the execution speed. For computationally efficient models, their FLOP and M should be small while FPS being large.

D. Performance Comparison

As illustrated in Fig. 5, the results obtained by Cloud-EGAN achieved lower spectral distortion and more significant SSIM with the ground truth in the thin-cloud-covered scenarios. Moreover, the results obtained by cGAN and CloudGAN suffered from much loss of texture details with some blurring areas while failing to thoroughly restore the land surface information in the generated cloud-free images. Compared with cGAN and CloudGAN, SpAGAN and MSDA-CR showed better results with more explicit texture details. However, some color distortions have been observed. As a result, the color information of the ground surface could not be fully restored. Finally, despite the fact that the results of CVAE and MMGAN achieved color compositions similar to the ground truth, some slightly-blurred edges were noticed in several patches.

In contrast, Cloud-EGAN performed best among all methods under evaluation in the thick-cloud-covered scenarios of the RICE2 dataset, as shown in Fig. 6. It generated images with better texture structures and color compositions. In comparison, cGAN and CloudGAN could not remove clouds thoroughly, which generated some edge-blurring and color-distortion areas in noncloudy regions. Moreover, the results obtained by SpAGAN exhibited severe loss of details since the structure information of ground scenarios could not be completely recovered. Furthermore, it was observed that the results of CVAE, MSDA-CR, and MMGAN showed more spatial features similar to the ground truth, though some slight color distortions were observed.

For the WHUS2-CR dataset, as depicted in Fig. 7, the color compositions and the texture details of the cloud-free images generated by Cloud-EGAN were more similar to the ground truth. In contrast, cGAN and CloudGAN showed the worst performance due to their limited feature extraction capability. Compared with cGAN and CloudGAN, SpAGAN and MSDA-CR obtained better results with more evident backgrounds and details, but some color distortion scenes remained in cloudless regions. Finally, the color tones in the results of CVAE and MMGAN were visually close to the ground truth. However, some contextual information was lost, and clouds were not segregated thoroughly in these models.

Quantitative results on the RICE1, RICE2, and WHUS2-CR datasets are shown in Table I. Cloud-EGAN achieved higher PSNR and SSIM values than other DL-based methods due to the

