Diffusion Enhancement for Cloud Removal in Ultra-Resolution Remote Sensing Imagery

Jialu Sui, Yiyang Ma, Wenhan Yang, *Member, IEEE*, Xiaokang Zhang, *Member, IEEE*, Man-On Pun, *Senior Member, IEEE*, and Jiaying Liu, *Senior Member, IEEE*

***Abstract*— The presence of cloud layers severely compromises the quality and effectiveness of optical remote sensing (RS) images. However, existing deep-learning (DL)-based cloud removal (CR) techniques, which usually take the fidelity-driven losses as constraints, e.g., L_1 or L_2 losses, tend to generate smooth results, often failing to reconstruct visually pleasing results and cause semantic loss. To tackle this challenge, this work proposes to encompass enhancements at the data and methodology fronts. On the data side, an ultra-resolution benchmark named CUHK cloud removal (CUHK-CR) of 0.5 m spatial resolution is established. This benchmark incorporates rich detailed textures and diverse cloud coverage, serving as a robust foundation for designing and assessing CR models. From the methodology perspective, a novel diffusion-based framework for CR named diffusion enhancement (DE) is introduced. This framework aims to gradually recover texture details, leveraging a reference visual prior providing foundational structure of the images to enhance inference accuracy. Additionally, a weight allocation (WA) network is developed to dynamically adjust the weights for feature fusion, thereby further improving performance, particularly in the context of ultra-resolution image generation. Furthermore, a coarse-to-fine training strategy is applied to effectively expedite training convergence while reducing the computational complexity required to handle ultra-resolution images. Extensive experiments on the newly established CUHK-CR and existing datasets such as RICE confirm that the proposed DE framework outperforms existing DL-based methods in terms of both perceptual quality and signal fidelity.**

***Index Terms*— Cloud removal (CR), denoising diffusion probabilistic model, remote sensing (RS) images.**

Manuscript received 10 March 2024; revised 27 April 2024; accepted 4 June 2024. Date of publication 10 June 2024; date of current version 21 June 2024. This work was supported in part by Guangdong Provincial Key Laboratory of Future Networks of Intelligence under Grant 2022B1212010001, in part by the National Natural Science Foundation of China under Grant 42371374 and Grant 41801323, in part by DeRUCCI Company Ltd., and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010454. *(Corresponding authors: Man-On Pun; Xiaokang Zhang.)*

Jialu Sui is with Shenzhen Future Network of Intelligence Institute (FNii-Shenzhen), the School of Science and Engineering (SSE), and Guangdong Provincial Key Laboratory of Future Networks of Intelligence, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China (e-mail: jialusui@link.cuhk.edu.cn).

Yiyang Ma and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China (e-mail: myy12769@pku.edu.cn; liujiaying@pku.edu.cn).

Wenhan Yang is with the PengCheng Laboratory, Shenzhen 518066, China (e-mail: yangwh@pcl.ac.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).

Man-On Pun is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China (e-mail: SimonPun@cuhk.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3411671

I. Introduction

REMOTE sensing (RS) images play a crucial role in a variety of applications, including change detection [1], semantic segmentation [2], and object detection [3]. However, the imaging capabilities of satellite sensors, characterized by their ultralong-range nature, make them susceptible to degradation, resulting in quality distortions in the captured images. One significant factor contributing to such degradation is the presence of cloud cover. Clouds significantly reduce visibility and saturation in the images, undermining the effectiveness of RS images, especially in the optical domain. This cloud-induced degradation hampers the clarity and detail of the images, impacting their practical utility. Consequently, there is a pressing need for the development of restoration methods aimed at enhancing land surface information obscured by cloud layers, thereby improving the effectiveness of RS images.

Traditional methods for cloud removal (CR) can be broadly categorized into two main groups, namely multispectral and multitemporal techniques. More specifically, multispectral methods [4], [5], [6], [7] primarily rely on variations in wavelength-dependent absorption and reflection to recover obscured landscapes caused by haze and thin cirrus clouds. However, in scenarios involving thick and filmy clouds that entirely obstruct optical signals, the efficacy of multispectral methods may be compromised due to the absence of supplementary information. In contrast, multitemporal methods [8], [9] integrate clear sky conditions from reference images captured at different time instances. While the results derived from the multitemporal methods are more reliable in general as they stem from actual cloud-free observations, the rapid changes in the landscape significantly impact the accuracy of the reconstructed images.

In recent years, deep-learning (DL)-based methods have gained significant popularity for their extraordinary ability to generate high-quality, cloud-removed results. These approaches within the realm of DL can be further categorized into CNN-based models [10], generative adversarial network (GAN)-based models [11], [12], and diffusion-based models [13]. More specifically, CNN-based models operate by inputting cloudy images into a network and updating parameters based on loss functions calculated from the output

and the corresponding cloud-free image. Along the same direction, Meraner et al. [14] introduced a deep residual neural network designed to reconstruct an optical representation of the underlying land surface structure. Notably, SAR imagery was incorporated into the CR process to offer additional information on surface characteristics beneath clouds. Additionally, Ma et al. [15] utilized a two-step convolution network to extract transparency information from clouds and determine their positions. However, the feature representation capability of CNN-based models is limited, resulting in less precise prediction of detailed texture information. Since clouds often obscure a substantial portion of an image, the inferior prediction of detailed texture within cloudy regions directly contributes to the diminished perceptual quality of the generated cloud-removed images.

To address this limitation, GAN-based models employ unique training strategies to enhance the model's capability for detailed prediction, which incorporate two key components, namely the generator and the discriminator. The generator creates cloud-removed images, while the discriminator evaluates whether the generated images meet desired quality standards, providing gradients for updating the generator's parameters through an additional GAN loss function. For instance, CloudGAN [12] preserved color composition and texture by learning a bidirectional mapping of feature representation between cloudy images and their corresponding cloud-free counterparts in a cyclic structure. Nevertheless, GAN-based models face persistent challenges such as model collapse, unstable training dynamics, and vanishing gradients, all of which detrimentally impact their performance across various applications. Moreover, GAN-based models for CR continue to depend on pixel-level loss functions to some extent, limiting their ability to accurately predict intricate textures.

Recently, a novel branch of generative models, known as diffusion models [16], has been introduced to computer vision tasks. These models have demonstrated remarkable performance in generating detailed textures across various low-level tasks compared with the GAN-based models, including super-resolution [17], [18], [19], deblurring [20], [21], and inpainting [22]. Optimal integration of the gradual learning and refinement features of diffusion models into the generation process is expected to pave the way for more advanced and effective approaches in CR. The diffusion model aims to learn the data distribution of the cloud-free image under the condition of the cloudy image instead of learning the change from cloudy image to cloud-free image, which improves its flexibility in detailed texture generation. In comparison with GAN-based models, the diffusion model demonstrates a more remarkable capability in predicting detailed information due to its specific training strategy. However, it is noteworthy that the outcomes obtained from pure diffusion models for CR are often inaccurate with undesirable fake textures and misalignment. Consequently, the current applications of diffusion models in CR primarily focus on feature extraction [13], limiting their inherent capabilities for gradual learning and refinement in this context.

In this study, based on the diffusion architecture, we propose a novel network named diffusion enhancement (DE) for CR, aiming to leverage the inherent strengths of the diffusion model to improve the quality of images. In sharp contrast to existing diffusion-based methods that only rely on progressive refinement for reconstructing fine-grained texture details, this work proposes to integrate a reference visual prior. In this way, the global visual information from reference visual prior can be effectively integrated into the progressive diffusion process to mitigate the training difficulty, which results in improved inference accuracy. Besides, a weight allocation (WA) network is introduced to optimize the dynamic fusion of the reference visual prior and intermediate denoising images derived from the diffusion models. To expedite the diffusion model convergence, we further propose a coarse-to-fine training strategy. More specifically, the network is first trained on smaller patches before being fine-tuned using larger patches. Finally, taking advantage of recent satellite observations of high quality and resolution [23], [24], [25], an ultra-resolution benchmark containing clear spatial texture information of the location and intrinsic features of the landscape is established for CR algorithm design and performance evaluation.

In summary, the main contributions of this work are summarized as follows.

1) A novel network called DE is proposed in this work to restore land surface under cloud cover. The proposed DE network, which merges global visual information with progressive diffusion recovery, offers an enhanced capability of capturing data distribution. As a result, it excels in predicting detailed information by utilizing reference visual prior during the inference process.
2) A WA network is devised to compute adaptive weighting coefficients for the fusion of the reference visual prior and intermediate denoising images derived from the diffusion models. As a result, the reference visual prior refinement predominantly contributes to coarse-grained content reconstruction in the initial steps, while the diffusion model focuses its efforts on incorporating rich details in the subsequent stages. In addition, a coarse-to-fine training strategy is applied to stabilize the training while accelerating the convergence speed of DE.
3) Finally, an ultra-resolution benchmark called CUHK-CR is established to evaluate the CR methods against different types of cloud coverage. Our benchmark consists of 668 images of thin clouds and 559 images of thick clouds with multispectral information. To the best of our knowledge, our benchmark stands for the CR dataset of the highest spatial resolution, i.e., 0.5 m, among all existing CR datasets. The data and code can be downloaded from GitHub.[1]

The remainder of this article is structured as follows: an overview of existing CR datasets and methods is first presented in Section II before Section III outlines in detail our dataset CUHK-CR. After that, Section IV introduces the proposed DE network whereas experimental findings and insights are deliberated in Section V. Finally, concluding remarks are offered in Section VI.

