Robust Haze and Thin Cloud Removal via Conditional Variational Autoencoders

Haidong Ding, Fengying Xie, *Member, IEEE*, Linwei Qiu, Xiaozhe Zhang, *Graduate Student Member, IEEE*, and Zhenwei Shi, *Senior Member, IEEE*

***Abstract*— Existing methods for remote-sensing image dehazing and thin cloud removal treat this image restoration task as a clear pixel estimation problem, yielding a single prediction result through a deterministic pipeline. However, image restoration is a highly ill-posed problem, as the sharp pixel value corresponding to the input cannot be uniquely determined solely from the degraded image. In this article, we present a novel algorithm for haze and thin cloud removal using conditional variational autoencoders (CVAEs) to generate multiple realistic restored images for each input. By sampling from the latent space to capture the pixel diversity, the proposed method mitigates the limitations arising from inaccuracies in a single estimation. In this uncertainty pipeline, we can generate a more accurate restored image based on these multiple predictions. Furthermore, we have developed a dynamic fusion network (DFN) for combining multiple plausible outcomes to obtain a more accurate result. DFN dynamically predicts the kernels used for restored result generation conditioned on inputs, improving haze and thin cloud thanks to its adaptive nature. Quantitative and qualitative experiments demonstrate that the proposed method outperforms existing state-of-the-art techniques by a significant margin on dehazing and thin cloud removal benchmarks.**

***Index Terms*— Conditional variational autoencoders (CVAEs), remote-sensing image dehazing, thin cloud removal.**

I. Introduction

IMAGES captured by remote-sensing satellites often suffer from absorption and scattering effects caused by haze and thin clouds, which ultimately leads to image degradation. The low quality of these images hampers their usefulness for subsequent high-level computer vision tasks, such as object detection [1], [2], [3], segmentation [4], [5], [6], image super-resolution [7], [8], [9], [10], and environmental protection [11], [12], [13]. Therefore, it is crucial to develop an effective method for removing haze and thin clouds from single remote-sensing images. During the imaging process, clouds may obscure ground scenes, affecting image quality, and the accuracy of analysis results. Cloud removal processing endeavors to detect and eliminate clouds within images, enhancing the visibility of ground-level information. Haze in the atmosphere causes blur and distortion of scenes in images. Dehazing technology aims to eliminate image blurring and low-contrast issues, making images more realistic.

To address this issue, numerous experts have proposed various methods. These methods can generally be classified into two categories: prior-based and data-driven approaches. Prior-based cloud removal models [14], [15], [16], [17], [18] are primarily based on the atmospheric scattering model, incorporating different physical priors from image statistics. However, these prior-based methods may not perform effectively when the statistical prior does not hold in real-world images.

To address the accuracy limitations of prior-based methods, data-driven approaches employ deep learning techniques to train networks using a supervised learning paradigm. Several techniques [19], [20], [21], [22], [23] utilize the strong data fitting capabilities of neural networks to directly generate clear images from their corresponding degraded counterparts in an end-to-end fashion. These algorithms [24], [25] are trained on extensive datasets and can yield satisfactory outcomes. However, directly learning the mapping relationship from low-quality images to clear images will result in limited interpretability. To mitigate this problem, other approaches [26], [27], [28] integrate convolutional neural networks (CNNs) with the imaging model. These approaches primarily concentrate on constructing a neural network with trainable parameters to substitute a portion of the physical model used in conventional methods. Data-driven-based methods leverage the robust representation capabilities of neural networks and depend on substantial amounts of training data, enabling them to consistently outperform prior-based methods. Consequently, they have emerged as the dominant approach for remote-sensing image restoration. Despite the remarkable results achieved by numerous outstanding studies, the task of haze and thin cloud removal from single image still presents various challenges and misconceptions. Hence, it is imperative to approach this problem from a broader perspective, including a reevaluation of its ill-posed nature. When a remote-sensing image becomes degraded, it loses crucial scene radiance information, and attempting to restore a fully clear image from such limited information renders the problem highly ill-posed [29], [30], [31]. Consequently, this low-level computer vision problem inherently involves uncertainty. Obtaining the exact pixel values in the clear image solely from the degraded image, without additional auxiliary information, is not feasible. To the best of our knowledge, none of the existing methods consider this aspect;

Manuscript received 30 November 2023; revised 25 December 2023; accepted 30 December 2023. Date of publication 16 January 2024; date of current version 22 January 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1510905 and in part by the National Natural Science Foundation of China under Grant 61871011. *(Corresponding author: Fengying Xie.)*

The authors are with the Department of Aerospace Information Engineering, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: xfy_73@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3349779

instead, they all rely on establishing a direct mapping from a degraded image to a restored image. Hence, incorporating the notion of uncertainty holds significant potential to enhance the performance of thin cloud removal algorithms.

Following this idea, we propose an uncertainty framework via conditional variational autoencoders (CVAEs) for remote-sensing image haze and thin cloud removal. Based on the above analysis of uncertainty, we tackle this problem from the perspective of multisolution. Each time the output from the proposed method is a sample of possible solutions. Overall, we develop a simple, yet effective multi-input multi-output (MIMO) U-shaped architecture, consisting of a stack of residual blocks (ResBlocks). For more efficient feature fusion, we introduce a selective feature fusion module (SFFM), which works among the channel dimension and leverage a selective mechanism to fuse intermediate features generated by the encoder and decoder. Furthermore, we propose a scheme to fuse multiple reasonable solutions to obtain a more accurate solution. We design a dynamic fusion network (DFN) to establish the mapping relationship between multiple reasonable solutions and final restored result. The DFN conducts convolution operations in a dynamic manner: the kernels are predicted dynamically and conditioned on the input. The diverse output and the proposed dynamic fusion strategy can ultimately enhance the generalization ability of the proposed thin cloud removal network.

This article presents an expanded version of our previous conference paper [32]. We extend our approach by introducing a dynamic fusion network that combines multiple plausible solutions to achieve a more refined restoration result. Furthermore, we broaden the scope of our method by applying it to remote-sensing image dehazing, allowing us to thoroughly validate the superior performance of our proposed algorithm beyond thin cloud removal, which was the focus of our conference paper.

The main contributions of the proposed method can be summarized as follows.

1) To address the inherent uncertainty in haze and thin cloud removal, we present a probabilistic model based on CVAE for restoring remote-sensing images. This approach tackles the challenge of multiple possible solutions by considering the problem from a probabilistic standpoint. The network generates multiple interpretable results, accommodating the inherent variations in potential solutions to the problem.
2) We propose an MIMO U-shaped architecture to restore the degraded images and introduce an SFFM to fusion the intermediate features based on a selective mechanism.
3) We propose an DFN to fuse multiple solutions resulting from the proposed uncertainty framework in a dynamic manner. It dynamically predicts convolution kernels to process different inputs, resulting in improved flexibility and robustness.
4) We introduce a new benchmark dataset specifically designed for single image thin cloud removal. The dataset consists of pairs of cloud and clear images captured at different instances of the same real scene.

II. Related Work

A. Haze and Thin Cloud Removal

The existing haze and thin cloud removal algorithms can be broadly classified into two categories: prior-based approaches and data-driven approaches.

1) Prior-Based Methods: Most conventional methods are based on the physical prior. They estimate some important quantities in the imaging model (e.g., the transmission map) and then recover a clear image from its degraded counterpart. Chavez Jr. [33] proposed an additive model to describe the generation principle of low-quality images under the assumption that the distance between the sensor and the ground is fixed. He et al. [16] proposed a dark channel prior based on statistical laws, showing that the pixel value of one or more color channels tends to zero in the nonsky area of the image, which is used to estimate the transmission map. Fattal [34] proposed a color-lines prior to estimate the transmission map based on the distribution of images in the RGB color space. Berman et al. [30] assumed that the color of a clear image can be approximated by hundreds of distinct colors and proposed a dehazing algorithm based on this novel nonlocal prior. Xu et al. [35] proposed a method based on signal transmission and airspace hybrid analysis, combined with atmospheric scattering theory to remove clouds. While prior knowledge-based methods demonstrate superior statistical properties in specific scenarios, they are prone to failure in real-world images where physical assumptions are not applicable.

2) Data-Driven-Based Methods: In recent years, the establishment of large-scale datasets and advancements in deep-learning techniques have led to the emergence of numerous data-driven supervised methods for haze thin cloud removal, aiming to address the limitations of traditional approaches. Mao et al. [36] proposed a deep encoder–decoder framework, which uses the multilayer convolution and deconvolution operators, and adds skip connections to improve the efficiency of image restoration. Singh and Komodakis [37] proposed an adversarial training-based network named cloud removal using a cyclic consistent generative adversarial network (Cloud-GAN) to directly learn the mapping relationship between cloudy and clear images. Qin et al. [38] described thin clouds as haze coverings for each band and used a multiscale deblurring CNN with the residual structure to remove the clouds. Li et al. [20] proposed RSC-Net, using an end-to-end residual symmetric connection network for thin cloud removal, which estimates cloud-free results directly from cloud images. Xu et al. [19] introduced a generative adversarial network based on the attention mechanism, using an adaptively generated attention map of the recurrent network to guide the network focus on more valuable matters. Ding et al. [39] introduced a compact thin cloud removal network utilizing a feedback mechanism that enables gradual improvement of the restoration outcome. Zi et al. [40] proposed a novel wavelet-integrated CNN, named WaveCNN-CR, designed specifically for thin cloud removal in remote-sensing images. This network achieves a larger receptive field, ensuring no information loss during the process.

Building upon traditional methods, some researchers integrate deep-learning technology with physical models to enhance the accuracy of the networks and achieve more precise results. Cai et al. [27] showed that medium transmission estimation can be reformulated as a learnable end-to-end system. Ren et al. [41] used deep learning to learn the transmission map and solve atmospheric scattering models. Zhang and Patel [42] proposed the densely connected pyramid dehazing network (DCPDN), a model that can estimate the transmission map and atmospheric light simultaneously. Moreover, they introduced a joint-discriminator to enhance the details in the resulting images. Zheng et al. [43] combined the atmospheric scattering model with UNet, learned the necessary cloud thickness distribution map and directly used UNet to remove thin clouds. According to the additive model of cloud images, Zi et al. [26] utilized deep neural networks combined with the imaging model to achieve thin cloud removal.

Despite the significant progress that data-driven methods have made in enhancing haze and thin cloud removal performance, their results are limited to a one-to-one mapping with respect to the input image. This indicates that they have identified a single reasonable solution from the multiple potential solutions to the multisolution problem. Differing from existing methods, we approach this restoration task as an indeterminate solution problem. To address this challenge effectively, we integrate haze and thin cloud removal with CVAE to provide a more robust solution to this ill-posed problem.