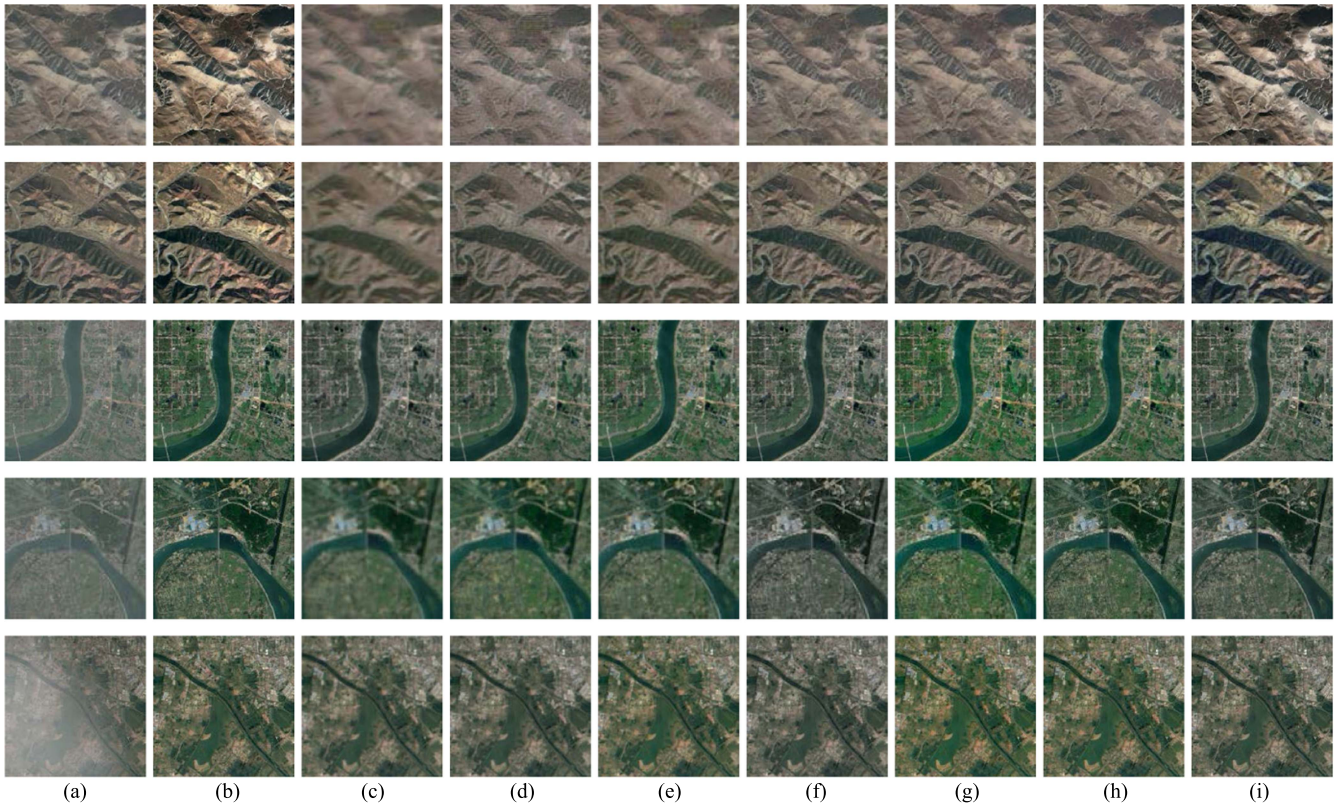

Fig. 5. Visual comparison of cloud removal results obtained by different models in thin-cloud-covered scenarios of the RICE1 dataset. (a) Cloudy images, (b) Ground Truth, (c) cGAN [25], (d) CloudGAN [12], (e) SpAGAN [27], (f) CVAE [24], (g) MSDA-CR [29], (h) MMGAN [28], and (i) Proposed Cloud-EGAN.

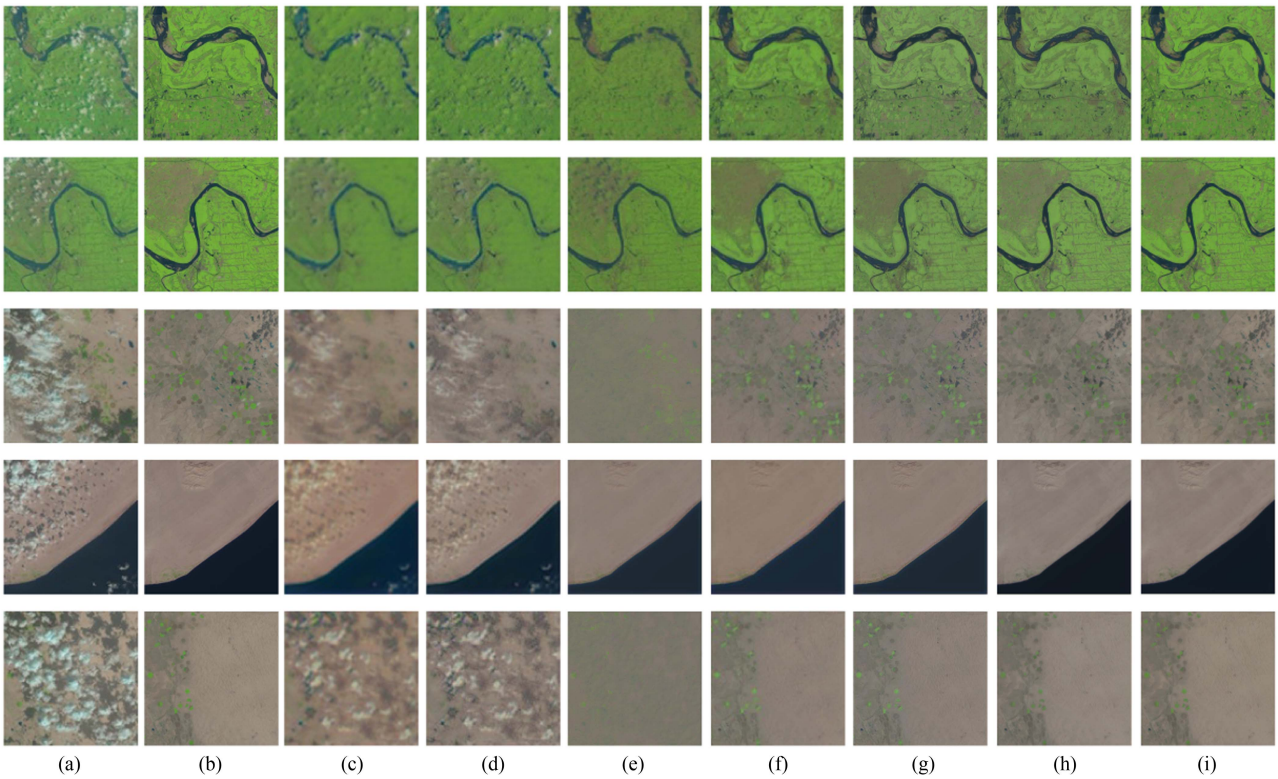

Fig. 6. Visual comparison of cloud removal results obtained by different models in thick-cloud-covered scenarios of the RICE2 dataset. (a) Cloudy images, (b) Ground Truth, (c) cGAN [25], (d) CloudGAN [12], (e) SpAGAN [27], (f) CVAE [24], (g) MSDA-CR [29], (h) MMGAN [28], (i) Proposed Cloud-EGAN.