[1]https://github.com/littlebeen/Diffusion-Enhancement-for-CR

TABLE I
COMPARISON BETWEEN EXISTING CR DATASETS AND CUHK-CR

Dataset	Acquired Time gap	Image Size	Number	Resolution (m)	Spectrum	Source
T-Cloud [10]	16 days	256	2,939	30	3	Landsat 8
RICE1 [26]	15 days	512	500	30	3	Google Earth
RICE2 [26]	15 days	512	736	30	3	Landsat 8
WHUS2-CR [27]	12 days	large	36	10	10	Sentinel-2A
SEN12MS-CR [28]	14 days	256	122,218	10	13	Sentinel-2
WHU Cloud Dataset [29]	6 months	512	859	30	3	Landsat 8
CUHK-CR1	17 days	512	668	0.5	4	Jilin-1
CUHK-CR2	17 days	512	559	0.5	4	Jilin-1

II. RELATED WORK

A. Conventional End-to-End Method for CR

End-to-end CR models including CNN-based models and GAN-based models are specifically designed to take a cloudy image as input and directly generate a cloud-removed image during the inference process. These models excel in swiftly producing inference results, primarily focusing on discerning the differences between the cloudy image and its corresponding cloud-free counterpart. CVAE [10] delved into the image degradation process using a probabilistic graphical model, whereas SpAGAN [30] emulated the human visual mechanism by employing a local-to-global spatial attention approach to detect and highlight cloud regions. Furthermore, AMGAN-CR [31] removed clouds using an attentive residual network guided by an attention map. Despite their merits, the visual outcomes of these end-to-end models consistently replace clouds with neighboring colors, lacking the capability to predict the underlying texture obscured by clouds. This limitation adversely impacts the effectiveness of these CR methods, particularly in cases of dense cloud coverage.

B. Diffusion Architecture and Prior Guidance

Recently, the diffusion model [16], [32], [33] has garnered significant attention with the improved capability on high-resolution image generation. This model gradually generates the ultimate result, denoted as $\mathbf{x}_0$, from a latent variable $\mathbf{x}_T$, where T represents the total number of diffusion steps in a parameterized Markov chain. The diffusion model comprises two key components, namely the forward process and the reverse process. More specifically, the forward process transforms the data distribution into a latent variable distribution through a step-by-step progression, leveraging the parameters of the Markov chain. Conversely, the reverse process aims to revert the latent variable distribution back to the original data distribution, recovering the initial data and providing a comprehensive understanding of the underlying data distribution.

In contrast to previously discussed end-to-end methods, the diffusion model [34], [35] offers a higher level of detailed information, beneficial for restoring the landscape under cloud coverage. However, the conventional diffusion model tends to generate unreliable fake textures and misalignment, since it endeavors to provide more detailed texture information using the restricted data available from cloudy images. In the absence of effective solutions to this issue, current diffusion model-based methods like DDPM-CR [13] primarily employ the diffusion model as a feature extractor, which overlooks the potential to leverage the diffusion model's inherent strengths in gradual learning and refinement. Alternatively, some pioneering attempts [36], [37] have been made to incorporate prior guidance to guide and regularize the generated results. Aiming to fully exploit the potential of the diffusion model for incremental learning and iterative refinement, while simultaneously minimizing the generation of spurious textures, our DE is crafted to leverage the diffusion process in conjunction with a reference visual prior.

TABLE II
JILIN-1KF01B SENSOR BANDS

Band Name	Spectral Coverage (nm)	Spatial Resolution (m)
Panchromatic color	450-800	0.5
Blue	450-510	2
Green	510-580	2
Red	630-690	2
Near-infrared	770-895	2

C. Datasets for CR

Table I lists several of the most representative existing image datasets for optical-based CR. As shown in Table I, all the datasets share a common drawback, i.e., their low spatial resolution of around 10–30 m. This limitation significantly compromises the level of spatial detail they can provide. Furthermore, despite the fact that multispectral information is necessary for satellite image analysis, datasets such as T-Cloud [10] and RICE [26] only contain RGB bands. In addition, it is advantageous to minimize the "acquired time gap" as significant landscape changes can occur between the time instances of taking the cloudy image and its corresponding clear image. However, popular datasets like WHU Cloud Dataset [29] possess a large "acquired time gap," which can be an issue of concern in practice. Finally, all datasets listed in Table I were generated with open-source satellites such as Landsat 8 and Sentinel-2. It is highly desirable to have datasets from more satellites with different sensor characteristics for CR algorithm design and performance assessment.

III. PROPOSED CUHK-CR DATASET

A. CUHK-CR

Driven by the ever-increasing resolution of RS imagery, we have established a new ultra-resolution benchmark named CUHK cloud removal (CUHK-CR). This benchmark is characterized by its ultra-high spatial resolution of 0.5 m and four multispectral bands with data acquisition confined to

TABLE III
SUMMARY OF THE STUDY SITES OF CUHK-CR

Data	Location	Size (km^2)	Covers	Cloudy Data Acquired Time	Cloud-free Data Acquired Time
I	Shenzhen Guangdong Province	35.72	City, Forest	2022/08/24	2022/09/03
II	Jinan Shandong Province	34.76	City, River	2022/08/24	2022/09/10
III	Hangzhou Zhejiang Province	25.62	City, Lack	2023/01/31	2023/01/21

a period of 17 days. Such an ultra-high spatial resolution benchmark can facilitate the training and evaluation of various CR methods specifically designed for ultra-resolution images. As a result, the benchmark can mitigate the gap between the low-resolution images during training and the high-resolution acquired in the real world, which is shown particularly critical for good CR performance in Section V. Furthermore, the benchmark comprises two subsets, a thin cloud subset, namely CUHK-CR1, and a thick cloud subset, namely CUHK-CR2, facilitating training and evaluation on varying cloud coverage. More specifically, the thin cloud subset includes 668 images, while the thick cloud subset includes 559 images. These images are cropped into smaller segments for convenience and directly compatible with DL models. Unless specified otherwise, a training-to-testing set ratio of 8 : 2 is employed in the sequel, resulting in 534 and 448 images for training, and 134 and 111 images for testing in the thin and thick subsets, respectively. Finally, it is worth pointing out that our dataset is based on a new commercial satellite, Jilin-1, instead of those frequently utilized satellites like Landsat 8 and Sentinel-2. The distinct image contexts provided by the Jilin-1 satellite sensors contribute to the uniqueness of our dataset.

B. Data Collection

Jilin-1 satellite constellation is the core project of Chang Guang Satellite Technology Company Ltd. (CGSTL). The constellation is composed of 138 high-performance optical RS satellites, covering high resolution, large width, video, and multispectrum information. Our dataset was collected by a satellite named Jilin-1KF01B equipped with a 0.5 m resolution push broom camera. Launched in 2021, Jilin-1KF01B incorporates advanced technology to acquire more than 2 million km^2 of high-definition images every day with a width greater than 150 km. As shown in Table II, the push broom camera covers four spectral bands, namely red (R), green (G), blue (B), and near-infrared, as well as a high-resolution panchromatic color band. Pen-sharpening using the complementary information from the multispectral and panchromatic images is applied to improve the spatial resolution of the spectral bands from 2 to 0.5 m. Following data processing, we obtain high-resolution satellite images with four bands: blue, green, red, and near-infrared. These images are fed into the model to generate cloud-removed images of the same size. The optical RGB bands can reflect the color characteristics of land surface in line with human perception. In addition to the optical bands, the near-infrared band encounters fewer disturbances from thin clouds, thereby enhancing the extraction of precise cloud distortion layers and enabling a more accurate reconstruction of the background signal for visible bands. As a result, the multispectral data [27] enhances CR by the supplementary assistance from the near-infrared band. Table III refers to the location, size, coverage, and acquired time of the cloudy images and their corresponding cloud-free images. The location of the satellite images is chosen from the north to the south of China while the gap in acquisition time is limited to 17 days.

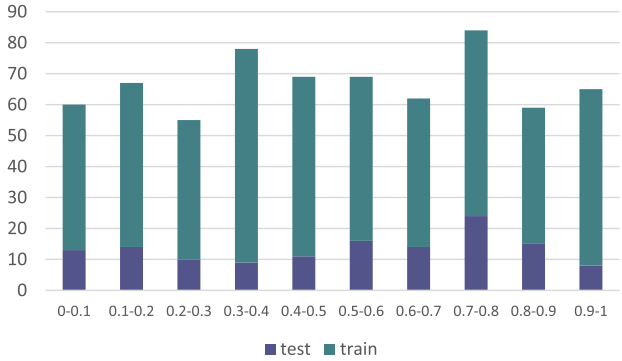

Fig. 1. Distribution of images on different CCPs of CUHK-CR1 training and test datasets computed via the detector of Cloud-Net [38]. The average probability of cloud coverage is 50.7%.