B. Conditional Variational Autoencoders

Following the pioneering contributions of Kingma and Welling [44] and Rezende et al. [45], VAEs and CVAEs have gained widespread usage across a range of computer vision tasks, such as image generation [46], data augmentation [47], and modeling inherent ambiguities of the image [48], [49], [50]. Since the framework analyzes problems from the view of probabilistic, several works introduce CVAE to deal with uncertainty in vision problems. In the image saliency detection, Zhang et al. [50] proposed a network structure based on the CVAE framework to generate probabilistic saliency maps according to the uncertainty of human perception of natural scenes. Our proposed algorithm utilizes the CVAE framework to infer the latent variable from the low-quality input images, combining the intermediate features to generate multiple plausible results.

C. Dynamic Neural Network

In traditional convolutional layers, the learned kernels are static after training and do not adapt to different inputs. Consequently, the uniform parameters across various input images constrain the representation capability of deep networks. Therefore, recent work [51] has explored the idea of introducing more flexibility into network architectures. Li et al. [1] introduced a seminal method known as the spatial transformer (ST), which enables adaptive feature map transformations based on the input data. ST incorporates a localization network to generate transformation parameters that are applied to the feature maps, resulting in the recovery of the input through corresponding variations. Importantly, ST can be seamlessly incorporated into existing convolutional architectures without the need for additional supervision during optimization. Jia et al. [52] proposed the DFNs which utilize a filter generation network to predict the parameters of a convolution layer. This approach enhances flexibility and adaptability by allowing the network to dynamically adjust the parameters of the filters based on the input data. Deformable convolutional networks (DCNs) [53] have been proposed to address the limitations of shape-fixed convolutions. DCNs dynamically adapt the sampling locations by predicting offsets for each location, enabling a more accurate capture of complex spatial patterns in the input features. For vision tasks, the application areas of dynamic networks have been further expanded, including image recognition [54], [55], [56], [57], [58], image segmentation [59], [60], [61], [62], [63], [64], objection detection [65], [66], [67], [68], super resolution [69], [70], [71], image restoration [24], [72], and more.

In our algorithm, we utilize the dynamic networks to enhance the performances of haze and thin cloud removal. Leveraging the capability of our proposed uncertainty framework to generate multiple plausible solutions, we employ a dynamic fusion scheme to individualize each sample. This approach aims to yield more precise restoration results and enhance the effect of haze and thin cloud removal in remote-sensing images.

III. Methodology

A. Overview

Fig. 1 presents the whole pipeline of the proposed method during training and testing. Let $\mathcal{D} = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{N}$ be the training dataset, in which $\mathbf{X}_i$ denotes the cloudy and hazy image obtained by the satellite sensor, $\mathbf{Y}$ is the clear ground scene information, and N denotes the total number of image pairs in the dataset. Our network consists of the following main modules: 1) prior network and posterior network, which map input $\mathbf{X}_i$ (for prior network) or input $\mathbf{X}_i$ and $\mathbf{Y}_i$ (for posterior network) to low-dimensional latent variable z_i; 2) the restoration network that employs the latent variable z_i and the degraded input $\mathbf{X}_i$ to restore a reasonable clear image $\mathbf{Y}_i$; and 3) the dynamic fusion network that fuses multiple reasonable outputs into a more accurate solution during the test phase.

Our framework integrates a CVAE that can generate multiple clear image candidates for a degraded remote-sensing image instead of a single prediction. Within the prior network and posterior network, a low-dimensional latent space is responsible for encoding potential latent variables z. The restoration network then takes a random sample from this latent space, alongside the degraded input $\mathbf{X}$, to generate the corresponding sharp image. This architecture's main feature lies in its capacity to implement one-to-many mapping. Unlike previous frameworks that only produce a single clear result, our algorithm can recover multiple reasonably clear images. These multiple results are then fused through the dynamic fusion network to enhance the robustness of haze and thin cloud removal.

In the standard CVAE pipeline, the prior distribution is modulated as a Gaussian distribution with parameters that

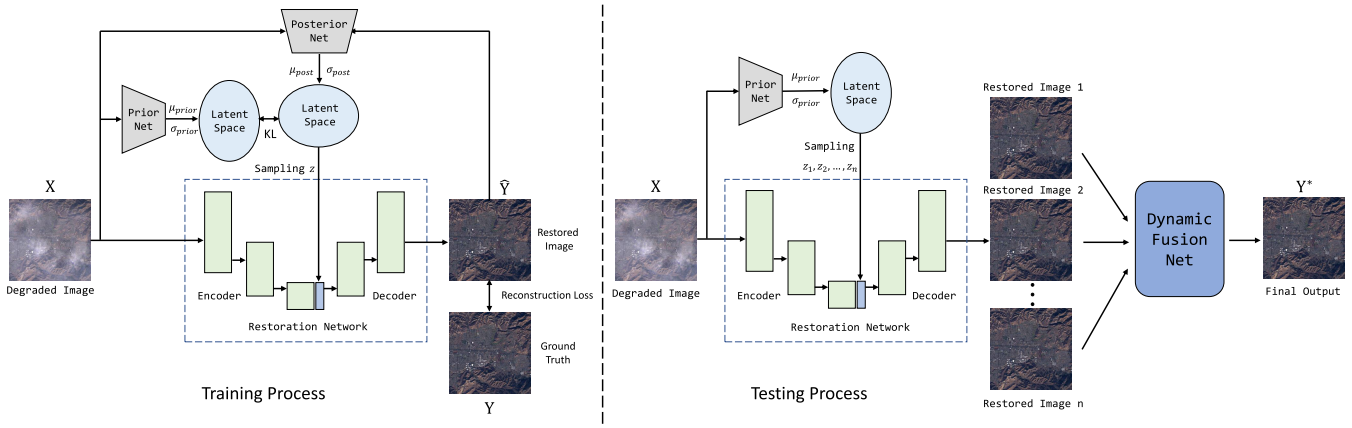


Fig. 1. Training and testing pipeline of the proposed algorithm.

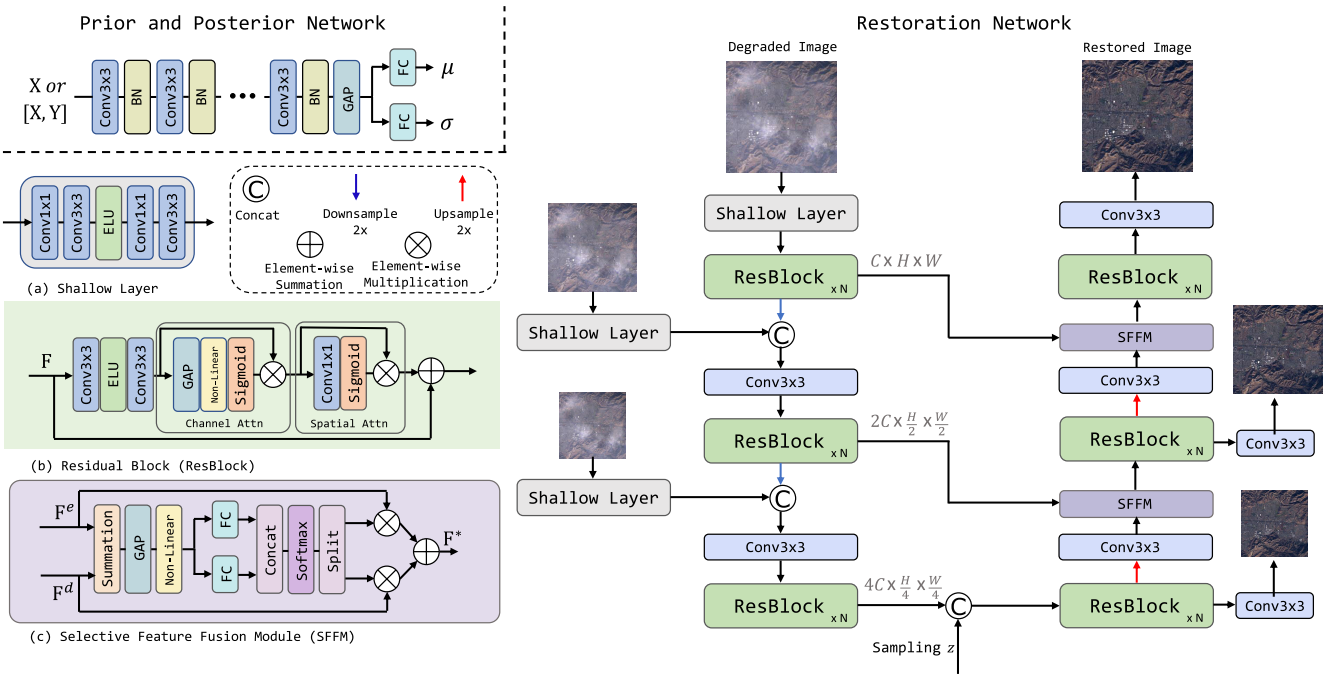


Fig. 2. Detailed architecture of the prior and posterior network and the restoration network. 1) Posterior network shares the same structure as the prior network, employing BN and GAP. 2) Restoration network is built on an MIMO U-shaped framework, where (a) shallow layer captures shallow features from the input. (b) ResBlock incorporates an attention mechanism to enhance informative feature channels and regions. Additionally, (c) SFFM combines intermediate features from the encoder and decoder.

are conditioned on the input data $\mathbf{X}$. In our haze and thin cloud removal framework, there are three types of variables: conditioning variable $\mathbf{X}$ (the hazy and cloudy image), latent variable z, and restored output $\mathbf{Y}$. To restore a clear image, the latent variable z is drawn from the Gaussian distribution $P_\theta(z|\mathbf{X})$ and then the restored output $\mathbf{Y}$ is generated from $P_\omega(\mathbf{Y}|\mathbf{X}, z)$. The posterior of z is formulated as $Q_\phi(z|\mathbf{X}, \mathbf{Y})$. The variational lower bound of the model is as follows:

$$\begin{aligned}\log P(\mathbf{Y}|\mathbf{X}) &= \int_z Q_\phi(z|\mathbf{X}, \mathbf{Y}) \log P(\mathbf{Y}|\mathbf{X}) dz \\ &\geq -D_{\mathrm{KL}}(Q_\phi(z|\mathbf{X}, \mathbf{Y})||P_\theta(z|\mathbf{X})) \\ &\quad + \mathbb{E}_{Q_\phi(z|\mathbf{X},\mathbf{Y})}[\log P_\omega(\mathbf{Y}|\mathbf{X}, z)] \end{aligned} \tag{1}$$

where $P_\omega(\mathbf{Y}|\mathbf{X}, z)$ is the likelihood of $P(\mathbf{Y})$ given the latent variable z and degraded input data $\mathbf{X}$. θ, ϕ, and ω represent the parameter set of distribution. Our CVAE framework is composed of a Prior Network $P_\theta(z|\mathbf{X})$, a Posterior Network $Q_\phi(z|\mathbf{X}, \mathbf{Y})$, and a Restoration Network $P_\omega(\mathbf{Y}|\mathbf{X}, z)$. The Kullback–Leibler (KL) Divergence $D_{\mathrm{KL}}(Q_\phi(z|\mathbf{X}, \mathbf{Y})||P_\theta(z|\mathbf{X}))$ work as a regularization loss to narrow the gap between the prior $P_\theta(z|\mathbf{X}))$ and posterior $Q_\phi(z|\mathbf{X}, \mathbf{Y})$.