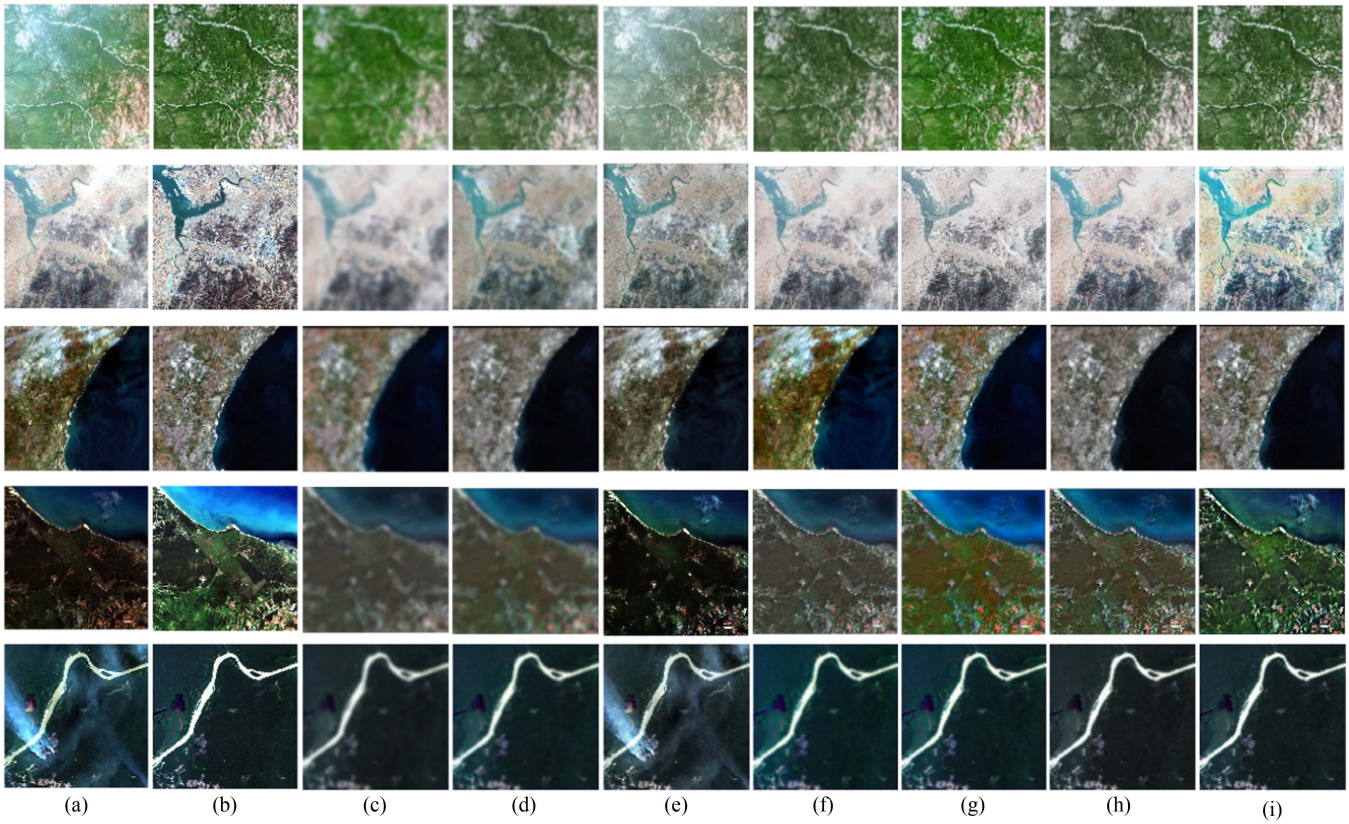

Fig. 7. Visual comparison of cloud removal results obtained by different models in the WHUS2-CR dataset. (a) Cloudy images, (b) Ground Truth, (c) cGAN [25], (d) CloudGAN [12], (e) SpAGAN [27], (f) CVAE [24], (g) MSDA-CR [29], (h) MMGAN [28], (i) Proposed Cloud-EGAN.