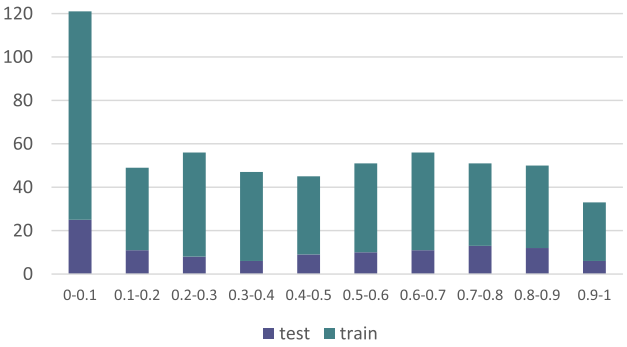

Fig. 2. Distribution of images on different CCPs of CUHK-CR2 training and test datasets computed via the detector of Cloud-Net [38]. The average probability of cloud coverage is 42.5%.

C. Data Analysis

To analyze the cloud coverage statistics in the CUHK-CR dataset, we calculate the widely used cloud coverage probability (CCP) [28] on two distinct sets. We visualize the distribution of image counts for different CCP values in Figs. 1 and 2.

For each optical image, the Cloud-Net detector [38], [39] is applied to produce binary masks with pixelwise values of either 0 or 1 with 0 and 1 indicating the cloudy and cloud-free places, respectively. It is important to note that the detector fails to differentiate between thin and thick cloud layers. It simply detects the presence of cloud cover at the pixel level.

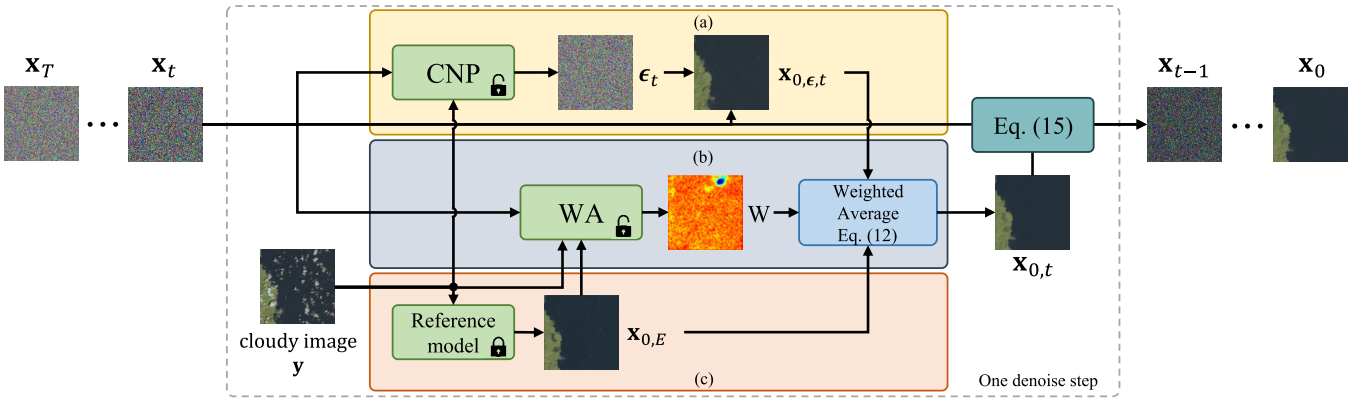

Fig. 3. Architecture of our DE for CR. (a) *Diffusion branch* performs the diffusion step that removes noise progressively, which is capable of restoring fine-grained textures. (b) *Weighting branch* performs the dynamic fusion of results from both the reference and diffusion branches with the result $\mathbf{x}_{0,t}$, capturing the merits of both excellent global estimations and fine details. (c) *Reference branch* generates a cloud-removed image based on the cloudy image $\mathbf{y}$, offering substantial global context. Ultimately, $\mathbf{x}_{0,t}$ and $\mathbf{x}_t$ are utilized in the generation of $\mathbf{x}_{t-1}$.

Thin clouds typically extend over a broader area, whereas thick clouds occupy a smaller portion of the image, including richer reference information used for predicting background ground. We remove those images whose landscapes are totally obscured by the dense clouds through visual observation. As a result, the average CCP for the set with thin clouds is higher than that for the set with thick clouds. Notably, the images with CCP between 0 and 0.1 account for the largest proportion in the CUHK-CR2.

IV. DE for CR

A. Architecture

Similar to the denoising diffusion probabilistic model [16], the proposed DE network proceeds in the following two processes.

1) Forward Process: It transforms the initial data distribution $q(\mathbf{x}_0)$ into a latent variable distribution $q(\mathbf{x}_T)$, where T represents the total number of the time steps. This transformation follows a fixed Markov chain that can be modeled as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{1}$$

where $\mathcal{N}$, $\{\beta_1, \ldots, \beta_T\} \in (1, 0)$, and $\boldsymbol{I}$ stands for the Gaussian distribution, a set of hyperparameters and the identity matrix, respectively.

By exploiting (1), we have

$$q(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{2}$$

As a result, the forward process can be represented as

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Subsequently, we can express $\mathbf{x}_t$ as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ is a standard Gaussian noise.

2) Reverse Process: It transforms the latent variable distribution $p_{\boldsymbol{\theta}}(\mathbf{x}_T)$ back to the data distribution $p_{\boldsymbol{\theta}}(\mathbf{x}_0)$ through a network parameterized by $\boldsymbol{\theta}$. The reverse process is defined as a Markov chain with learned Gaussian transitions starting with a Gaussian distribution

$$p_{\boldsymbol{\theta}}(\mathbf{x}_0, \ldots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{5}$$

where

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \sigma_{\boldsymbol{\theta}}(\mathbf{x}_t, t)^2 \boldsymbol{I}) \tag{6}$$

with $\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ and $\sigma_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ being the mean and variance of the Gaussian distribution at the t-th step.

During the training process, we propose to minimize the mean square error (MSE) loss between the random noise $\boldsymbol{\epsilon}$ added to the clean image and the predicted noise $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{y})$ derived from $\mathbf{x}_t$, t and cloudy image $\mathbf{y}$. Since the DE network predicts the noise information based on the cloudy images, it is named conditional noise predictor (CNP). In summary, the loss function employed takes the following form:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_t, \boldsymbol{\epsilon}, t, \mathbf{y}}\left[||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{y})||^2\right]. \tag{7}$$

B. Reference Visual Prior Integration

Inspired by Fei et al. [36] and Zhou et al. [37], the proposed DE network incorporates reference visual priors, comprising a weighting branch and a reference branch, to direct the inference process toward obtaining refined outcomes, as illustrated in Fig. 3. The reference visual prior is intended to offer the global image structure, thereby diminishing the production of unwanted fake textures generated by the pure diffusion model. We outline the implementation process and motivation behind our DE approach.

For the t-th step of the reverse process, $\mathbf{x}_{t-1}$ is calculated based on $\mathbf{x}_t$ and $\mathbf{x}_{0,t}$. $\mathbf{x}_t$ and $\mathbf{x}_{0,t}$ are noise image at time step t and a calculated clear image at the intermediary stage of the diffusion model, respectively. For the calculation of $\mathbf{x}_{0,t}$, we first predict the noise $\boldsymbol{\epsilon}_t$ based on the state $\mathbf{x}_t$, the time-step t, and the cloudy image $\mathbf{y}$

$$\boldsymbol{\epsilon}_t = \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t, \mathbf{y}). \tag{8}$$

After that, as a reverse process of (4), we obtain $\mathbf{x}_{0,\boldsymbol{\epsilon},t}$ in the current step t based on the predicted noise $\boldsymbol{\epsilon}_t$ and the noisy image $\mathbf{x}_t$

$$\mathbf{x}_{0,\boldsymbol{\epsilon},t} = (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t)/\sqrt{\bar{\alpha}_t}. \tag{9}$$

For the pure diffusion model, $\mathbf{x}_{0,t}$ is equal to $\mathbf{x}_{0,\boldsymbol{\epsilon},t}$. In our approach, we utilize a reference visual prior to refine $\mathbf{x}_{0,\boldsymbol{\epsilon},t}$, obtaining an improved $\mathbf{x}_{0,t}$. The improved $\mathbf{x}_{0,t}$ approaches the genuine cloud-free image more closely, yielding superior final outcomes. Specifically, we begin by utilizing the reference model denoted as $\mathbf{E}$ to produce a cloud-removed output denoted as $\mathbf{x}_{0,E}$

$$\mathbf{x}_{0,E} = \mathbf{E}(\mathbf{y}). \tag{10}$$

The output $\mathbf{x}_{0,E}$ generated by the reference model serves as the primary structural foundation of the image, while $\mathbf{x}_{0,\boldsymbol{\epsilon},t}$ predicted by the diffusion model introduces abundant details and textures. A comprehensive formula for this refinement process is presented as follows:

$$\mathbf{x}_{0,t} = \Gamma(\mathbf{x}_{0,E}, \mathbf{x}_{0,\boldsymbol{\epsilon},t}) \tag{11}$$

where Γ means the fusion function.

In practice, we utilize a pixelwise linear combination of the two predictions

$$\mathbf{x}_{0,t} = (\mathbf{1} - \mathbf{W}) \odot \mathbf{x}_{0,\boldsymbol{\epsilon},t} + \mathbf{W} \odot \mathbf{x}_{0,E} \tag{12}$$

where $\odot$ represents the element-wise multiplication, $\mathbf{1}$ means the all-one matrix, and $\mathbf{W} \in \mathbb{R}^{C\times H\times W}$ is a pixel-wise fusion ratio which will be further described in the next part.