B. Prior and Posterior Network

One of the core components of our architecture is a low-dimensional latent space $\mathbb{R}^N$ ($N = 9$ in our experiments). We define $P_\theta(z|\mathbf{X}))$ as the Prior Network that maps the input hazy and cloudy data $\mathbf{X}$ to a low-dimensional latent feature space. The Posterior Network $Q_\phi(z|\mathbf{X}, \mathbf{Y})$ has the same structure as the Prior Network. As shown in Fig. 2 (top

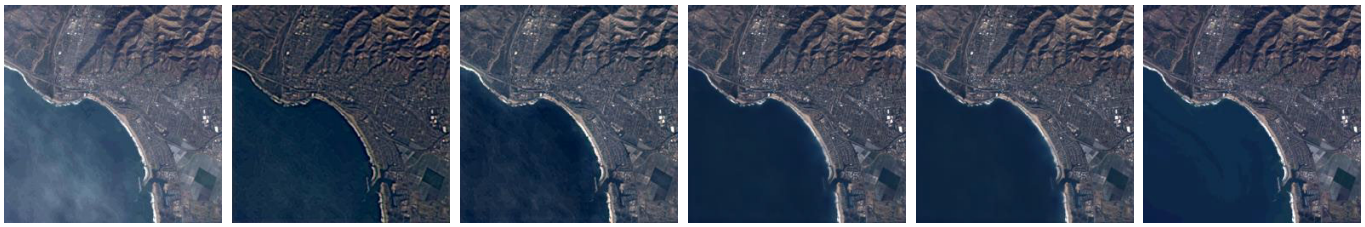


Fig. 3. Results of the proposed algorithm. Our method can achieve a one-to-many mapping.

left), we use `Conv3x3` and batch normalization (`BN`) layers to extract the features of input data $\mathbf{X}$, and then two fully connected layers are employed to estimate the mean μ and standard deviation σ of the prior latent Gaussian variable. The prior and posterior probability distribution is modeled as an axis-aligned Gaussian, for example, $z \sim \mathcal{N}(\mu, \text{diag}(\sigma^2))$.

To allow the network can be trained using the gradient descent algorithm, z is drawn with the reparameterization trick, which is written as (4)

$$z = \mu + \sigma \cdot \varepsilon \tag{2}$$

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This trick allows error backpropagation through the Gaussian latent variables, which is essential in the training process.

During training, we sample from the latent space determined by the Posterior Network to recover a reasonably sharp image. The KL divergence $D_{\text{KL}}(Q_\phi(z|\mathbf{X}, \mathbf{Y})||P_\theta(z|\mathbf{X}))$ penalizes the difference between the posterior distribution $Q_\phi(z|\mathbf{X}, \mathbf{Y})$ and the prior distribution $P_\theta(z|\mathbf{X}))$, pulling the posterior and prior distributions toward each other. In this way, we can sample from the prior distribution and get a similar hidden variable z to restore the image during the test. The whole pipeline of the algorithm during training and testing is illustrated in Fig. 1.

C. Restoration Network

The restoration network comprises three encoders, each taking the downsampled degraded image at different scales (i.e., original scale, 1/2, and 1/4) as inputs. Additionally, there are three decoders that utilize a 3×3 convolution layer to predict the corresponding scale output. The MIMO mechanism is employed to alleviate the difficulty of training [73]. Specifically, each encoder and decoder comprises multiple ResBlocks [refer to Fig. 2(b), and $N = 3$ in our experiments]. To strengthen the connection between the encoder and the decoder, we employ feature-level fusion using the SFFM [refer to Fig. 2(c)]. Downsampling and upsampling operations are achieved through stride and transpose convolutions, respectively.

1) Residual Block: Let $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ be the input feature map of ResBlock, where C is the number of channels and $H \times W$ represents the spatial resolution. Small-kernel convolutions have been observed to play a crucial role in the performance of UNet-like networks. As a result, we employ `Conv3x3` layers and activation functions for feature extraction. To facilitate the propagation of more informative features, channel attention and spatial attention mechanisms are adopted to enhance the representational capability of the spatial branch. Channel attention [74] is achieved using global average pooling (GAP), nonlinear projection (`Conv1x1-ELU-Conv1x1`), and the sigmoid activation function. Spatial attention is implemented using `Conv1x1` and sigmoid to generate an attention map with dimensions of $[1, H, W]$. The final output of the spatial branch is obtained through elementwise multiplication.

2) Selective Feature Fusion Module: Feature aggregation is a widely adopted technique that facilitates the training of deep networks using gradient-based methods, commonly employing simple concatenation and summation operations [75], [76]. However, this aggregation strategy may not adequately enhance the adaptation ability of neurons [77]. Inspired by selective kernel networks (SKNets) [77], we propose the SFFM, which operates along the channel dimension to fuse the intermediate features generated by both the encoder and the decoder [refer to Fig. 2(c)].

Formally, given two intermediate feature maps, $\mathbf{F}^e$ and $\mathbf{F}^d$, we first perform the fuse operation as follows:

$$\mathbf{h} = \text{NL}\big(\text{GAP}(\mathbf{F}^e + \mathbf{F}^d)\big) \tag{3}$$

where NL represents the nonlinear projection. $\mathbf{h} \in \mathbb{R}^{1 \times C/r}$ is the fused feature and r is the reduction factor (with $r = 8$). And then we employ two separate fully connected layers to derive the weights for channelwise feature selection. Afterward, the weights corresponding to the same channels are normalized using a softmax operator, which can be formalized as follows:

$$\left[w_c^e, w_c^d\right] = \left[\frac{e^{\mathbf{E}_c\mathbf{h}}}{e^{\mathbf{E}_c\mathbf{h}} + e^{\mathbf{D}_c\mathbf{h}}}, \frac{e^{\mathbf{D}_c\mathbf{h}}}{e^{\mathbf{E}_c\mathbf{h}} + e^{\mathbf{D}_c\mathbf{h}}}\right] \tag{4}$$

where w_c^e and w_c^d are channelwise attention weights for features from the encoder and decoder, respectively. $\mathbf{E}^c, \mathbf{D}^c \in \mathbb{R}^{1 \times C/r}$ are the parameters of fully connected layers and c is the channel index. The channel weights for each feature map can be obtained through a split operation, and the features are recalibrated and aggregated utilizing elementwise multiplication and summation. The final aggregated feature, denoted as $\mathbf{F}^*$, is defined as follows:

$$\mathbf{F}^* = \mathbf{w}^e \cdot \mathbf{F}^e + \mathbf{w}^d \cdot \mathbf{F}^d. \tag{5}$$

D. Dynamic Fusion Network

The latent variables z enable the modeling of multiple modes, allowing the decoder network to effectively capture one-to-many mappings. Fig. 3 demonstrates that the clear

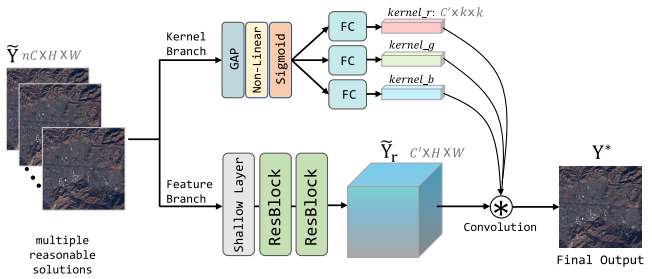


Fig. 4. Structure of the DFN consists of dual branches running in parallel. The feature branch takes the input and projects it into a $[C', H, W]$ feature map. Subsequently, fully connected layers are employed in the kernel branch to predict dynamic convolution kernels. The final output is generated through a convolution operation denoted by $\circledast$, where k represents the kernel size (with $k = 3$).

images obtained from different sampling results exhibit variations in terms of the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) scores. This diversity promotes the integration of multiple plausible solutions, resulting in a more accurate restoration. Consequently, we propose an DFN to establish the mapping relationship between these multiple reasonable solutions and the final restored result.

In traditional convolution layers, the learned kernels remain fixed and are independent of the input [52], [60]. Consequently, the kernel parameters are shared across arbitrary input images. However, since different degraded images exhibit variations in color distribution and degree of degradation, adhering to a fixed mode restricts the flexibility of the fusion model. To address this limitation, we propose the DFN which enhances feature representation and dynamically generates the final restored result. This is illustrated in Fig. 4.

The DFN is composed of a feature branch and a kernel branch. Specifically, given the concatenation (in channel dimension) of multiple reasonable solutions $\widetilde{\mathbf{Y}} \in \mathbb{R}^{nC\times H\times W}$, where $n = 6$ is the number of samples, the feature branch utilizes a shallow layer and two ResBlock to generate the refined feature $\widetilde{\mathbf{Y}_r} \in \mathbb{R}^{C'\times H\times W}$. Next, we predict convolution kernels using the kernel branch conditioned on $\widetilde{\mathbf{Y}}$. The kernel branch comprises three independent fully connected layers to predict the kernels for R, G, and B channels, respectively. The kernel prediction can be expressed as

$$\mathbf{K}_i = \mathrm{FC}_i(\sigma(\mathrm{NL}(\mathrm{GAP}(\widetilde{\mathbf{Y}})))). \tag{6}$$

Here, $\mathbf{K}_i$ is the predicted kernels, which are reshaped as $[C', k, k]$, σ denotes sigmoid function, and FC_i is the fully connected layer. The index i represents each channel (i.e., R, G, and B), and k is the kernel size.

To generate the final restored result, the DFN performs a dynamic convolution operation between the predicted kernels and refined features. This operation can be formulated as follows:

$$\mathbf{Y}^* = \mathbf{K} \circledast \widetilde{\mathbf{Y}_r}. \tag{7}$$

$\mathbf{Y}^*$ is the final removal result in RGB space and $\circledast$ indicates the convolution operation.