TABLE I
QUANTITATIVE RESULTS FOR PSNR AND SSIM VALUES

Method	RICE1		RICE2		WHUS2-CR	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
cGAN [25]	25.1276	0.7926	26.2374	0.8056	15.3472	0.7169
CloudGAN [12]	27.2568	0.8258	28.5561	0.8378	18.6826	0.7482
SpAGAN [27]	30.7125	0.8891	31.8679	0.8971	20.8764	0.7983
CVAE [24]	32.1247	0.9245	34.2258	0.9282	22.8274	0.8265
MSDA-CR [29]	34.2386	0.9352	34.9247	0.9447	23.7375	0.8357
MMGAN [28]	35.1253	0.9473	35.6712	0.9512	24.0889	0.8504
Cloud-EGAN	**37.8231**	**0.9751**	**38.1125**	**0.9792**	26.4125	0.8872
Cloud-EGAN*	-	-	-	-	**30.1342**	**0.9326**

The values in bold are the best.

exploration and aggregation of enriched local and global features in hierarchical and deep contextualized space. In particular, the results labeled as Cloud-EGAN* were generated with the proposed Cloud-EGAN using *four* input bands, i.e., RGB and NIR. It is evident from Table I that the proposed Cloud-EGAN can also work well in multispectral scene. Furthermore, it is shown that the NIR band could indeed further improve the performance of cloud removal. Therefore, the cloud-covered and cloud-free regions could be more accurately characterized, which aided in maintaining the recovered images close to the ground truth.

E. Ablation Study

1) Cycle-Consistence: In order to evaluate the necessity of cycle-consistence that requires two generators and discriminators, we conducted an ablation experiment as shown in the second line from the bottom in Table II, which is the result of the conventional GAN framework with only one generator-discriminator pair. The experiment results showed that the cycle-consistent mechanism enabled the generator to learn better global representations to promote the prediction of the ground objects of cloud-free areas.

TABLE II
QUANTITATIVE RESULTS WITH DIFFERENT MODULES FOR THE GENERATOR

SE	HFE		RICE1		RICE2		WHUS2-CR	
	THFE	CHFE	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
			34.1265	0.9428	35.2276	0.9473	23.3121	0.8274
✓			35.2347	0.9532	36.4357	0.9572	24.2938	0.8477
	✓		35.8213	0.9576	36.8224	0.9612	25.5376	0.8534
		✓	36.2466	0.9635	37.2453	0.9657	25.5489	0.8568
✓	✓		36.7124	0.9657	37.7623	0.9694	25.7124	0.8654
✓		✓	37.1326	0.9702	37.8312	0.9756	25.8682	0.8712
	✓	✓	37.4252	0.9723	37.8847	0.9779	26.0784	0.8805
✓	✓	✓	35.5481	0.9527	35.8126	0.9474	23.9174	0.8581
			37.8231	**0.9751**	**38.1125**	**0.9792**	**26.4125**	**0.8872**

The values in bold are the best.

TABLE III
QUANTITATIVE RESULTS WITH DIFFERENT LOSS FUNCTIONS

Dataset	original loss	perceptual loss	PSNR	SSIM
RICE1	✓		35.7374	0.9586
	✓	✓	**37.8231**	**0.9751**
RICE2	✓		37.1326	0.9612
	✓	✓	**38.1125**	**0.9792**
WHUS2-CR	✓		24.1658	0.8645
	✓	✓	**26.4125**	**0.8872**

The values in bold are the best.

TABLE IV
COMPARISON ON COMPUTATIONAL COMPLEXITY MEASURED BY A 256×256 INPUT ON A SINGLE NVIDIA GEFORCE RTX 3090 GPU

Model	Complexity (G)	Parameters (M)	Speed (FPS)
cGAN	70.548	6.968	1.143
Cloud-GAN	227.328	11.378	0.824
SPAGAN	**68.338**	**1.211**	**1.724**
CVAE	185.736	15.275	0.925
MSDA-CR	214.426	2.938	1.512
MMGAN	309.278	15.026	0.731
Cloud-EGAN	95.733	23.983	0.742

Bold values are the best.