Finally, according to the posterior distribution of diffusion model [16], we could sample $\mathbf{x}_{t-1} \sim p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ based on $\mathbf{x}_t$ and the refined $\mathbf{x}_{0,t}$ from the distribution shown in (6) with the mean value $\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ and variance $\sigma_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$

$$\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_{0,t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \tag{13}$$

$$\sigma_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \tilde{\beta}_t^{\frac{1}{2}} \tag{14}$$

where $\tilde{\beta}_t = (1-\bar{\alpha}_{t-1}/1-\bar{\alpha}_t)\beta_t$.

Combining (6), (13), and (14), the formula of $\mathbf{x}_{t-1}$ is as follows:

$$\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_{0,t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \tilde{\beta}_t z, z \sim \mathcal{N}(0, \boldsymbol{I}). \tag{15}$$

In Fig. 4, we present an example of $\mathbf{x}_{0,t}$ from time step T to 1 demonstrating the impact of integrating the reference visual prior. As depicted in the initial line of Fig. 4, the diffusion model, guided by the loss function shown in (7), primarily concentrates on learning the distribution of the entire image set rather than the fine pixel-level information. This approach inspires its ability to generate diverse texture information. However, the generated textures often lack authenticity failing to accurately align with the actual scene due to the absence of direct structure constraints at the pixel level. Notably, discrepancies are evident in features such as the lake outline and background texture, which do not align closely with the ground truth.

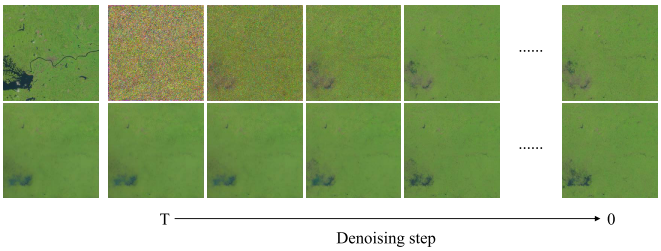

Fig. 4. Style of $\mathbf{x}_{0,t}$ from denoising time step T to 0. The first line and second line represent the result of the vanilla diffusion model and our DE, respectively. The ground truth and cloud-removed image generated by reference model are presented on the left side.

On the other hand, based on the result of the reference model illustrated in Fig. 4, the reference model implemented in an end-to-end manner primarily relies on fidelity-driven loss functions during training to minimize pixel disparities between cloud-removed and cloud-free images. Consequently, they can swiftly reconstruct the accurate underlying structure of cloud-removed images in a single step. This characteristic renders them effective for low-resolution datasets with limited texture information. However, when dealing with higher-resolution scenes boasting richer textures, the reference model struggles to capture and replicate those fine-grained details. As a result, it is challenging to faithfully restore complicated landscapes beneath the cloud cover.

Considering these pros and cons, our DE introduces the reference visual prior to the diffusion model. We utilize guidance from an approximately cloud-removed image generated by a reference model, denoted as $\mathbf{x}_{0,E}$, to steer the denoising process. $\mathbf{x}_{0,E}$, predicted by the reference model, establishes the fundamental image structure, while $\mathbf{x}_{0,\boldsymbol{\epsilon},t}$ generated by the diffusion model introduces details and textures. As a result, the accurate structure produced by the reference model helps mitigate the generation of fake details by the diffusion model, while the diffusion model contributes additional texture information to enhance cloud-removed image reconstruction, particularly for high-resolution scenes. As depicted in the second line of Fig. 4, our DE effectively addresses the limitations of both the pure diffusion model and the reference model through the reference visual prior.

C. Dynamic Fusion Among Diffusion Steps

We employ a WA network, which is trained to dynamically balance the fusion of results from the diffusion model and reference model throughout the progressive diffusion process steps. As shown in Fig. 5, the WA takes inputs consisting of the concatenation of $\mathbf{x}_t$, $\mathbf{y}$, and $\mathbf{x}_{0,E}$, with the time step t guiding the network across all layers. The UNet architecture of WA is inspired by CNP [40]. Consequently, the training objective enables the WA dynamically to determine the fusion ratio $\mathbf{W}$ based on the noise strength from time step t and the image features from $\mathbf{x}_t$, $\mathbf{y}$, and $\mathbf{x}_{0,E}$, allowing for temporal and spatial adaptation.

As shown in the first line of Fig. 4, the images $\mathbf{x}_{0,\boldsymbol{\epsilon},t}$ produced by the diffusion model initially contain a significant amount of noise. The noise in the image gradually decreases

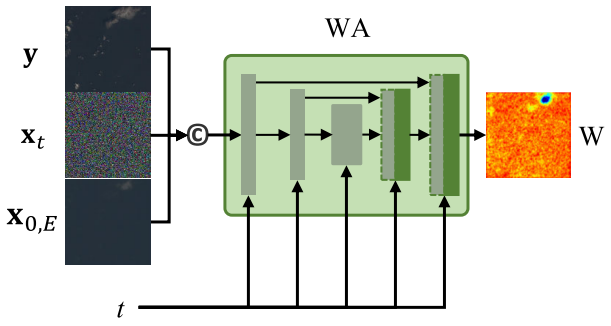

Fig. 5. Architecture of WA. WA learns to dynamically determine the weighting matrix based on the image features and the noise strength.

as the time step approaches one. $\mathbf{x}_{0,E}$ closely approximates the ground truth compared with $\mathbf{x}_{0,\epsilon,t}$, especially at the first few time steps. In order to gain the high-quality $\mathbf{x}_{0,t}$, according to (12), the fusion ratio should make temporal adaptation by achieving a high value at first to facilitate the establishment of precise image structure based on the reference model and gradually diminish as t decreases to enhance various texture generation based on the diffusion model. Furthermore, despite that the image structure $\mathbf{x}_{0,E}$ from the reference model is roughly accurate, it still may contain minor errors. And the image noise from $\mathbf{x}_{0,\epsilon,t}$ which has not been totally removed is randomly distributed across the entire image. Hence, the fusion ratio is also critical for adaptation in the spatial domain to detect the noise and errors from the results of both the diffusion model and reference model.

To tackle these challenges, we obtain the fusion ratio $\mathbf{W}$ from the WA network based on the time step t and the image restoration result. This enables WA to generate a specific fusion ratio for each time step t and each pixel, thereby providing detailed pixel-level weight information for the refinement process. Moreover, to prevent $\mathbf{x}_{0,t}$ from becoming overly reliant on $\mathbf{x}_{0,\epsilon,t}$ with a low value of $\mathbf{W}$, which could lead to unguided generation results, we introduce a limiting factor η to confine the range of $\mathbf{W}$ from η to 1 in the inference process to ensure the constraint from the reference model to the final result. Further details regarding the hyperparameter η are provided in Section V-C1. In summary, the WA network encourages the diffusion model to focus on generating more detailed texture information based on the image structure provided by the reference model. Additionally, it still aids in identifying and rectifying errors originating from the reference model.

D. Coarse-to-Fine Training and Inference

To accelerate the convergence speed of our DE during the training phase, we implement a coarse-to-fine training strategy. Initially, the image is resized to 1/4 of its original dimensions and processed by a sole diffusion model. Throughout this process, the employed loss function is given in (7). The fine-tuning process takes place after the diffusion model reaches near convergence at this smaller scale.

Once the network convergences on smaller images, we introduce and train the WA network with the full-size images, leveraging the knowledge from the well-converged diffusion network. The WA achieves initial convergence based on the locked diffusion model trained on the downscaled images, laying a foundation for the subsequent joint training of the diffusion model and the WA. In this context, the corresponding loss function of the DE is defined as

$$\begin{aligned}\mathcal{L}_{\text{WA}} &= |\tilde{\mathbf{x}}_0 - \mathbf{x}_{0,t}| \\ &= |\tilde{\mathbf{x}}_0 - (\mathbf{1} - \mathbf{W}) \odot (\mathbf{x}_{0,\epsilon,t})_{sg} + \mathbf{W} \odot \mathbf{x}_{0,E}| \end{aligned} \tag{16}$$

where $\tilde{\mathbf{x}}_0$ means the real cloud-free image and $(\cdot)_{sg}$ means stop gradient. Only the gradient of $\mathbf{W}$ is calculated while $\mathbf{x}_{0,\epsilon,t}$'s is disabled.

Ultimately, the CNP and WA are jointly trained using the full-size images. The loss function for this joint training is defined as

$$\mathcal{L}_{\text{joint}} = \lambda \cdot \mathcal{L}_{\text{DDPM}} + \mathcal{L}_{\text{WA}} \tag{17}$$

where λ is the weight proportion coefficient to balance the value gap between the two parts of the loss function. The detailed setting of λ is provided in Section V-B.

Notably, since the training process of the diffusion model is known to be quite unstable, in (16) and the second segment of (17), the gradient of $\mathbf{x}_{0,\epsilon,t}$ remains deactivated to prevent any adverse effects on the CNP. The CNP consistently maintains its original training strategy for larger images, while the WA adapts its approach based on the training outcomes of the CNP.