TABLE I
COMPARISON BETWEEN VARIOUS DATASETS

Dataset	Training	Testing	Image Size	Cloud Type
T-CLOUD	2,351	588	256 × 256	Real-world
RICE-I	400	100	512 × 512	Real-world
WHUS2-CR	4,000	1,000	256 × 256	Real-world
SateHaze1k	960	135	512 × 512	Synthetic
RS-Haze	51,300	2,700	512 × 512	Synthetic

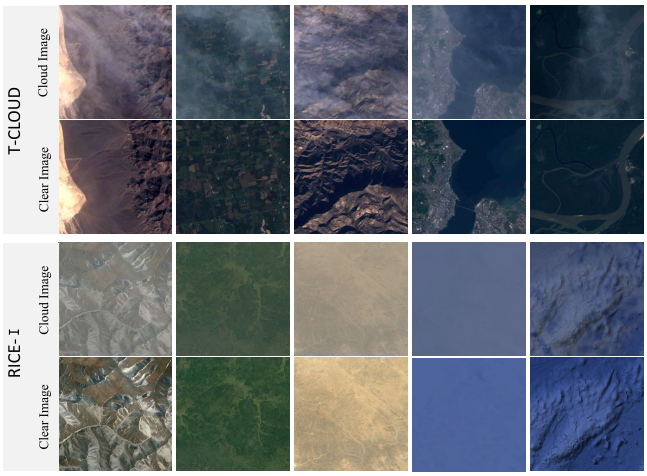


Fig. 5. Visual comparisons between our proposed T-CLOUD and RICE-I.

E. Loss Function

Fig. 1 illustrates that during the training stage, a single sampled latent variable z and degraded input $\mathbf{X}$ are employed to restore a clear image. The reconstruction loss is then calculated between this output and the ground truth. The DFN is solely used in the testing phase. To train the proposed network, we utilize a two-stage approach. In the first stage, we only sample a single latent variable z to train the prior network, posterior network, and restoration network. The objective function in this stage consists of the reconstruction loss and KL divergence. Subsequently, in the second stage, we freeze the network parameters of these three parts, generate training data for DFN through multiple random sampling, and solely utilize the reconstruction loss for DFN training. During testing, all subnetworks are combined to merge multiple potential solutions and achieve more precise and clear results.

For stage one, the losses are combined as a weighted sum with a weighting parameter β, formulated as

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = \beta \cdot D_{\mathrm{KL}}(Q_\phi(z|\mathbf{X}, \mathbf{Y})||P_\theta(z|\mathbf{X})) + \mathbb{E}_{Q_\phi(z|\mathbf{X},\mathbf{Y})}[-\log P_\omega(\mathbf{Y}|\mathbf{X}, z)]. \tag{8}$$

The first term in the loss function is the KL divergence and we set $\beta = 1.0$.

The second term is the reconstruction loss. In supervised training, the restoration performance can be quantified by counting the differences between the restoration network output $\hat{\mathbf{Y}}$ with its corresponding reference clear image $\mathbf{Y}$ under some proper loss L, for example, mean square error (mse).

In our method, we choose the combination of $\mathcal{L}_1$ loss and the frequency loss $\mathcal{L}_{\text{fre}}$ as the criterion to optimize the parameters

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_1 + \lambda_1 \mathcal{L}_{\text{fre}}. \tag{9}$$

We use $\lambda_1 = 0.1$ as the weighting factor. The $\mathcal{L}_1$ loss can be expressed as

$$\mathcal{L}_1 = \sum_{s=1}^{3} \omega_s \left\| \hat{\mathbf{Y}}_s - \mathbf{Y}_s \right\|_1 \tag{10}$$

where s denotes different scales (corresponds to multiple output), $\omega = [1.0, 0.3, 0.1]$ are the coefficients for each scale, and $\mathbf{Y}$ denotes the ground truth. $\mathcal{L}_{\text{fre}}$ is the frequency domain loss [73], [78], [79] that enforces high-frequency details

$$\mathcal{L}_{\text{fre}} = \sum_{s=1}^{3} \omega_s \left\| \mathcal{F}(\hat{\mathbf{Y}}_s) - \mathcal{F}(\mathbf{Y}_s) \right\|_1 \tag{11}$$

where $\mathcal{F}(\cdot)$ is the fast Fourier transform (FFT) operator.

For stage two, the reconstruction loss used to train the DFN can be expressed as

$$\mathcal{L}_{\text{dfn}} = \left\| \mathbf{Y}^* - \mathbf{Y} \right\|_1 + \lambda_1 \left\| \mathcal{F}(\mathbf{Y}^*) - \mathcal{F}(\mathbf{Y}) \right\|_1. \tag{12}$$

$\mathbf{Y}^*$ is the final fusion result, which is illustrated in (7).

IV. Experiments

A. T-CLOUD for Thin Cloud Removal

Obtaining image pairs containing both thin clouds and cloud-free regions in real-world scenes is a challenging task. Therefore, some previous algorithms [26], [38], [43] utilize a synthetic approach to construct image pairs containing cloud-contaminated and cloud-free regions for training purposes. However, there is a significant difference between simulated and real-world images, which can cause the network to learn the laws of data synthesis rather than the essence of image degradation during the optimization process.

To overcome the limitation of synthetic datasets for thin cloud removal, we collect a real scene image dataset called T-CLOUD. Both training and test sets are from Landsat 8 RGB images. Our dataset contains 2939 doublets of cloud images and their clear counterparts separated by one satellite reentry period (16 days). We select the image pairs which has similar lighting conditions and crop them into 256 × 256 patches. We split the dataset with a ratio of 8:2, with 2351 images in the training set and 588 images in the test set.

T-CLOUD is a novel benchmark dataset for single remote-sensing image declouding. Our dataset is different from the existing dataset RICE-I [80] in the following points. First, T-CLOUD is a large-scale dataset. Our dataset contains 2939 doublets of cloud images and their clear counterpart while RICE-I only contains 500 image pairs. The large-scale dataset can effectively improve the performance of the thin cloud removal algorithms. Second, the ground scenes in our dataset have much finer texture details. T-CLOUD includes many different ground scenarios such as cities, rivers, and deserts while RICE-I is relatively simple. Third, the thin clouds exhibited by T-CLOUD are nonhomogeneous which is consistent with the characteristics of remote-sensing images occluded by thin clouds. Fig. 5 shows some visual comparisons between T-CLOUD and RICE-I, it can be observed that our dataset is more realistic. Table I summarizes the similarities and differences between the two datasets.

It should be noted that our constructed dataset comprised cloudy and clear pairs captured by the same satellite sensor but at different times. The presence of illumination noise is inevitable due to changes in ambient light. Despite our efforts to choose images with as similar lighting conditions as possible, achieving substantial results on this dataset remained a formidable task.

B. Experiment Setting

1) Datasets: We evaluate our uncertainty-based framework on the proposed T-CLOUD, RICE-I, WHUS2-CR [81], SateHaze1k [82], and RS-Haze [25]. The RICE-I dataset contains 500 image pairs from Google Earth, and each pair has hazy and haze-free images with the sizes of 512 × 512. From RICE-I, 400 pairs were randomly allocated for training, while the remaining 100 pairs were reserved for testing purposes. The WHUS2-CR dataset comprises cloudy and corresponding cloud-free images captured by the Sentinel-2A satellite. From the original high-resolution image pairs, we randomly cropped 5000 image patches sized at 256 × 256 pixels. For our experiments, 4000 pairs were allocated for training purposes, while the remaining 1000 pairs were reserved for testing. SateHze1k contains three levels of haze, that is, thin haze, moderate haze, and thick haze. Each of them consists of 320 pairs for training, 35 pairs for validation, and 45 pairs for testing, respectively. RS-Haze is a larger-scale synthetic dataset for remote-sensing image dehazing, in which the cloud is nonhomogeneous. It contains 51 300 image pairs for training and 2700 pairs for testing. Table I presents the summary of these five datasets in our experiments.

2) Evaluation Metrics: To quantitatively assess the effectiveness of our algorithm against other thin cloud removal methods, we employed three full-reference metrics: PSNR, SSIM [83], and the CIEDE2000 [84]. These metrics are chosen for their ability to provide objective measurements of performance. Larger PSNR and SSIM and smaller CIEDE2000 indicate better restoration performance.

3) Implementation Details: The proposed algorithm was implemented using the PyTorch framework. The computing platform consists of an Intel Gold 6252 CPU and an NVIDIA A100 GPU. For optimization of the network, the Adam optimizer [91] was utilized with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In the first stage, the proposed algorithm was trained for 300 epochs on the T-CLOUD dataset, 1000 epochs on the RICE-I and SateHze1k datasets, and 50 epochs on the RS-Haze dataset. A batch size of 16 was used. In the second stage, the training epochs were set to 50 on the T-CLOUD dataset, 300 on the RICE-I and SateHze1k datasets, and 15 on the RS-Haze dataset. During training, the image pairs were randomly cropped into 256 × 256 patches as input. The initial learning rate was set to 0.0001 and gradually reduced to 1×10^{-6} using the cosine annealing strategy.