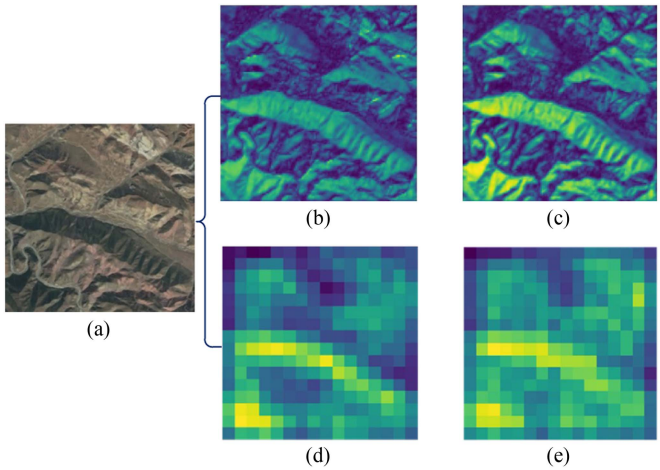

Fig. 8. Comparison of the feature maps via introducing SE and HFE in the proposed Cloud-EGAN. Note that the brighter regions are paid higher attention to during the training process, where more contextualized features will be exploited. (a) Original image. (b) Feature map of the output through the first convolutional block in the first SE module. (c) Feature map of the output through the first SE module. (d) Feature map of the input of HFE. (e) Feature map of the output through HFE.

2) Model Components: We compared the quantitative results by eliminating various modules in the proposed Cloud-EGAN framework, as shown in Table II. Note that the elimination of the SE module by only utilizing the convolutional layers followed by instance normalization and leaky ReLU function still follows the same UNet-based architecture as Fig. 1(c). Inspection of Table II reveals that integrating all feature enhancement modules resulted in the best performance in terms of both PSNR and SSIM. Accordingly, this confirmed the benefits of these modules in aggregating enriched contextualized features and restoring ground surface information sufficiently. Notably, SE can be further developed by other channel-based or spatial-based attention modules. We provided a unique perspective on using squeeze-and-excitation module [52] to enhance convolutional networks comprehensively. Considering the versatility and complexity, we finally chose this classical channel attention module as the feature enhancement structure in this work. Moreover, the quantitative results without the SE modules were better than those obtained without HFE. In other words, HFE plays a critical role in cloud removal performance, further demonstrating the necessity of enhancing high-level features for remote sensing images. More specifically, THFE enables the model to learn more global representations, facilitating the model to better predict the objects under the cloudy area. As shown in Table II, the results generated with THFE were better than those without THFE. Similar observations regarding CHFE can be made in Table II, which suggests models with CHFE can learn more detailed representations.

3) Effectiveness of Adding Perceptual Loss: To evaluate the effectiveness of the perceptual loss, we compared the proposed hybrid loss function with the loss function in the classical CycleGAN, as shown in Table III. The adjustable weights λ_{cyc} and λ_{id} of the loss function in the classical CycleGAN were set to 10 and 9, respectively. It is observed that there was a