Throughout the inference process, at each step, the diffusion model predicts the noise $\boldsymbol{\epsilon}_t$ and computes $\mathbf{x}_{0,\epsilon,t}$ using (9). Subsequently, the reference model generates its cloud-removed output, $\mathbf{x}_{0,E}$, which is then utilized by the WA to determine the fusion ratio, $\mathbf{W}$. $\mathbf{x}_{0,t}$ is calculated through a pixelwise linear combination of the predictions of $\mathbf{x}_{0,E}$ and $\mathbf{x}_{0,\epsilon,t}$ based on $\mathbf{W}$ produced by the WA. Ultimately, $\mathbf{x}_{t-1}$ is generated and the denoising cycle concludes when $t = 1$.

TABLE IV
DETAILED MODEL SETTING OF THE CNP AND WA

Model setting	CNP	WA
Channels	96	64
Depth	2	2
Channels multiple	1, 1, 2, 2, 3	1, 1, 2
Attention resolution	4,8	4,8
Heads	4	1
Dropout	0.0	0.0

V. EXPERIMENTS

A. Datasets and Metrics

To evaluate the efficiency of our proposed method, we utilize two datasets: RICE [26] and the newly introduced CUHK-CR, for validation. The RICE dataset comprises 500 images with thin cloud covers and 736 images with thick cloud covers in RGB channel and sized at 512×512 pixels. The training and test sets are randomly partitioned in an 8:2 ratio. Further details about our CUHK-CR dataset are provided in Section IV.

We employ three widely recognized metrics for quantitative evaluation of CR performance: peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual

TABLE V
QUANTITATIVE EXPERIMENTAL RESULTS ON THE RICE1 AND RICE2 DATASETS. ↑ AND ↓ REPRESENT HIGHER BETTER AND LOWER BETTER, RESPECTIVELY

Method	RICE1			RICE2		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SpAGAN	28.509	0.9122	0.0503	28.783	0.7884	0.0963
AMGAN-CR	26.497	0.9120	0.0447	28.336	0.7819	0.1105
CVAE	31.112	0.9733	0.0149	30.063	0.8474	0.0743
MemoryNet	34.427	0.9865	0.0067	32.889	0.8801	0.0557
MSDA-CR	34.569	0.9871	0.0073	33.166	0.8793	0.0463
DE-MemoryNet	**35.525**	**0.9882**	**0.0055**	**33.637**	0.8825	0.0476
DE-MSDA	35.357	0.9878	0.0061	33.598	**0.8842**	**0.0452**

TABLE VI
QUANTITATIVE EXPERIMENTAL RESULTS ON THE CUHK-CR1 AND CUHK-CR2 DATASETS. ↑ AND ↓ REPRESENT HIGHER BETTER AND LOWER BETTER, RESPECTIVELY

Method	CUHK-CR1			CUHK-CR2		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SpAGAN	20.999	0.5162	0.0830	19.680	0.3952	0.1201
AMGAN-CR	20.867	0.4986	0.1075	20.172	0.4900	0.0932
CVAE	24.252	0.7252	0.1075	22.631	0.6302	0.0489
MemoryNet	26.073	0.7741	0.0315	24.224	0.6838	0.0403
MSDA-CR	25.435	0.7483	0.0374	23.755	0.6661	0.0433
DE-MemoryNet	**26.183**	**0.7746**	**0.0290**	**24.348**	**0.6843**	**0.0369**
DE-MSDA	25.739	0.7592	0.0321	23.968	0.6737	0.0372

image patch similarity (LPIPS) [41]. PSNR evaluates the generated image by comparing it with the ground truth at the pixel level. SSIM primarily assesses structural differences, while LPIPS aligns more closely with human perception.

B. Implementation Details

Our DE is based on the guided diffusion [40]. The hyperparameters of the UNet for CNP and the WA are listed in Table IV.

In DE, the CNP and WA undergo training employing the L_2 and L_1 loss, respectively, with a consistent learning rate of 10^{-5}. We maintain a weight proportion coefficient, λ, set at 1. To enhance efficiency in inference, we implement DDIM [42] with 50 steps, and the limiting factor η is set to 0.3 which means that the values of $\mathbf{W}$ are confined within the range of 0.3–1. All images, for both training and testing, are standardized to dimensions of 256 × 256 pixels. Initially, CNP is trained by smaller images measuring 64 × 64 pixels, utilizing a batch size of 64. As the training dataset shifts to standard-sized 256 × 256 pixel images, the batch size is adjusted to 16. For our CUHK-CR dataset, we perform model training and testing using four-band multispectral images. All experiments are executed using PyTorch on a single NVIDIA GeForce RTX 4090 GPU equipped with 24 GB of RAM.

C. Performance Comparison

We conduct a comprehensive comparison between our DE and several state-of-the-art CR networks, including two CNN-based models, namely MemoryNet [43], CVAE [10], and three GAN-based models, namely SpAGAN [30], AMGAN-CR [31], and MSDA-CR [44]. We choose two types of reference models, MSDA-CR and MemoryNet, to train and evaluate our DE. To differentiate between the DE variants trained on these models, we label them as DE-MSDA and DE-MemoryNet, respectively. To ensure a fair evaluation, all of these methods are thoroughly optimized using our training and test datasets to achieve their peak performance.

The quantitative results of these experiments on the RICE and CUHK-CR datasets are presented in Tables V and VI, respectively. Since the visual differences of the thin clouds are not readily discernible, we have chosen to display visual comparisons solely for thick cloud datasets at Figs. 6 and 7.

1) RICE: As indicated in Table V, our method demonstrates a substantial improvement compared to its corresponding reference model. Notably, our DE-MSDA and DE-MemoryNet achieve superior performance among these end-to-end models. For MSDA-CR, which achieves the best results on both RICE datasets, our DE-MSDA exhibits an improvement of 0.8 dB and 0.001, 0.4 dB, and 0.01 in PSNR and LPIPS for RICE1 and RICE2, respectively. These gains in PSNR and LPIPS indicate that our results not only achieve accurate landscape predictions but also align with human perception. Our diffusion-based approach significantly enhances the generation of fine textures, closely matching the ground truth, within the framework provided by the corresponding reference visual prior. The enhancements on LPIPS are especially obvious in the context of RICE2, where the dense cloud cover raises a hard challenge for cloud-removed image reconstruction. This scenario demands a heightened capacity for generating complicated and visually authentic textural details, given the considerable amount of obscured texture concealed by the clouds. Consequently, the model's capability to predict and generate textures is highlighted in such conditions. Though the end-to-end models such as MemoryNet and MSDA-CR also gain promising results, our DE could make additional improvements based on them.

Visual results are presented in Fig. 6. SpAGAN and AMGAN-CR exhibit obvious shortcomings in image style and color. Despite the superior results achieved by MSDA-CR

