Uncertainty-Based Thin Cloud Removal Network via Conditional Variational Autoencoders

Haidong Ding, Yue Zi, and Fengying Xie(✉)

School of Astronautics, Beihang University, Beijing 100191, China
{dinghaidong, ziyue91, xfy_73}@buaa.edu.cn

Abstract. Existing thin cloud removal methods treat this image restoration task as a point estimation problem, and produce a single cloud-free image following a deterministic pipeline. In this paper, we propose a novel thin cloud removal network via Conditional Variational Autoencoders (CVAE) to generate multiple reasonable cloud-free images for each input cloud image. We analyze the image degradation process with a probabilistic graphical model and design the network in an encoder-decoder fashion. Since the diversity in sampling from the latent space, the proposed method can avoid the shortcoming caused by the inaccuracy of a single estimation. With the uncertainty analysis, we can generate a more accurate clear image based on these multiple predictions. Furthermore, we create a new benchmark dataset with cloud and clear image pairs from real-world scenes, overcoming the problem of poor generalization performance caused by training on synthetic datasets. Quantitative and qualitative experiments show that the proposed method significantly outperforms state-of-the-art methods on real-world cloud images. The source code and dataset are available at https://github.com/haidong-Ding/Cloud-Removal.

1 Introduction

Remote sensing images often suffer from absorption and scattering effects caused by thin clouds, resulting in degraded images. These low-quality images limit their utilization on subsequent high-level computer vision tasks, *e.g.*, object detection [1,2,3] and segmentation [4,5,6]. Therefore, it is significant to develop an effective method for single remote sensing image de-clouding.

Existing methods can generally be divided into two categories: prior-based approaches and data-driven approaches. Prior-based cloud removal models [7,8,9] are mainly built upon the atmospheric scattering model with various physical assumptions imposed on image statistics. These prior-based methods are more explanatory but can not perform well when the statistical prior does not hold in real-world images.

To alleviate these limitations, data-driven methods adopt the deep learning approach to train the network in supervised learning paradigm. Several methods [10,11,12] directly learn how to generate clear images from their cloud counterparts in an end-to-end manner. Such algorithms are based on learning from large amounts of data and can produce decent results. However, the end-to-end training fashion usually regards de-clouding as a black box problem, making it poorly interpretable. To avoid this issue, other methods [13,14,15] combine the imaging model with Convolutional Neural Networks (CNNs). They mainly focus on building a neural network to replace part of the physical model in the conventional methods.

Although many excellent works have shown outstanding results, there are still many difficulties and misconceptions about the thin cloud removal task. Therefore, it is necessary to examine this problem in a broader context, two of which are highlighted below.

1) *The role of synthetic datasets.* It is hard to obtain image pairs with and without thin clouds in real-world scenes, so most algorithms [13,16,17] are trained on synthetic datasets. The differences between the synthetic and the real-world images make the network learn the law of data synthesis rather than the essence of image degradation during the optimization process. In addition, the synthetic dataset is based on the physical model of image degradation, so it is worth considering whether to rely on this physical model when designing the algorithm. As a result, this type of method can show excellent performance on synthetic data while performing poorly in real-world scenes. It demonstrates that the use of synthetic datasets inhibits the generalization performance of the model to a certain extent.

2) *The diversity of solutions.* A single cloud image loses some essential information of scene radiance, and recovering a completely clear image from it is equivalent to using little information to recover the whole scene structure, which makes this problem highly ill-posed. Therefore, this low-level computer vision problem is inherently uncertain. The exact value of each pixel in the clear image cannot be obtained by using only the degraded image without other auxiliary information. To our best knowledge, none of the existing methods take this into consideration. All of them established a one-to-one mapping from cloud image to cloud-free image. Therefore, considering uncertainty has great potential to improve the performance of cloud removal algorithms.

To address the aforementioned challenges, we propose a probabilistic model via CVAE for remote sensing image thin cloud removal. Based on the above analysis of uncertainty, we tackle this problem from the perspective of multi-solution. Each time the output from the proposed method is a sample of possible solutions. Our method is not based on an explicit prior model to avoid the insufficient understanding of the image degradation mechanism. Meanwhile, we do not use synthetic datasets so that the underlying de-clouding principles can be learned directly from real-world cloud images. The diverse output can ultimately enhance the generalization ability of the proposed thin cloud removal network.