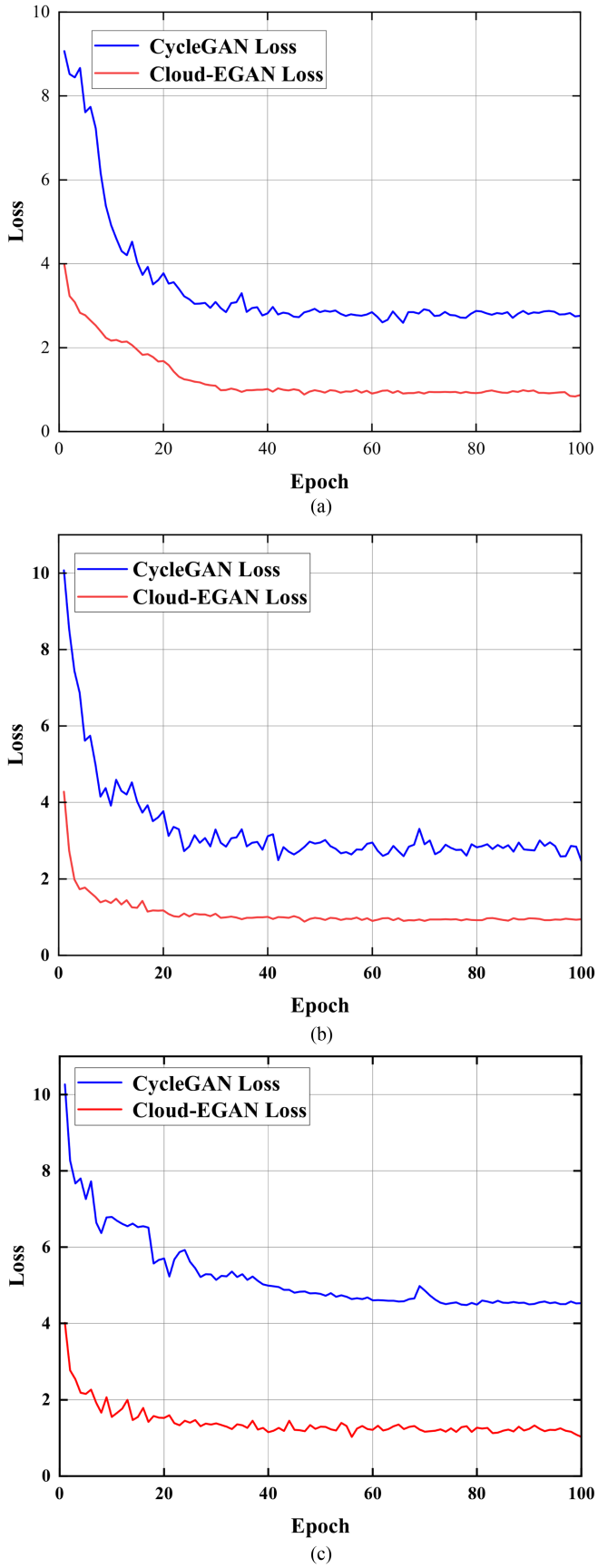

Fig. 9. Comparison of the training convergence of the classical CycleGAN and Cloud-EGAN on the (a) RICE1 dataset, (b) RICE2 dataset, and (c) WHUS2-CR dataset.

non-negligible improvement in terms of PSNR and SSIM after incorporating perceptual loss.

F. Model Complexity Analysis

Table IV shows the complexity evaluation results of all methods conducted in our work. SPAGAN achieved the best performance on these metrics since it only used simple CNN-based modules while the generation performance is poor. Compared to most other methods, the proposed Cloud-EGAN achieves significantly improved performance with low computational complexity by exploiting the convolution operations and the high-efficient WMSA module. Meanwhile, we added more modules including the CHFE module and a THFE module to enhance the high-level features. Therefore, the proposed Cloud-EGAN achieved a better cloud removal performance at the cost of a larger number of parameters and a lower inference speed.

G. Discussion

The experimental results have demonstrated that Cloud-EGAN performs better than existing DL-based models in the cloud removal task. This superior performance can be attributed to the cyclic structure and the integration of the SE and HFE modules. More specifically, Cloud-EGAN learns the mapping of feature representations between cloudy images and the corresponding cloud-free images in a cyclic-consistent way, which is conducive to strengthening the model capability of feature representation. Moreover, the combination of SE and HFE can effectively extract and aggregate contextual information, which is conducive to generating high-quality cloud-free images similar to the ground truth. The effectiveness of introducing SE and HFE can be validated from the feature maps shown in Fig. 8. Notably, the informative feature details are further enhanced through SE and HFE. As a result, cloud-removed scenes with enriched ground information can be preserved in Cloud-EGAN.

In addition, we compared the training loss convergence using Cloud-EGAN and the classical CycleGAN on the RICE1, RICE2, and WHUS2-CR datasets. It is observed in Fig. 9(a)–(c) that Cloud-EGAN obtained better convergence performance than CycleGAN due to the novel framework and the incorporation of the perceptual loss.

IV. Conclusion

In this work, a novel CycleGAN-based architecture, named Cloud-EGAN, has been proposed to perform supervised cloud removal tasks, which can effectively remove thin and thick clouds while preserving spectral and spatial consistency with the land surface. Compared with existing DL-based models developed for removing clouds, the proposed Cloud-EGAN utilizes a cyclic architecture while integrating the SE and HFE modules to enhance the ability to identify remote sensing images with complex ground objects. While the cyclic architecture is designed to recalibrate the weights of hierarchical channels, the integration of the SE and HFE modules is employed to further aggregate local and global high-level contextualized features. As a result, the proposed Cloud-EGAN can more effectively exploit multilevel enriched features with more saliency to highlight

ground information while suppressing cloud components and blurred edges through the integration of CNN and transformer. Extensive simulation results on the RICE and WHUS2-CR datasets have confirmed the superior cloud removal performance achieved by Cloud-EGAN as compared to existing DL-based methods for removing thin and thick clouds.

There are several extensions of this study that can be further explored. First, it is of great practical interest to further investigate how to construct a more computationally efficient model for various cloud-covered scenarios. Furthermore, it is interesting to consider applying the proposed Cloud-EGAN to large-scale remote sensing datasets, such as Sentinel-2 and Landsat-9 images in an unsupervised or semisupervised manner. Finally, end-to-end designs of cloud removal and other downstream tasks, such as semantic segmentation, will be explored in future research.