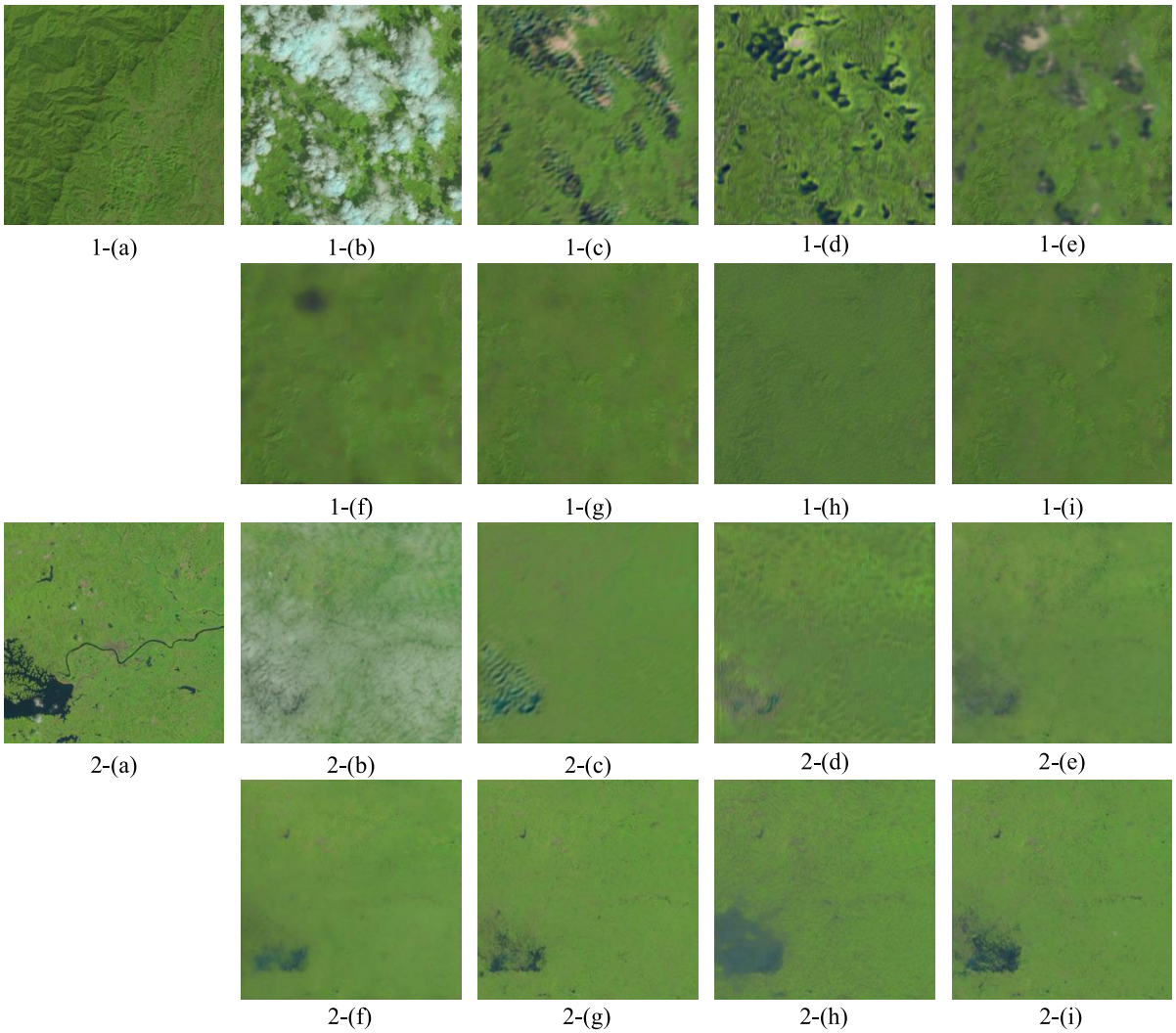

Fig. 6. Visual comparisons on RICE. (a) Label. (b) Cloudy image. (c) SpAGAN. (d) AMGAN-CR. (e) CVAE. (f) MemoryNet. (g) DE-MemoryNet. (h) MSDA-CR. (i) DE-MSDA.

and MemoryNet, there are still some errors present, including residual noise and cloud cover. Additionally, these models make relatively few predictions regarding texture information. In contrast, our DE is capable of error correction and accurate detailed predictions. For instance, our DE-MSDA and DE-MemoryNet exhibit enhanced reconstruction of the lake outline in the second image compared to MemoryNet and MSDA-CR.

2) CUHK-CR: The restored results of our CUHK-CR dataset are generally less satisfactory compared to RICE. The highest PSNR achieved by the end-to-end model in the RICE dataset surpasses 30 dB, but it decreases to 26 and 24 dB in the CUHK-CR1 and CUHK-CR2 datasets, respectively. The results signify that our ultra-resolution dataset presents greater challenges. Despite the increased difficulty, our DE-MSDA still yields superior results, achieving nearly a 0.3 dB PSNR improvement in both CUHK-CR1 and CUHK-CR2. On the CUHK-CR dataset, the limitations of certain models like SpAGAN and AMGAN-CR become more conspicuous when confronted with such ultra-resolution images, underscoring their unsuitability for high-resolution CR tasks in the realm of RS. They exhibit limited effectiveness in removing clouds, with an improvement of less than 1 dB improvement over the cloudy image.

Visual results for CUHK-CR are provided in Fig. 7. SpAGAN and AMGAN-CR struggle with such high-resolution CR tasks, particularly in the presence of thick clouds. In the case of CVAE and MemoryNet, despite a reasonable outline, it struggles with severe color deviations. Our DE primarily introduces subtle texture changes and correct color when compared to their corresponding reference models. For instance, the color of the roof in the output of DE-MSDA more closely resembles the ground truth than that of MSDA-CR. Furthermore, the results after our enhancement DE-MemoryNet appear clearer and more accurate, in contrast to the rather blurry outputs of MemoryNet, especially in areas obscured by dense clouds.

D. WA Analysis

1) Spatial Adaptation: In Fig. 8, we present an example of the attention heat map that depicts the behavior of the WA. Notably, the reference model falls short of completely removing the cloud cover, as indicated by the highlighted

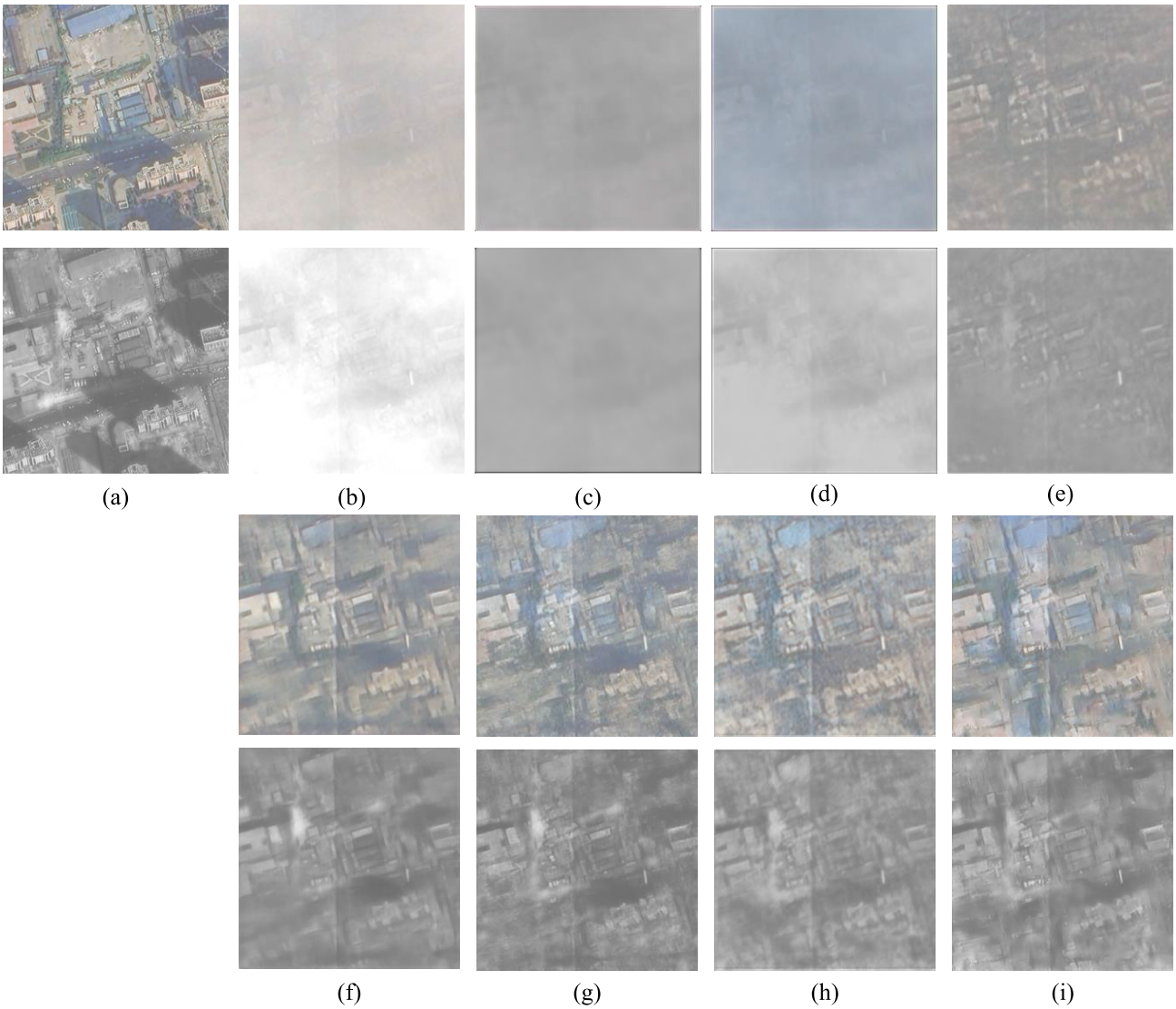

Fig. 7. Visual comparisons on CUHK-CR. The first line and second line present the RGB images and near-infrared images, respectively. (a) Label. (b) Cloudy image. (c) SpAGAN. (d) AMGAN-CR. (e) CVAE. (f) MemoryNet. (g) DE-MemoryNet. (h) MSDA-CR. (i) DE-MSDA.

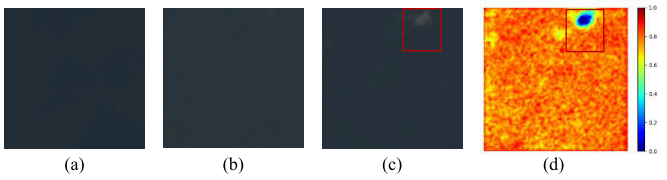

Fig. 8. Example presents the attention heatmap of $\mathbf{W}$. As the value approaches 1, its reliance on $\mathbf{x}_{0,E}$ becomes more pronounced. Conversely, as it nears 0, it exhibits a stronger dependence on $\mathbf{x}_{0,\epsilon,t}$. (a) Label. (b) $\mathbf{x}_{0,\epsilon,t}$. (c) $\mathbf{x}_{0,E}$. (d) Heatmap of $\mathbf{W}$.

area within the red box. As depicted in Fig. 8(d), our WA diligently addresses this discrepancy by reducing the weight allocated to this specific area. Moving to the domain of $\mathbf{x}_{0,\epsilon,t}$, we observe that some regions still retain residual noise that has not been eliminated. In response, the value of $\mathbf{W}$ is notably higher in these challenging areas, denoting slight adjustments that correspond to the noise distribution. This attention heat map serves as a compelling visual representation of the WA's capacity to dynamically fine-tune the strength of the reference visual prior in the spatial domain. The results demonstrate that this fine-tuning process could generate superior $\mathbf{x}_{0,t}$ based on the assessment of the quality of both $\mathbf{x}_{0,E}$ and $\mathbf{x}_{0,\epsilon,t}$. The refined $\mathbf{x}_{0,E}$ is closer to the ground truth, consequently leading to better final cloud-removed results.