TABLE II

QUANTITATIVE EVALUATIONS ON THE T-CLOUD AND RICE-I DATASETS, WHERE BOLD TEXTS AND UNDERLINED TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY. ↑: THE LARGER THE BETTER. ↓: THE SMALLER THE BETTER

Method	T-CLOUD			RICE-I			WHUS2-CR			Param (M)	FLOPs (G)
	PSNR↑	SSIM↑	CIEDE2000↓	PSNR↑	SSIM↑	CIEDE2000↓	PSNR↑	SSIM↑	CIEDE2000↓		
RSC-Net [20]	23.98	0.7596	7.0502	21.34	0.8150	8.3078	29.03	0.9056	4.6571	0.11	14.84
MCRN [85]	26.60	0.8091	5.5816	31.09	0.9465	3.3767	28.81	0.9163	4.7939	1.41	94.90
MSAR-DefogNet [21]	28.84	0.8432	4.1862	33.58	0.9534	2.7066	29.89	0.9168	5.2028	0.80	104.90
RCA-Net [86]	28.69	0.8443	4.3708	32.49	0.9537	2.2334	29.57	0.9128	4.4211	2.27	401.79
SPA-GAN [87]	27.15	0.8145	4.9107	29.62	0.8844	4.3374	28.78	0.8887	4.7904	0.21	33.97
Zheng *et al.* [43]	23.71	0.7630	7.6156	23.92	0.8085	7.6766	29.58	0.9008	5.1388	3.31	11.83
MS-GAN [88]	24.04	0.7228	7.8543	27.74	0.8796	5.6267	27.59	0.8560	6.2101	8.08	44.27
Color-GAN [89]	24.01	0.7490	6.9769	21.57	0.8065	8.5284	29.24	0.9020	4.7212	0.51	9.95
AMGAN-CR [19]	27.85	0.8317	4.5691	29.05	0.8965	4.4694	28.82	0.8672	4.9061	0.29	96.96
Cycle-SNSPGAN [90]	22.95	0.7714	8.1552	25.13	0.8229	10.4286	28.52	0.9122	6.0930	4.72	134.42
Ours	**31.52**	**0.8893**	**3.6275**	**36.29**	**0.9671**	**1.8316**	**30.23**	**0.9303**	**4.1672**	5.91	34.40

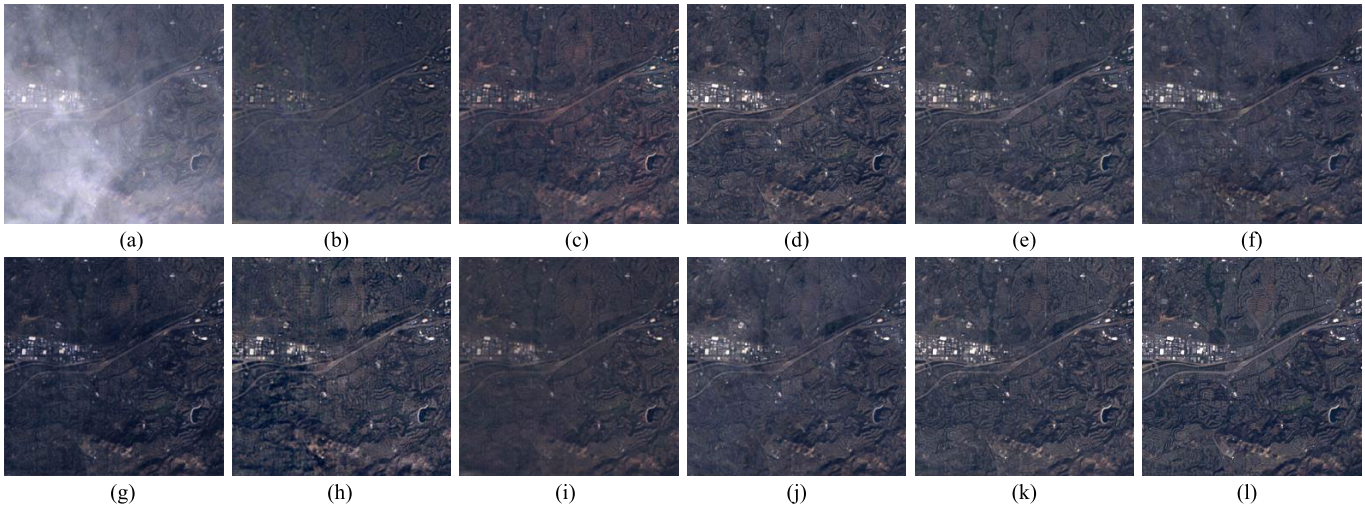


Fig. 6. Thin cloud removal results on the T-CLOUD dataset. Zoomed-in view for the best view. (a) Input. (b) RSC-Net. (c) MCRN. (d) MSAR-DefogNet. (e) RCA-Net. (f) SPA-GAN. (g) Zheng et al. (h) MS-GAN. (i) Color-GAN. (j) AMGAN-CR. (k) Ours. (l) Reference.

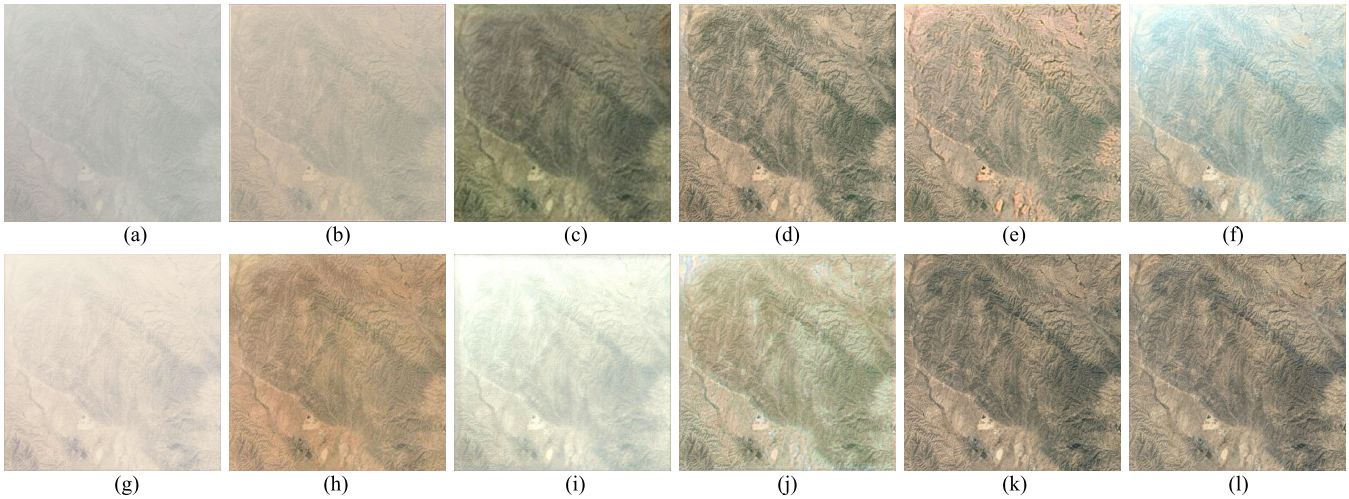


Fig. 7. Thin cloud removal results on the RICE-I dataset. Zoomed-in view for the best view. (a) Input. (b) RSC-Net. (c) MCRN. (d) MSAR-DefogNet. (e) RCA-Net. (f) SPA-GAN. (g) Zheng et al. (h) MS-GAN. (i) Color-GAN. (j) AMGAN-CR. (k) Ours. (l) Reference.

C. Thin Cloud Removal Results

To evaluate the performance of the proposed method, we compare it with several CNN-based thin cloud removal techniques, including multiscale distortion-aware networks (MCRNs) [85], spatial attention generative adversarial network (SPA-GAN) [87], RSC-Net [20], multiple scale attention residual network using for cloud remove (MSAR-DefogNet) [21], residual channel attention network (RCA-Net) [86], Zheng et al. [43], multiscale generative

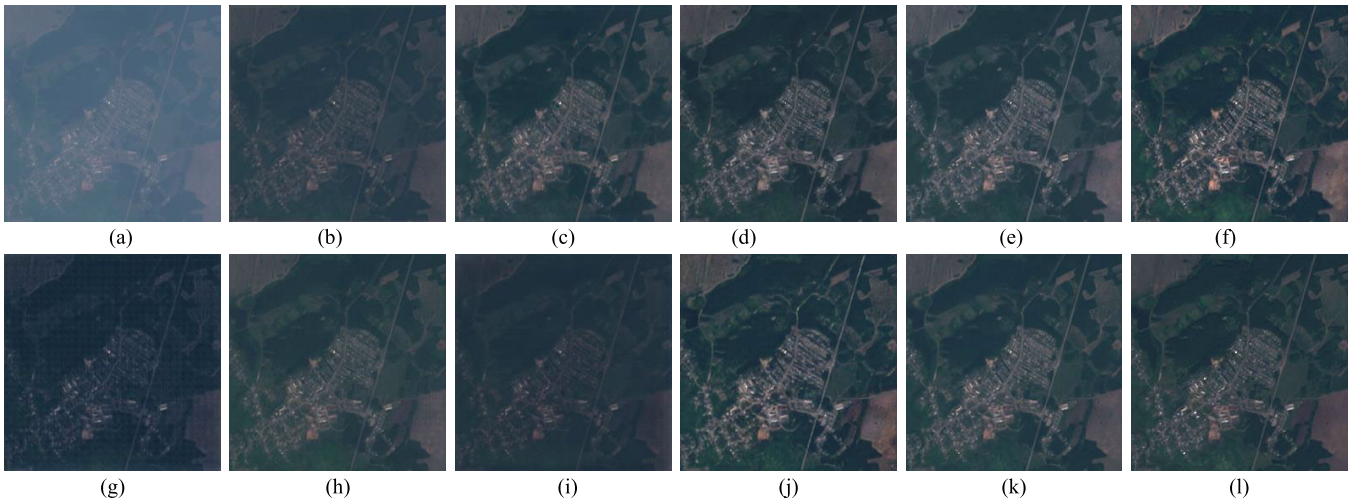


Fig. 8. Thin cloud removal results on the WHUS2-CR dataset. Zoomed-in view for the best view. (a) Input. (b) RSC-Net. (c) MCRN. (d) MSAR-DefogNet. (e) RCA-Net. (f) SPA-GAN. (g) Zheng et al. (h) MS-GAN. (i) Color-GAN. (j) AMGAN-CR. (k) Ours. (l) Reference.

adversarial net (MS-GAN) [88], Color-GAN [89], attention mechanism-based generative adversarial networks for cloud removal (AMGAN-CR) [19], and cycle spectral normalized soft likelihood estimation patch GAN (Cycle-SNSPGAN) [90].

The first three columns of Table II present a quantitative comparison between the proposed algorithm and existing restoration methods on the T-CLOUD dataset. Our algorithm outperforms all other methods in terms of all metrics. Specifically, compared to the previously top-performing model MSAR-DefogNet [21], our method achieves gains of 2.68 dB in PSNR and 0.0461 in SSIM. Visual comparisons of the evaluated models on T-CLOUD can be seen in Fig. 6. It is evident that some CNN-based models exhibit poor visual quality. For example, MCRN [85] and Color-GAN [89] fail to preserve much of the detailed information, while RSC-Net [20] and AMGAN-CR [19] struggle to remove the clouds. Additionally, Zheng et al. [43] displays noticeable grid artifacts and color distortion.

In addition, we performed a comparison of our proposed uncertainty framework with existing cloud removal algorithms using the widely used benchmark dataset RICE-I. The results, displayed in the middle three columns of Table II, demonstrate that our model outperforms previous methods across all metrics. Notably, our method achieves the lowest color difference score of 1.1836 on the CIEDE2000 metric, indicating that our results closely match the patterns of the reference image. Visualization results of the cloud removal process for each algorithm on the RICE-I dataset are presented in Fig. 7. It can be observed that certain algorithms, including RSC-Net [20], Zheng et al. [43], SPA-GAN [87], and Color-GAN [89], suffer from significant cloud residues. Likewise, MCRN [85] and AMGAN-CR [19] exhibit noticeable color distortion.