Our contributions can be summarized as follows:

- Aiming at the uncertainty of thin cloud removal, we propose a probabilistic model based on CVAE for remote sensing image de-clouding, which solves this multi-solution problem from a probabilistic perspective. The network outputs multiple interpretable results, which fits the property of indefinite solutions to the problem.

- We propose an encoder network based on Vision Transformer (ViT) and a multi-scale feature fusion decoder network to achieve a one-to-many mapping from cloud image to clear image.
- We create a new benchmark dataset for single image thin cloud removal. The cloud and clear image pairs are from different moments of the same real scene to overcome the low generalization performance of the model due to training on synthetic datasets.

2 Related Work

2.1 Prior-Based Methods

Most conventional methods are based on the physical prior, estimate some important quantities (*e.g.*, the transmission map) in the model, and then recover a clear image from its cloud counterpart. Chavez [18] proposed an additive model to describe the generation principle of cloud images under the assumption that the distance between the sensor and the ground is fixed. He *et al.* [9] 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 non-sky area of the image, which is used to estimate the transmission map. Fattal *et al.* [19] proposed a color-lines prior to estimate the transmission map based on the distribution of images in the RGB color space. Berman *et al.* [20] 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 non-local prior. Xu *et al.* [21] proposed a method based on signal transmission and airspace hybrid analysis, combined with atmospheric scattering theory to remove clouds.

These prior knowledge-based methods show superior statistical properties in specific scenarios but fail easily in real-world images where physical assumptions do not hold.

2.2 Data-Driven Methods

In recent years, thanks to the establishment of large-scale datasets and the development of deep learning techniques, many data-driven supervised cloud removal methods have been proposed to overcome the shortcomings of traditional methods. Mao *et al.* [22] proposed a deep encoder-decoder framework and added skip connections to improve the efficiency of image restoration. Praveer Singh *et al.* [23] proposed an adversarial training-based network named Cloud-GAN to directly learn the mapping relationship between cloud and clear images. Qin *et al.* [16] proposed a multi-scale deblurring convolutional neural network with the residual structure to remove the thin cloud. Li *et al.* [11] designed an end-to-end residual symmetric connection network for thin cloud removal. Xu *et al.* [10] introduced a generative adversarial network based on the attention mechanism to guide the network to invest more efforts in denser cloud regions.

Inspired by traditional methods, some researchers combine deep learning technology with imaging physical models. Cai *et al.* [14] showed that medium transmission estimation can be reformulated as a learnable end-to-end system. Ren *et al.* [24] used deep learning to learn the transmission map and solve atmospheric scattering model. Zheng *et al.* [17] combined the existing atmospheric scattering model with UNet to remove thin clouds. According to the additive model of cloud images, Zi *et al.* [13] utilized deep neural networks combined with the imaging model to achieve thin cloud removal.

Although these data-driven methods have made immense progress in thin cloud removal performance, all of them achieve a one-to-one mapping with respect to the input. They only found a reasonable one out of all the solutions to this multi-solution problem. Unlike existing methods, we treat single image cloud removal as an indeterminate solution problem. We combine cloud removal with uncertainty analysis to better solve this ill-posed problem.

3 Uncertainty Cloud Removal Framework

3.1 Analysis of Image Degradation Process

Several researchers [18,25,26] have proposed that the satellite image degradation process can be described as an additive model (1):

$$S = G + C \ . \tag{1}$$

or a non-linear model (2):

$$S = (1 - C) \cdot G + C \ . \tag{2}$$

where S, G, and C represent cloud images acquired by satellite sensors, clear ground scene images, and thin cloud thickness maps, respectively.

However, designing a network based on the explicit model raises two questions: whether the model fully conforms to the degradation process and whether the unknown variables can be estimated accurately. Both of these factors can affect the outcome of recovery and even lead to failure.

To tackle these problems, firstly, we do not rely on the atmosphere scattering model of satellite images. We note that these model-based algorithms have a common idea: estimate the unknown variables from the observed images and then combine the imaging model to restore clear images. In our method, we redefine the unobservable variables as degeneration factors, which results in low-quality images. The image degradation system is determined by the cloud image, clear image, and degradation factors. Instead of giving the relationship between these three variables, we leverage the powerful representation ability of the neural network to express implicitly.