2) Temporal Adaptation: The change of the mean value of $\mathbf{W}$ over each time step is visually represented in Fig. 9. Initially, the mean value of $\mathbf{W}$ is relatively high and gradually decreases as the time step decreases, approaching nearly 0 in the later stages. This trend indicates that, at the outset, $\mathbf{x}_{0,t}$ primarily relies on the guidance provided by $\mathbf{x}_{0,E}$, while the influence of $\mathbf{x}_{0,\epsilon,t}$ becomes more prominent as the time step approaches 0.

The fluctuation in the mean value of $\mathbf{W}$ reveals the underlying assumption that reference visual prior establishes the groundwork for the overall structure of $\mathbf{x}_{0,t}$ in the first few denoising steps, outlining the likely shapes of the images. Subsequently, the diffusion architecture intervenes, making fine adjustments by introducing additional texture information and correcting errors based on the guidance. This dynamic

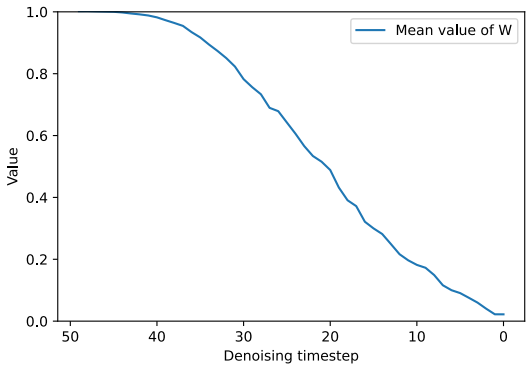

Fig. 9. Mean value generated by the WA in each time step on RICE2 with reference model MSDA-CR.

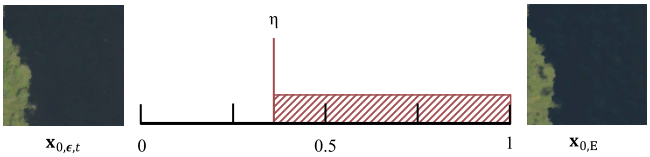

Fig. 10. Schematic representation of the adjustment of the limiting factor η. The red box means the limited value range of $\mathbf{W}$.

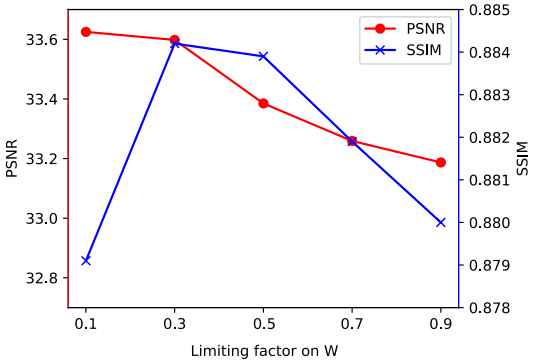

Fig. 11. Experimental comparisons on RICE with different limiting factors η on $\mathbf{W}$.

shift in the mean value of $\mathbf{W}$ underscores the collaborative relationship between the reference visual prior and the diffusion architecture, leading to a leap in reconstruction performance.

3) Parameter Analysis: Our investigation delves into the impact of the limiting factor η on $\mathbf{W}$. The schematic representation of η adjustment is illustrated in Fig. 10. It means that each value in $\mathbf{W}$ is limited at the range of η–1. In the training process, we set η to 0, thereby effectively allowing $\mathbf{W}$ to range from 0 to 1 without any constraints. The WA network flexibly learns the balance between $\mathbf{x}_{0,\epsilon,t}$ and $\mathbf{x}_{0,E}$. In the inference process, with low values of $\mathbf{W}$, $\mathbf{x}_{0,t}$ becomes overwhelmingly dependent on $\mathbf{x}_{0,\epsilon,t}$. In this scenario, $\mathbf{x}_{0,t}$ may incorporate a substantial amount of inaccurate information from $\mathbf{x}_{0,\epsilon,t}$. To address this concern, we set the limiting factor η to a value more than 0 to restrict the range of values for $\mathbf{W}$. In theory, η serves to control the maximum influence that $\mathbf{x}_{0,\epsilon,t}$ can exert based on the reference visual prior refinement. Our evaluation of various η values, including $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, reveal interesting insights. We note that our DE achieves the highest PSNR when η is set to 0.1, while the SSIM is maximized when η is set to 0.3, as illustrated in Fig. 11. In summary, our DE appears to yield the most favorable results when η is set to 0.3, achieving the performance balance between structural detail and global contour preservation. This optimized setting of η ensures that both $\mathbf{x}_{0,E}$ and $\mathbf{x}_{0,\epsilon,t}$ contribute effectively to the cloud-removed image generation process.

TABLE VII

RESULT OF MEMORYNET TRAINED WITH THE DIFFERENT RESOLUTION IMAGES. TRAINING AND TESTING REPRESENT THE SPATIAL RESOLUTION OF IMAGES FOR TRAINING AND TESTING

Training	Testing	PSNR	SSIM	LPIPS
2m	1m	22.773	0.6238	0.0438
2m	0.5m	21.393	0.5934	0.0529
1m	1m	24.819	0.7098	0.0392
1m	0.5m	22.575	0.6503	0.0443

TABLE VIII

VERTICAL ABLATION STUDY ON THE RICE2 WITH MSDA-CR

No.	Image size	Diffusion Fix	WA	PSNR	SSIM	LPIPS
1	64	No		31.408	0.8454	0.0518
2	256	Yes	✓	33.573	0.8792	0.0465
3	256	No	✓	33.598	0.8842	0.0452

E. Gap Between the High-Resolution and Low-Resolution Datasets

We perform extra experiments to demonstrate the significant influence of differences in data resolution on the model's performance. Essentially, models trained with low-resolution images yield less favorable results when tested on high-resolution datasets. This emphasizes the necessity for an ultra-resolution CR dataset.

Our approach initiates by training the model with images of various resolutions in the same size, followed by evaluating its performance on high-resolution sets. Specifically, we resize our 512 × 512 images from 0.5 m to different spatial resolution, such as {1m, 2m} and crop all of them as 128 × 128 to train the model. Following the training phase, we utilize corresponding crop sizes 128 × 128 from the original and resized images with 0.5 and 1 m spatial resolution to assess the impact of resolution on the ultimate CR results. As depicted in Table VII, all metrics show degradation as the training image resolution decreases. When comparing the performance between the training spatial resolution of 1 and 2 m on 0.5 m test set, we observe a decrease of 1.2 dB in PSNR, 0.06 in SSIM, and 0.008 in LPIPS. These experimental findings underscore the importance of our efforts to construct an ultra-resolution CUHK-CR dataset.

F. Ablation Study

Table VIII illustrates the results of an ablation study that explores the impact of the coarse-to-fine training strategy, WA, and reference visual prior. The results are presented in the order of training steps, with all outcomes evaluated using the same test set comprising images of size 256 × 256. No. 1 represents the outcome of training the diffusion model solely with small images of 64 × 64, while No. 2 denotes the

TABLE IX
HORIZONTAL ABLATION STUDY ON THE RICE2 WITH MSDA-CR

Prior	Coarse-to-fine	WA	PSNR	SSIM	LPIPS
			32.667	0.8680	0.0480
✓			33.239	0.8832	0.0457
✓	✓		33.215	0.8835	0.0457
✓		✓	33.512	0.8835	0.0466
✓	✓	✓	33.598	0.8842	0.0452

results based on the pretrained model from No. 1, which is further trained with regular-sized images of 256 × 256 on WA. No. 3 is the final result, where the WA and the diffusion model are jointly fine-tuned with regular-sized images based on the weight from No. 2. In comparison to No. 1 and No. 2, the inclusion of WA and reference visual prior refinement results in a remarkable improvement of nearly 2.1 dB, 0.034, and 0.005 in terms of PSNR, SSIM, and LPIPS, respectively. The fine-tuning process on regular-sized images has a lesser impact on PSNR and LPIPS but contributes more significantly to SSIM with a 0.005 improvement. These experimental results emphasize the advantageous role of coarse-to-fine training strategy, WA, and reference visual prior in the training order.

In the previous paragraph, we illustrated the improvements achieved through our three-stage experimental process. Here, we conduct a horizontal comparison by presenting results without the reference visual prior, coarse-to-fine training strategy and the WA in Table IX. In the first row, the result refers to the pure diffusion model trained on normal-sized images and evaluated without the reference visual prior. In the second row, the reference visual prior is incorporated into the pure diffusion model. In the third row, the coarse-to-fine training strategy is introduced. The fourth row presents outcomes from models with WA trained solely with normal-sized images, excluding the coarse-to-fine training strategy. All the experiments with the reference visual prior but without WA replace WA with a simple linear combination using a fixed parameter of 0.5. In other words, both $\mathbf{x}_{0,\epsilon,t}$ and $\mathbf{x}_{0,\mathbf{E}}$ each contributes half to $\mathbf{x}_{0,t}$ at any time step. As indicated in Table IX, comparing the first and second rows, the incorporation of the reference visual prior leads to improvements of about 0.6 dB, 0.015, and 0.002 in PSNR, SSIM, and LPIPS, respectively. Moreover, the similarity in results between the second and third, fourth and fifth rows demonstrates that the coarse-to-fine training strategy effectively reduces computational costs without precision loss. Finally, when comparing the third and fifth rows, the addition of the WA results in an enhancement of nearly 0.4 dB in PSNR. This horizontal comparison objectively highlights the advantages of the reference visual prior, WA, and the coarse-to-fine training strategy.