Table II and Fig. 8 present the comparison on the WHUS2-CR dataset. In comparison to the previous best method, MSAR-DefogNet, our algorithm showcased enhancements of 0.34 dB in PSNR and 0.0135 in SSIM. Notably, our method demonstrated the most superior performance in CIEDE2000, indicating significant potential for improving thin cloud removal techniques. The visual comparison in Fig. 8 highlights the enhanced quality of the clear images recovered by our method. Compared with other algorithms, the clear images recovered by our method have the most similar patterns to the reference image, exhibiting superior accuracy in detail and consistent color rendition. In summary, our proposed algorithm outperforms in thin cloud removal, image detail restoration, and preserving color fidelity, demonstrating its overall superiority.

TABLE III

QUANTITATIVE EVALUATIONS ON THE COASTAL BAND AND NEAR INFRARED BAND, WHERE BOLD TEXTS AND UNDERLINED TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY. ↑: THE LARGER THE BETTER. ↓: THE SMALLER THE BETTER

Method	Coastal Band		Near Infrared Band	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
RSC-Net [20]	25.14	0.7523	18.73	0.6572
MCRN [85]	27.29	0.7940	20.48	0.7138
MSAR-DefogNet [21]	<u>30.49</u>	<u>0.8550</u>	<u>22.73</u>	<u>0.7579</u>
RCA-Net [86]	30.27	0.8505	22.32	0.7554
SPA-GAN [87]	29.33	0.8477	21.84	0.7497
Zheng *et al.* [43]	27.67	0.8199	20.46	0.7242
MS-GAN [88]	24.81	0.7819	18.03	0.6619
Color-GAN [89]	28.06	0.8282	20.67	0.7398
AMGAN-CR [19]	25.97	0.7902	18.33	0.5979
Ours	**32.04**	**0.8798**	**23.08**	**0.7744**

In the last two columns of Table II, we conduct a comparison of the model's parameter count and floating-point operations (FLOPs). The results indicate that RSC-Net, SPA-GAN, and AMGAN-CR possess relatively fewer parameters and computational demands; however, their performance is comparatively inferior. Despite the relatively higher computational cost of our method, it delivers outstanding performance. A comprehensive analysis underscores that our algorithm strikes a superior balance between computational overhead and performance compared to other methods.

These single-solution algorithms only manage to recover relatively clear images, which affects their performance and robustness, particularly on challenging examples. In contrast,

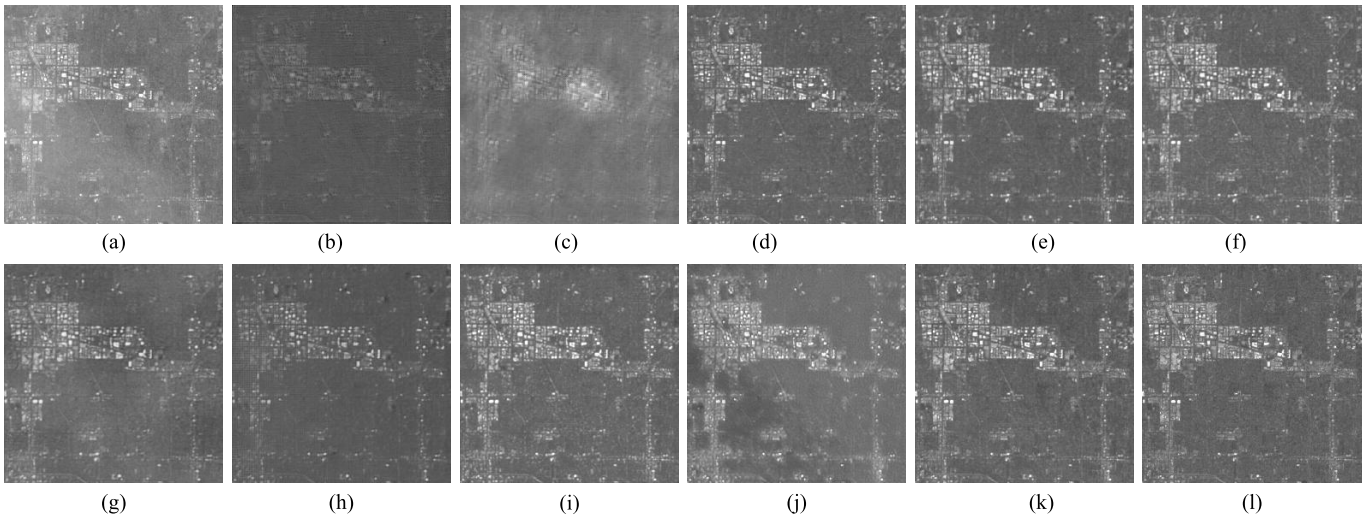


Fig. 9. Thin cloud removal results on the coastal band. Zoomed-in view for the best view. (a) Input. (b) RSC-Net. (c) MCRN. (d) MSAR-DefogNet. (e) RCA-Net. (f) SPA-GAN. (g) Zheng et al. (h) MS-GAN. (i) Color-GAN. (j) AMGAN-CR. (k) Ours. (l) Reference.

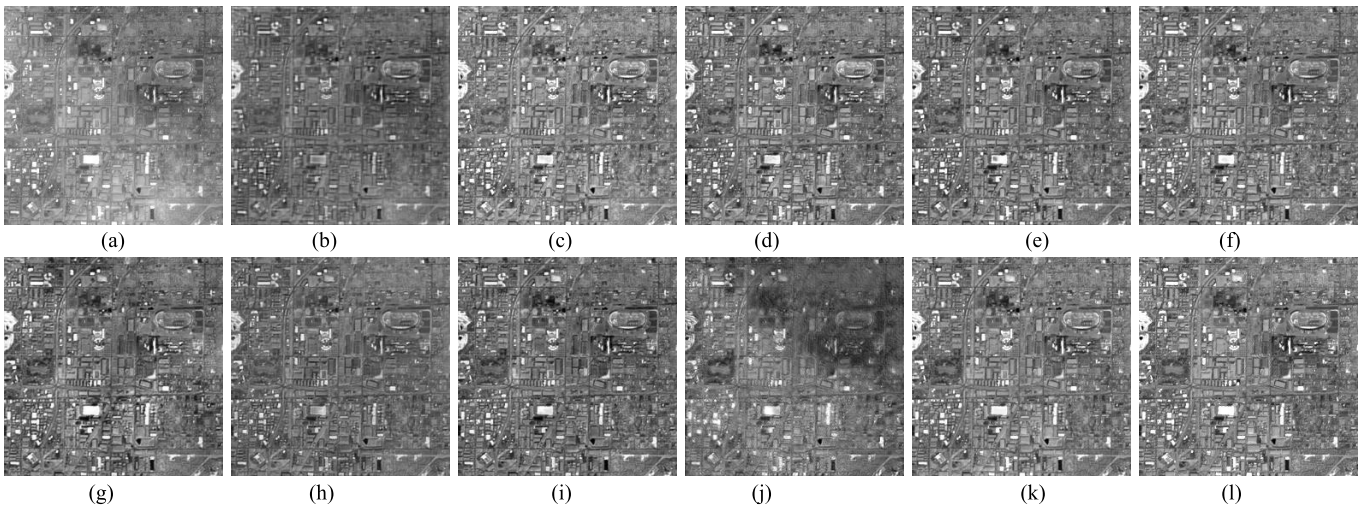


Fig. 10. Thin cloud removal results on the near infrared band. Zoomed-in view for the best view. (a) Input. (b) RSC-Net. (c) MCRN. (d) MSAR-DefogNet. (e) RCA-Net. (f) SPA-GAN. (g) Zheng et al. (h) MS-GAN. (i) Color-GAN. (j) AMGAN-CR. (k) Ours. (l) Reference.

our approach leverages multisolution fusion to enhance clarity and preserve intricate details. This highlights the potential of the multisolution fusion approach in enhancing the robustness of cloud removal algorithms.

D. Thin Cloud Removal on Other Bands

The multispectral configuration of satellite imagery is a fundamental feature different from common photographs. When we made T-CLOUD, we only used red band, green band, and blue band to synthesize RGB images. To verify the effectiveness of our algorithm on other bands, we selected the data of coastal band and near infrared band for experiments. Table III shows the quantitative comparison results. It suggests that our method still achieves the best results in these two bands. Figs. 9 and 10 show the visual comparison results. The results show that RSC-Net and MCRN have blur artifacts. SPA-GAN, UNet-GAN, and AMGAN-CR fail to remove the clouds completely. In contrast, our method achieves better visual results on both coastal band and near infrared band.

E. Dehazing Results

To further validate the effectiveness of our algorithm, we conducted experiments on two benchmark datasets for remote-sensing image dehazing: SateHaze1k and RS-Haze. The dehazing effect was quantitatively evaluated using the metrics of PSNR and SSIM. The comparison between our algorithm and existing methods on the SateHaze1k and RS-Haze test sets are presented in Table IV. The results clearly indicate that our algorithm achieves the highest scores in terms of PSNR and SSIM for both datasets. This suggests that the clear images generated by our method exhibit more similar patterns to the reference clear images, having richer texture details.