REFERENCES

[1] W. Ma et al., "Feature split–merge–enhancement network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616217.

[2] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5617712.

[3] G. Shi, J. Zhang, J. Liu, C. Zhang, C. Zhou, and S. Yang, "Global context-augmented objection detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10604–10617, Dec. 2021.

[4] W. Wang, Y. Chen, and P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5533918.

[5] Y. Hu, X. Huang, X. Luo, J. Han, X. Cao, and J. Zhang, "Variational self-distillation for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5627313.

[6] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, Jan. 2021.

[7] H. Li et al., "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5618014.

[8] L. Ding et al., "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 6512405.

[9] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7871–7886, Sep. 2020.

[10] Z. Sun, W. Zhou, C. Ding, and M. Xia, "Multi-resolution transformer network for building and road segmentation of remote sensing image," *ISPRS Int. J. Geo- Inf.*, vol. 11, no. 3, 2022, Art. no. 165.

[11] X. Wen, Z. Pan, Y. Hu, and J. Liu, "An effective network integrating residual learning and channel attention mechanism for thin cloud removal," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Mar. 2022, Art. no. 6507605.

[12] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *Proc. IEEE IGARSS Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1772–1775.

[13] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5222414.

[14] J. Zhou, X. Luo, W. Rong, and H. Xu, "Cloud removal for optical remote sensing imagery using distortion coding network combined with compound loss functions," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3452.

[15] F. Xu et al., "GLF-CR: SAR-enhanced cloud removal with global–local fusion," *ISPRS J. Photogrammetry Remote Sens.*, vol. 192, pp. 268–278, 2022.

[16] T.-Y. Ji, D. Chu, X.-L. Zhao, and D. Hong, "A unified framework of cloud detection and removal based on low-rank and group sparse regularizations for multitemporal multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5303015.

[17] J. Li et al., "Thin cloud removal fusing full spectral and spatial features for Sentinel-2 imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8759–8775, Oct. 2022.

[18] Q. Xiong, G. Li, X. Yao, and X. Zhang, "SAR-to-optical image translation and cloud removal based on conditional generative adversarial networks: Literature survey, taxonomy, evaluation indicators, limits and future directions," *Remote Sens.*, vol. 15, no. 4, 2023, Art. no. 1137.

[19] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2998–3006, May 2016.

[20] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan, "An effective thin cloud removal procedure for visible remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 96, pp. 224–235, 2014.

[21] M. Xu, M. Pickering, A. J. Plaza, and X. Jia, "Thin cloud removal based on signal transmission principles and spectral mixture analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1659–1669, Mar. 2016.

[22] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[23] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, Jun. 2019.

[24] H. Ding, Y. Zi, and F. Xie, "Uncertainty-based thin cloud removal network via conditional variational autoencoders," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 469–485.

[25] X. Wang, G. Xu, Y. Wang, D. Lin, P. Li, and X. Lin, "Thin and thick cloud removal on remote sensing image by conditional generative adversarial network," in *Proc. IEEE IGARSS Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1426–1429.

[26] J. Hwang, C. Yu, and Y. Shin, "SAR-to-optical image translation using SSIM and perceptual loss based cycle-consistent GAN," in *Proc. IEEE Int. Conf. Inf. Commun. Technol. Convergence*, 2020, pp. 191–194.

[27] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," 2020, *arXiv:2009.13015*.

[28] Y. Zhao, S. Shen, J. Hu, Y. Li, and J. Pan, "Cloud removal using multimodal GAN with adversarial consistency loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jul. 2021, Art. no. 8015605.

[29] W. Yu, X. Zhang, and M.-O. Pun, "Cloud removal in optical remote sensing imagery using multiscale distortion-aware networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 5512605.

[30] M. Xu, F. Deng, S. Jia, X. Jia, and A. J. Plaza, "Attention mechanism-based generative adversarial networks for cloud removal in Landsat images," *Remote Sens. Environ.*, vol. 271, 2022, Art. no. 112902.

[31] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[32] 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.

[33] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[34] 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.

[35] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.

[36] X. Zhang, W. Yu, M.-O. Pun, and W. Shi, "Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 1–17, 2023.

[37] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400515.