G. Computational Complexity Analysis

We conduct a thorough comparison of the computational complexity among the models, in terms of model complexity, memory usage, parameter count, and processing speed. The specific details are presented in Table X. The results demonstrate that our model achieves superior results without significantly increasing computational complexity.

TABLE X
COMPUTATIONAL COMPLEXITY OF THE COMPARED METHODS

Models	Complexity (G)	Memory (MB)	Parameters (M)	Speed (FPS)
SpAGAN	33.97	6196	0.22	29.37
AMGAN-CR	96.97	5668	0.24	28.94
CVAE	92.88	6068	15.56	37.73
MN	1097.30	11822	3.64	17.61
MSDA-CR	106.89	11318	3.91	17.80
DE	397.65	13112	36.80	12.04

VI. CONCLUSION

In this article, the DE method is introduced for reconstructing cloud-removed images. DE incorporates the diffusion architecture under the basis guidance of an reference visual prior, aiming to capture the merits of progressive diffusion process and end-to-end network to achieve both fine-grained detailed reconstruction and excellent global context modeling. To adaptively fuse the information from both branches, a WA network is trained to make balance based on their outputs across the whole denoising steps. Additionally, a coarse-to-fine training strategy is employed to accelerate convergence while obtaining superior results within a limited number of iterations. Finally, we introduce an ultra-resolution benchmark that provides a new basis with well-defined spatial landscape textures to train and evaluate the performance of CR models. Our experimental results on both the RICE and our CUHK-CR datasets demonstrate its superior performance. For future works, various conditions, such as feature maps of cloudy images and semantic information, may substitute the cloudy image to offer improved guidance for constructing more effective diffusion models.

REFERENCES

[1] X. Zhang, W. Yu, and M.-O. Pun, "Multilevel deformable attention-aggregated networks for change detection in bitemporal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621518.

[2] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417.

[3] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[4] M. Xu, X. Jia, M. Pickering, and S. Jia, "Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 215–225, Mar. 2019.

[5] G. Hu, X. Li, and D. Liang, "Thin cloud removal from remote sensing images using multidirectional dual tree complex wavelet transform and transfer least square support vector regression," *J. Appl. Remote Sens.*, vol. 9, no. 1, Sep. 2015, Art. no. 095053.

[6] 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, 2022, Art. no. 5303015.

[7] 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.

[8] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang, "Blind cloud and cloud shadow removal of multitemporal images based on total variation regularized low-rank sparsity decomposition," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 93–107, Nov. 2019.

[9] J. Wang, P. A. Olsen, A. R. Conn, and A. C. Lozano, "Removing clouds and recovering ground observations in satellite image sequences via temporally contiguous robust matrix completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2016, pp. 2754–2763.

[10] 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.

[11] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

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

[13] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, "Denoising diffusion probabilistic feature-based network for cloud removal in Sentinel-2 imagery," *Remote Sens.*, vol. 15, no. 9, p. 2217, Apr. 2023.

[14] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[15] D. Ma, R. Wu, D. Xiao, and B. Sui, "Cloud removal from satellite images using a deep learning model with the cloud-matting method," *Remote Sens.*, vol. 15, no. 4, p. 904, Feb. 2023.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[17] J. Sui, X. Ma, X. Zhang, and M. O. Pun, "GCRDN: Global context-driven residual dense network for remote sensing image super-resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4457–4468, May 2023.

[18] J. Sui, X. Ma, X. Zhang, and M.-O. Pun, "DTRN: Dual transformer residual network for remote sensing super-resolution," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jun. 2023, pp. 6041–6044.

[19] Y. Ma, H. Yang, W. Yang, J. Fu, and J. Liu, "Solving diffusion ODEs with optimal boundary conditions for better image super-resolution," 2023, *arXiv:2305.15357*.

[20] X. Tao, H. Gao, X. Shen, and J. Wang, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 8174–8182.

[21] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4641–4650.

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[23] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Sci. Rev.*, vol. 232, Sep. 2022, Art. no. 104110.

[24] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.

[25] 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. Photogramm. Remote Sens.*, vol. 197, pp. 1–17, Mar. 2023.

[26] 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*.

[27] 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, p. 157, Jan. 2021.

[28] 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, 2022.

[29] S. Ji, P. Dai, M. Lu, and Y. Zhang, "Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 732–748, Jan. 2021.

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

[31] 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, Mar. 2022, Art. no. 112902.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2022, pp. 10684–10695.

[33] Y. Benny and L. Wolf, "Dynamic dual-output diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11472–11481.

[34] X. Zou et al., "DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images," 2023, *arXiv:2308.04417*.

[35] X. Zhao and K. Jia, "Cloud removal in remote sensing using sequential-based diffusion models," *Remote Sens.*, vol. 15, no. 11, p. 2861, May 2023.

[36] B. Fei et al., "Generative diffusion prior for unified image restoration and enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9935–9946.

[37] D. Zhou, Z. Yang, and Y. Yang, "Pyramid diffusion models for low-light image enhancement," 2023, *arXiv:2305.10028*.

[38] S. Mohajerani and P. Saeedi, "Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 1029–1032.

[39] S. Mohajerani, T. A. Krammer, and P. Saeedi, "A cloud detection algorithm for remote sensing images using fully convolutional neural networks," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2018, pp. 1–5.

[40] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. NIPS*, vol. 34, 2021, pp. 8780–8794.

[41] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[42] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[43] X. F. Zhang, C. C. Gu, and S. Y. Zhu, "Memory augment is all you need for image restoration," 2023, *arXiv:2309.01377*.

[44] 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, pp. 1–5, 2022.

Jialu Sui received the B.S. degree in computer science and technology from Shandong University, Weihai, China, in 2021. She is currently pursuing the M.Phil. degree in computer and information engineering with The Chinese University of Hong Kong, Shenzhen, China.

Her research interests include remote sensing and deep learning.

Yiyang Ma received the B.S. degree in artificial intelligence from Peking University, Beijing, China, in 2022, where he is currently pursuing the M.S. degree with the Wangxuan Institute of Computer Technology.

His research interests include generative models and image enhancement.

Wenhan Yang (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018, respectively.

He is currently an Associate Researcher with the PengCheng Laboratory, Shenzhen, Guangdong, China. He has authored over 50 technical articles in refereed journals and proceedings and holds nine granted patents. His research interests include image/video processing/restoration, bad weather restoration, and human–machine collaborative coding.

Dr. Yang received the 2023 IEEE Multimedia Rising Star Runner-Up Award, the IEEE ICME-2020 Best Paper Award, the IFTC 2017 Best Paper Award, the IEEE CVPR-2018 UG2 Challenge First Runner-Up Award, and the MSA-TC Best Paper Award of ISCAS 2022. He was the Candidate of CSIG Best Doctoral Dissertation Award in 2019.

Xiaokang Zhang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018.

From 2019 to 2022, he was a Post-Doctoral 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. He has authored or coauthored more than 40 scientific publications in international journals and conferences. His research interests include remote sensing image analysis, computer vision, and machine learning.

Dr. Zhang is currently a reviewer for more than 20 renowned international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and 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, Hong Kong, 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), Los Angeles, CA, USA, in 2006.

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

Dr. Pun is the Founding Chair of the IEEE ComSoc-SPS joint chapter in Shenzhen. He has served as the Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the Guest Editor for *Remote Sensing*. He was the APSIPA Distinguished Lecturer from 2022 to 2023.

Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2010.

She was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. She was a Visiting Researcher with Microsoft Research Asia, Beijing, in 2015, supported by the Star Track Young Faculties Award. She is currently an Associate Professor and a Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University. She has authored more than 100 technical articles in refereed journals and proceedings and holds 70 granted patents. Her research interests include multimedia signal processing, compression, and computer vision.

Dr. Liu is a Senior Member of CSIG and a Distinguished Member of CCF. She has served as a member of the Multimedia Systems and Applications Technical Committee (MSA TC) and the Visual Signal Processing and Communications Technical Committee (VSPC TC) in the IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and IEEE MMSP 2015 Top10% Paper Award. She was the Technical Program Chair of ACM MM Asia-2023, IEEE ICME-2021, ACM ICMR-2021, and IEEE VCIP-2019; the Area Chair of CVPR-2021, ECCV-2020, and ICCV-2019; the ACM ICMR Steering Committee Member; and the CAS Representative at the ICME Steering Committee. She was the APSIPA Distinguished Lecturer from 2016 to 2017. She has served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS SYSTEMS FOR VIDEO TECHNOLOGY, and *Journal of Visual Communication and Image Representation*.