The visual comparisons on these two datasets are depicted in Figs. 11 and 12. All-in-one dehazing network (AOD-Net) [28] exhibits prominent residual haze and significant color distortion, which can be attributed to its simplistic network architecture. While GridDehazeNet [95] and multi-scale boosted dehazing network (MSBDN) [96] man-

TABLE IV

QUANTITATIVE EVALUATIONS ON THE SATEHAZE1K AND RS-HAZE DATASET, WHERE BOLD TEXTS AND UNDERLINED TEXTS INDICATE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY. ↑: THE LARGER THE BETTER. ↓: THE SMALLER THE BETTER

Method	SateHaze1k-Thin		SateHaze1k-Moderate		SateHaze1k-Thick		SateHaze1k-Average		RS-Haze	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DCP [16]	13.15	0.7246	9.78	0.5735	10.25	0.5850	11.06	0.6477	17.86	0.7340
AOD-Net [28]	19.54	0.8543	20.10	0.8854	15.92	0.7313	18.52	0.8234	27.09	0.8476
FCFT-Net [92]	23.59	0.9127	22.88	0.9272	20.03	0.8156	22.17	0.8852	33.28	0.9417
H2RL-Net [93]	20.91	0.8797	22.34	0.9061	17.41	0.7684	20.22	0.8514	31.18	0.9212
M2SCN [94]	25.21	0.9175	26.11	0.9416	21.33	0.8289	24.22	0.8960	37.75	0.9497
GridDehazeNet [95]	24.67	0.9075	25.62	0.9367	20.80	0.8414	23.71	0.8952	36.40	0.9600
MSBDN [96]	25.63	0.9195	26.62	0.9450	20.59	0.8350	24.28	0.8998	38.57	0.9650
Trinity-Net [97]	21.55	0.8842	23.35	0.8952	20.97	0.8226	21.96	0.8673	32.17	0.9186
Uformer [98]	22.82	0.9070	24.47	0.9393	20.36	0.8148	22.55	0.8864	38.89	0.9573
Restormer [99]	23.08	0.9116	24.73	0.9334	18.58	0.7616	22.13	0.8689	39.24	0.9576
UMWTransformer [100]	24.29	0.9190	26.65	0.9455	20.07	0.8252	23.67	0.8966	36.11	0.9464
FocalNet [101]	24.16	0.9162	25.99	0.9469	21.69	0.8474	23.95	0.9035	38.39	0.9539
C^2PNet [102]	19.62	0.8802	24.79	0.9399	16.83	0.7895	20.41	0.8699	34.78	0.9419
Ours	**27.36**	**0.9293**	**28.37**	**0.9566**	**22.83**	**0.8518**	**26.19**	**0.9126**	**39.52**	**0.9706**

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

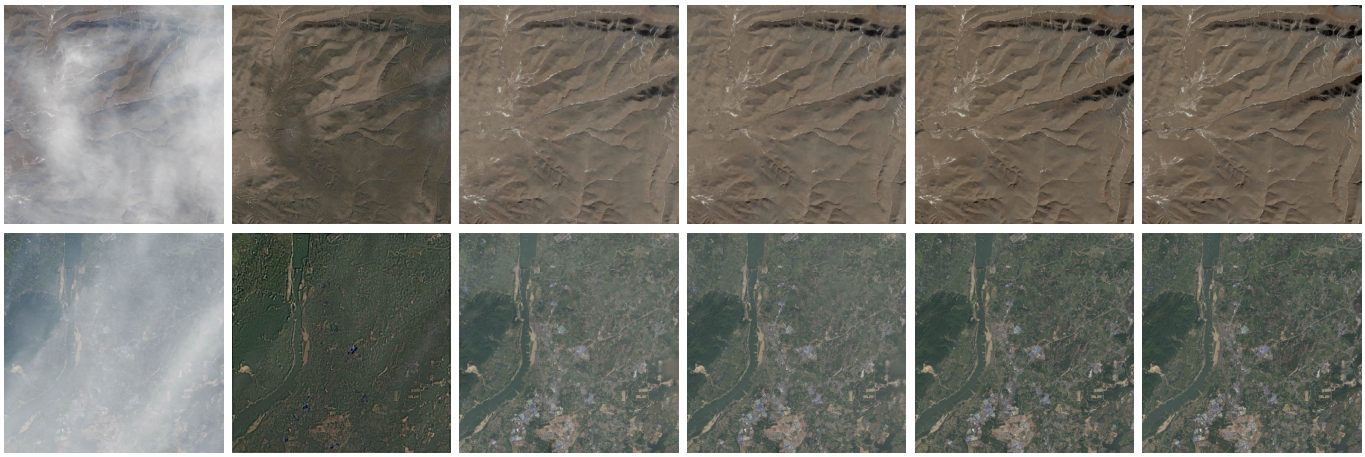
Fig. 11. Visual comparisons on the SateHaze1k dataset. The first row, the second row, and the third row show the visualization results in the SateHaze1k-Thin, SateHaze1k-Moderate, and SateHaze1k-Thick test sets, respectively. Zoomed-in view for the best view. (a) Input. (b) AOD-Net. (c) GridDehazeNet. (d) MSBDN. (e) Ours. (f) Reference.

age to produce competitive outputs, some traces of haze residue still persist. These supervised algorithms employ diverse network structures to obtain a single solution that closely aligns with the reference image. However, this approach compromises their robustness and constrains their effectiveness in restoring degraded images in challenging scenarios. In contrast, with a multiple solutions fusion strategy, our algorithm significantly improves the performance of image restoration, leading to visually pleasing and realistic outcomes.

F. Ablation Study

Here, we present ablation experiments to demonstrate the effectiveness of the proposed uncertainty-based algorithms. All evaluation is performed on the proposed T-CLOUD dataset. To streamline the experiments and align with prior research [40], we reduce the number of training rounds in the ablation experiment stage by half compared to the training stage.

1) Scale of Latent Variable z: First, we investigate the impact of the dimensionality of the latent variable z. We set z sizes to 3, 6, 9, and 12, respectively, and compare the quantitative indicators of these variant models on the test set. The quantitative results are shown in Table V. Different dimensions of z exhibit varying defogging performances. The performance of z is relatively poor at low dimensions, but it also diminishes if the dimension is too high. Based on our



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

Fig. 12. Visual comparisons on the RS-Haze dataset. Zoomed-in view for the best view. (a) Input. (b) AOD-Net. (c) GridDehazeNet. (d) MSBDN. (e) Ours. (f) Reference.

TABLE V
ABLATION STUDY RESULTS ON T-CLOUD DATASET FOR DIFFERENT DIMENSION OF z IN THE UNCERTAINTY FRAMEWORK

z	3	6	9	12
PSNR	28.72	29.20	**29.63**	29.07
SSIM	0.8512	0.8538	**0.8579**	0.8532

TABLE VI
ABLATION STUDY RESULTS ON T-CLOUD DATASET FOR DIFFERENT FUSION STRATEGY

Method	Average Restoration	SFN	DFN
PSNR	28.12	28.96	**29.63**
SSIM	0.8501	0.8526	**0.8579**

TABLE VII
RESULTS OF EACH LOSS ITEM

$\mathcal{L}_1$	$\mathcal{L}_{mse}$	$\mathcal{L}_{fre}$	PSNR	MSSIM
✓			29.37	0.8551
	✓		28.73	0.8516
✓		✓	**29.63**	**0.8579**

TABLE VIII
ABLATION STUDY RESULTS ON T-CLOUD DATASET FOR DIFFERENT AGGREGATION STRATEGY

Method	Concatenation	Summation	SFFM
PSNR	27.76	27.94	**29.63**
SSIM	0.8303	0.8298	**0.8579**

experimental findings, we set the dimension of z to 9 in our uncertainty framework.

2) Effect of DFN: Our proposed uncertainty framework has the advantage of generating multiple plausible solutions. Therefore, we design DFN to merge these solutions and achieve a more accurate overall solution. To assess the effectiveness of this fusion scheme, we introduce two additional fusion strategies. The first strategy, termed Average Restoration, computes the expectation of multiple reasonable solutions as the final fusion result [i.e., $Y^* = (1/n)\sum_{i=1}^{n} Y_i$]. The second strategy is a static fusion network (SFN), which employs convolution and activation layers to construct a fusion network. After training, all inputs are processed using the same parameters.

The quantitative comparison results of these three fusion strategies are presented in Table VI. Although the method of directly obtaining expectations is simple and efficient, it performs poorly in all quantitative indicators. The other two fusion strategies, which employ learnable parameters, exhibit significantly superior performance. In contrast, our proposed dynamic fusion strategy attains the highest PSNR and SSIM scores, highlighting its superiority in handling diverse samples. This dynamic fusion scheme can generate varying parameters for individual samples and align them for personalized processing. The Dynamic fusion strategy offers higher flexibility and greater robustness compared to static fusion.

3) Evaluation on Loss Function: Furthermore, we investigate the influence of the objective function on the network's final recovery performance during training. As a point of comparison, we select the mse loss function, and the corresponding experimental results are displayed in Table VII. L1 loss yields superior PSNR and SSIM scores, potentially because minimizing mse suppresses high-frequency details, leading to image blurring and excessive smoothing. Therefore, we adopt L1 loss as the primary reconstruction loss term. Additionally, Table VI demonstrates that utilizing the frequency domain loss as an auxiliary term resulted in PSNR and SSIM improvements of 0.26 and 0.0028, respectively. Based on these experiments, we ultimately select the combination of L1 loss and frequency domain loss as the objective function for optimizing the network parameters.

4) Effect of SFFM: To demonstrate the effectiveness of this feature aggregation strategy, we compare it with concatenation and summation. Table VIII shows that SFFM achieves a PSNR gain of 1.87 and 1.69 dB over concatenation and summation, respectively. This demonstrates that the selective mechanism

TABLE IX
ABLATION STUDY RESULTS ON T-CLOUD DATASET FOR THE ATTENTION MECHANISM AND THE MIMO STRATEGY

Method	w/o Attentino	w/o MIMO	Ours
PSNR	25.03	27.51	**29.63**
SSIM	0.7750	0.8128	**0.8579**

utilized in SFFM can integrate more informative features, resulting in improved performance.

5) Effect of Attention and MIMO Mechanism: The influence of the attention mechanism is evaluated in Table IX. The results suggest that the attention mechanism significantly improves performance. After introducing the channel attention and spatial attention mechanisms, the model achieved a gain of 4.60-dB PSNR and 0.0829 SSIM. The attention mechanism makes the network invest more learning effort in valuable patterns, achieving better performance. To demonstrate the effectiveness of the MIMO mechanism, we built a single-input single-output (SISO) variant. The MIMO mechanism eases the difficulty of training and leads to an additional 2.12-dB PSNR and 0.0451 SSIM gains.

V. CONCLUSION

In this article, we propose an algorithm for removing haze and thin clouds in remote-sensing images based on a probabilistic approach. Unlike previous deterministic restoration algorithms, our method is capable of recovering multiple clear images from a single degraded input. Our algorithm is built upon the CVAE framework and employs an MIMO U-shaped architecture for the restoration network. The restoration process involves feeding the sampled latent variable and degraded image into the network to obtain a reasonable solution. To enhance the performance, we introduce an SFFM that merges the intermediate features from the encoder and decoder. Furthermore, we design a dynamic fusion network utilizing dynamic convolution to combine multiple reasonable solutions and generate a more accurate restoration result. Through our experiments on multiple datasets, our algorithm has achieved state-of-the-art results in remote-sensing image dehazing and cloud removal tasks, demonstrating the substantial potential of incorporating uncertainties to enhance the quality of remote-sensing image restoration.

Our proposed CVAE-based uncertainty restoration framework holds the potential for extension to various image restoration tasks, including image super-resolution reconstruction, denoising, and rain removal. However, an inherent drawback lies in the increased computational load due to multiple samplings. Enhancing the model's efficiency stands as a focal point in our forthcoming research endeavors.

REFERENCES

[1] B. Li et al., "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[2] Y. Li, Q. Xu, Z. He, and W. Li, "Progressive task-based universal network for raw infrared remote sensing imagery ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610013.