[38] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, Sep. 2022.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[40] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[41] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[42] D. Christopoulos, V. Ntouskos, and K. Karantzalos, "Cloudtran: Cloud removal from multitemporal satellite images using axial transformer networks," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 1125–1132, 2022.
[43] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv:1912.12180*.
[44] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3065.
[45] L. Gao et al., "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, Oct. 2021.
[46] C. Wang et al., "Translution-snet: A semisupervised hyperspectral image stripe noise removal based on transformer and CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5533114.
[47] K. Jiang, Z. Wang, C. Chen, Z. Wang, L. Cui, and C.-W. Lin, "Magic ELF: Image deraining meets association learning and transformer," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1–10.
[48] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2441–2449.
[49] X. Ma, X. Zhang, and M.-O. Pun, "A crossmodal multiscale fusion network for semantic segmentation of remote sensing data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3463–3474, Apr. 2022.
[50] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of CNN and transformer for lightweight image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 1103–1112.
[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.
[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[53] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.
[54] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.
[55] C. Zhou, J. Zhang, J. Liu, C. Zhang, R. Fei, and S. Xu, "Perceppan: Towards unsupervised pan-sharpening based on perceptual loss," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2318.
[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[57] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," 2019, *arXiv:1901.00600*.
[58] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier, "Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2A imagery," *Remote Sens.*, vol. 13, no. 1, pp. 157–188, 2021.
[59] L. Sun, Y. Zhang, X. Chang, Y. Wang, and J. Xu, "Cloud-aware generative network: Removing cloud from optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 691–695, Apr. 2020.
[60] A. Singh and L. Bruzzone, "Sigan: Spectral index generative adversarial network for data augmentation in multispectral remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2021, Art. no. 6003305.
[61] S. Bock and M. Weiß, "A proof of local convergence for the Adam optimizer," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.
[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[63] 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.

Xianping Ma (Student Member, IEEE) received the bachelor's degree in geographical information science from Wuhan University, Wuhan, China, in 2019. He is currently working toward the Ph.D. degree in computer and information engineering with The Chinese University of Hong Kong, Shenzhen, China.

His research interests include remote sensing image processing, deep learning, and multimodal learning.

Yiming Huang received the bachelor's degree in information engineering from the Guangdong University of Technology, Guangzhou, China, in 2022. He is currently working toward the master's degree in communication engineering with The Chinese University of Hong Kong, Shenzhen, China.

His research interests include remote sensing and deep learning, and cloud removal.

Xiaokang Zhang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from The School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2018.

From 2019 to 2022, he was a Postdoctoral Research Associate with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong, Shenzhen, China. He is currently a specially appointed Professor with the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China. He has authored or coauthored more than 20 scientific publications in international journals and conferences. His research interests include remote sensing image analysis, computer vision, and deep learning.

Dr. Zhang is currently a Reviewer for more than ten renowned international journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Information Fusion*, and the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Man-On Pun (Senior Member, IEEE) received the B.Eng. degree in electronic engineering from The Chinese University of Hong Kong, Shenzhen (CUHKSZ), Shenzhen, China, in 1996, the M.Eng. degree in computer science from the University of Tsukuba, Tsukuba, Japan, in 1999, and the Ph.D. degree in electrical engineering from the University of Southern California (USC) at Los Angeles, Los Angeles, CA, USA, in 2006.

He was a Postdoctoral Research Associate with Princeton University, Princeton, NJ, USA, from 2006 to 2008. He held research positions with Huawei, NJ, USA, the Mitsubishi Electric Research Labs (MERL), Boston, MA, USA, and Sony, Tokyo, Japan. He is currently an Associate Professor with the School of Science and Engineering, CUHKSZ. His research interests include artificial intelligence (AI) Internet of Things (AIoT) and applications of machine learning in communications and satellite remote sensing.

Prof. Pun was the recipient of best paper awards from the IEEE Vehicular Technology Conference 2006 Fall, the IEEE International Conference on Communication 2008, and the IEEE Infocom'09. He is the Founding Chair of the IEEE Joint Signal Processing Society-Communications Society Chapter, Shenzhen. He was an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2010 to 2014.

Bo Huang received the Ph.D. degree in remote sensing and mapping from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 1997.

He is currently a Chair Professor with the Department of Geography, The University of Hong Kong, Hong Kong. His research interests include most aspects of GIScience, specifically the design and development of models and algorithms for unified satellite image fusion, spatiotemporal statistics, and multiobjective spatial optimization, and their applications in environmental monitoring and sustainable land use and transportation planning.

Dr. Huang is currently an Associate Editor for the *International Journal of Geographical Information Science* (Taylor & Francis) and the Editor-in-Chief of *Comprehensive GIS* (Elsevier), a three-volume GIS sourcebook.