[3] Q. Xu, Y. Li, M. Zhang, and W. Li, "COCO-Net: A dual-supervised network with unified ROI-loss for low-resolution ship detection from optical satellite image sequences," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3201530.

[4] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 113–140, Jul. 2005.

[5] J.-P. Tarel, N. Hautière, A. Cord, D. Gruyer, and H. Halmaoui, "Improved visibility of road scene images under heterogeneous fog," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2010, pp. 478–485.

[6] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, Sep. 2018.

[7] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sens.*, vol. 10, no. 11, p. 1700, Oct. 2018.

[8] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[9] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.

[10] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," 2023, *arXiv:2310.19288*.

[11] K. Li, M. Feng, A. Biswas, H. Su, Y. Niu, and J. Cao, "Driving factors and future prediction of land use and cover change based on satellite remote sensing data by the LCM model: A case study from Gansu Province, China," *Sensors*, vol. 20, no. 10, p. 2757, May 2020.

[12] Y. Chen, J. Huang, X. Song, H. Wen, and H. Song, "Evaluation of the impacts of rain gauge density and distribution on gauge-satellite merged precipitation estimates," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3037099.

[13] X. Hong et al., "Retrieval of global carbon dioxide from TanSat satellite and comprehensive validation with TCCON measurements and satellite observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3066623.

[14] R. Richter, "A spatially adaptive fast atmospheric correction algorithm," *Int. J. Remote Sens.*, vol. 17, no. 6, pp. 1201–1214, Apr. 1996.

[15] E. F. Vermote, D. Tanre, J. L. Deuze, M. Herman, and J.-J. Morcette, "Second simulation of the satellite signal in the solar spectrum, 6S: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 35, no. 3, pp. 675–686, May 1997.

[16] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[17] J. Li, Q. Hu, and M. Ai, "Haze and thin cloud removal via sphere model improved dark channel prior," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 472–476, Mar. 2019.

[18] Q. Guo, H.-M. Hu, and B. Li, "Haze and thin cloud removal using elliptical boundary prior for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9124–9137, Nov. 2019.

[19] 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.

[20] W. Li, Y. Li, D. Chen, and J. C.-W. Chan, "Thin cloud removal with residual symmetrical concatenation network," *ISPRS J. Photogramm. Remote Sens.*, vol. 153, pp. 137–150, Jul. 2019.

[21] Y. Zhou, W. Jing, J. Wang, G. Chen, R. Scherer, and R. Damaševičius, "MSAR-DefogNet: Lightweight cloud removal network for high resolution remote sensing images based on multi scale convolution," *IET Image Process.*, vol. 16, no. 3, pp. 659–668, Feb. 2022.

[22] L. Zhang and S. Wang, "Dense haze removal based on dynamic collaborative inference learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3207832.

[23] J. Guo, J. Yang, H. Yue, H. Tan, C. Hou, and K. Li, "RSDehazeNet: Dehazing network with channel refinement for multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2535–2549, Mar. 2021.

[24] J. Nie, W. Wei, L. Zhang, J. Yuan, Z. Wang, and H. Li, "Contrastive haze-aware learning for dynamic remote sensing image dehazing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3220940.

[25] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," 2022, *arXiv:2204.03883*.

[26] Y. Zi, F. Xie, N. Zhang, Z. Jiang, W. Zhu, and H. Zhang, "Thin cloud removal for multispectral remote sensing images using convolutional neural networks combined with an imaging model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3811–3823, 2021.

[27] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[28] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4780–4788.

[29] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.

[30] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1674–1682.

[31] J. Zhang and D. Tao, "FAMED-Net: A fast and accurate multi-scale end-to-end dehazing network," *IEEE Trans. Image Process.*, vol. 29, pp. 72–84, 2020.

[32] 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.

[33] P. S. Chavez Jr., "An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data," *Remote Sens. Environ.*, vol. 24, no. 3, pp. 459–479, Apr. 1988.

[34] R. Fattal, "Single image dehazing," *ACM Trans. Graph*, vol. 27, no. 3, pp. 1–9, Aug. 2008.

[35] 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.

[36] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2810–2818.

[37] 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.*, Jul. 2018, pp. 1772–1775.

[38] M. Qin, F. Xie, W. Li, Z. Shi, and H. Zhang, "Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1645–1655, May 2018.

[39] H. Ding, F. Xie, Y. Zi, W. Liao, and X. Song, "FeedBack network for compact thin cloud removal," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[40] Y. Zi, H. Ding, F. Xie, Z. Jiang, and X. Song, "Wavelet integrated convolutional neural network for thin cloud removal in remote sensing images," *Remote Sens.*, vol. 15, no. 3, p. 781, Jan. 2023.

[41] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 154–169.

[42] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3194–3203.

[43] J. Zheng, X.-Y. Liu, and X. Wang, "Single image cloud removal using U-Net and generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6371–6385, Aug. 2021.

[44] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[45] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1278–1286.

[46] P. Esser and E. Sutter, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8857–8866.

[47] L. Zhang and Y. Liu, "Remote sensing image generation based on attention mechanism and VAE-MSGAN for ROI extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[48] S. Kohl et al., "A probabilistic U-Net for segmentation of ambiguous images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6965–6975.

[49] C. F. Baumgartner et al., "PHiSeg: Capturing uncertainty in medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent*. Cham, Switzerland: Springer, 2019, pp. 119–127.

[50] J. Zhang et al., "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8579–8588.

[51] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

[52] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 667–675.

[53] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[54] L. Yang, Y. Han, X. Chen, S. Song, J. Dai, and G. Huang, "Resolution adaptive networks for efficient inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2366–2375.

[55] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: A saliency-based sampling layer for neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 51–66.

[56] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4476–4484.

[57] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.

[58] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5219–5227.

[59] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 282–298.

[60] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17721–17732.

[61] R. Guo, D. Niu, L. Qu, and Z. Li, "SOTR: Segmenting objects with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7137–7146.

[62] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3561–3571.

[63] F. Wu, F. Chen, X.-Y. Jing, C.-H. Hu, Q. Ge, and Y. Ji, "Dynamic attention network for semantic segmentation," *Neurocomputing*, vol. 384, pp. 182–191, Apr. 2020.

[64] Z. Zhong et al., "Squeeze-and-attention networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13062–13071.

[65] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308.

[66] T. Verelst and T. Tuytelaars, "Dynamic convolutions: Exploiting spatial sparsity for faster inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2317–2326.

[67] Z. Xie, Z. Zhang, X. Zhu, G. Huang, and S. Lin, "Spatially adaptive inference with stochastic feature sampling and interpolation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 531–548.

[68] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1913–1922.

[69] A. Bhowmik, S. Shit, and C. S. Seelamantula, "Training-free, single-image super-resolution using a dynamic convolutional network," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 85–89, Jan. 2018.

[70] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-SR: A magnification-arbitrary network for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1575–1584.

[71] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE Trans. Image Process.*, vol. 29, pp. 4027–4040, 2020.

[72] M. Chang, Q. Li, H. Feng, and Z. Xu, "Spatial-adaptive network for single image denoising," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, Aug. 2020, pp. 171–187.

[73] 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. 4621–4630.

[74] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[76] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 234–241.

[77] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[78] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13899–13909.

[79] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5759–5770.

[80] 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*.

[81] 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, 2022.

[82] B. Huang, Z. Li, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1795–1802.

[83] 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.

[84] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, Feb. 2005.

[85] W. Yu, X. Zhang, M.-O. Pun, and M. Liu, "A hybrid model-based and data-driven approach for cloud removal in satellite imagery using multi-scale distortion-aware networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 7160–7163.

[86] 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, pp. 1–5, 2022.

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

[88] Z. Xu, K. Wu, L. Huang, Q. Wang, and P. Ren, "Cloudy image arithmetic: A cloudy scene synthesis paradigm with an application to deep-learning-based thin cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3122253.

[89] C. Zhang, X. Zhang, Q. Yu, and C. Ma, "An improved method for removal of thin clouds in remote sensing images by generative adversarial network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 6706–6709.

[90] Y. Wang et al., "Cycle-SNSPGAN: Towards real-world image dehazing via cycle spectral normalized soft likelihood estimation patch GAN," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20368–20382, Nov. 2022.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[92] Y. Li and X. Chen, "A coarse-to-fine two-stage attentive network for haze removal of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1751–1755, Oct. 2021.

[93] X. Chen, Y. Li, L. Dai, and C. Kong, "Hybrid high-resolution learning for single remote sensing satellite image dehazing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[94] S. Li, Y. Zhou, and W. Xiang, "M2SCN: Multi-model self-correcting network for satellite remote sensing single-image dehazing," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[95] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7313–7322.

[96] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2154–2164.

[97] K. Chi, Y. Yuan, and Q. Wang, "Trinity-Net: Gradient-guided Swin transformer-based remote sensing image dehazing and beyond," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702914.

[98] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," 2021, *arXiv:2106.03106*.

[99] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[100] A. Kulkarni, S. S. Phutke, and S. Murala, "Unified transformer network for multi-weather image restoration," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 344–360.

[101] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Focal network for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 13001–13011.

[102] Y. Zheng, J. Zhan, S. He, J. Dong, and Y. Du, "Curricular contrastive regularization for physics-aware single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5785–5794.

Haidong Ding received the B.E. degree from the Department of Aerospace Information Engineering, School of Astronautics, Beihang University, Beijing, China, in 2021, where he is currently pursuing the M.S. degree.

His research interests include image processing and deep learning.

Fengying Xie (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent system from Beihang University, Beijing, China, in 2009.

From 2010 to 2011, she was a Visiting Scholar with the Laboratory for Image and Video Engineering, The University of Texas at Austin, Austin, TX, USA. She is currently a Professor with the Department of Aerospace Information Engineering, School of Astronautics, Beihang University. Her research interests include biomedical image processing, remote-sensing image understanding and applications, image quality assessment, and object recognition.

Linwei Qiu received the B.S. degree in detection, guidance, and control techniques from Beihang University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the School of Astronautics.

His research interests include low-level vision tasks and medical image analysis.

Xiaozhe Zhang (Graduate Student Member, IEEE) received the B.E. degree in electronic and information engineering from Southwest Jiaotong University, Chengdu, China, in 2023. He is currently pursuing the master's degree with the Image Processing Center, School of Astronautics, Beihang University, Beijing, China.

His research interests include deep learning and computer vision.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or coauthored over 200 scientific articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and the IEEE International Conference on Computer Vision (ICCV). His research interests include remote-sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and *Infrared Physics and Technology*. His personal website is http://levir.buaa.edu.cn/.