We analyze image degradation employing a probabilistic graphical model, as shown in Figure 1. The observable variable X is the cloud image obtained by

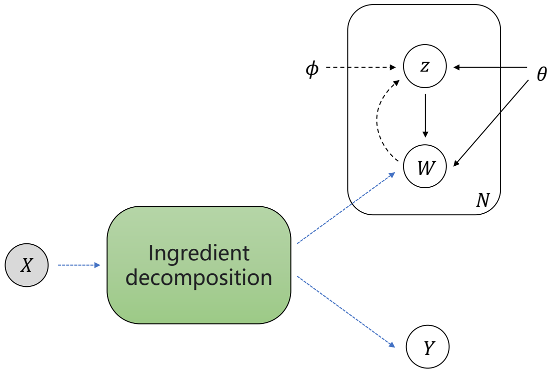


Fig. 1. The probabilistic graphical model for the degradation process of cloud images. The solid black line represents the generative model, and the dashed black line represents the variational approximation. θ and ϕ are the parameters of the model.

the satellite sensor, which can be decomposed into degeneration factors W and ground scene information Y. The ground scene Y is more complex and changeable, which results in describing its generation mechanism with a probability graph model is a challenging task. So it is more feasible to infer degeneration factors W from cloud images.

Since the essential intrinsic relationship between these three variables is unknown, variable W cannot be obtained explicitly. To get this crucial variable, we assume that the latent variable z determines its generation. Because the cloud image contains the information of degeneration factors, we estimate z through X and then generate W from a sampling of z.

Due to the diversity of the sampling process, multiple estimations can be generated. This diversity mitigates the problems caused by inaccurate estimations and makes the algorithm more robust.

The overall model framework is based on the CVAE, the inference network estimates the variational distribution, and the generative network achieves final cloud-free results.

The variational lower bound of the model is as follows (detailed derivation is available in supplementary material):

$$\log p_\theta(Y|X) = -D_{KL}(q_\phi(z|X)||p_\theta(z|X)) + \mathbb{E}_{q_\phi(z|X)}[\log p_\theta(Y|X,z)] \ . \tag{3}$$

where the proposal distribution $q_\phi(z|X)$ is introduced to approximate the posterior $p_\theta(z|X)$; the latent variable z drawn from $q_\phi(z|X)$. θ and ϕ represent the parameter set of distribution. Our CVAE framework is composed of a encoder network $q_\phi(z|X)$ and a generative encoder network $p_\theta(Y|X,z)$. The Kullback-Leibler (KL) Divergence $D_{KL}(q_\phi(z|X)||p_\theta(z|X))$ work as a regularization loss to narrow the gap between the posterior $p_\theta(z|X)$ and the proposal distribution $q_\phi(z|X)$. To simplify the network and reduce computation, we assume that the posterior follows the standard normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$.

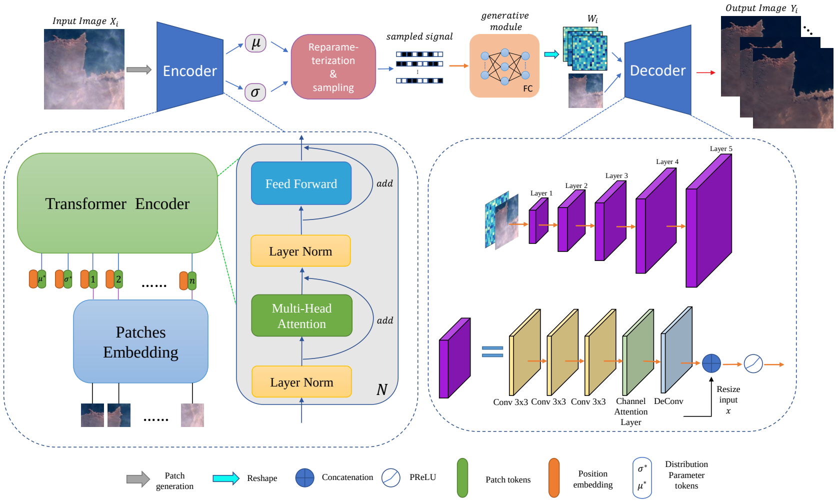


Fig. 2. An overview of the proposed cloud removal network. (1) The encoder network based on ViT learns the distribution of degradation factors W from cloud image X. (2) The decoder samples in the latent space with the reparameterization trick. Then the generative module utilizes the sampled signal to yield the image manifold of W. (3) Then, the combination of X and W is fed into the decoder network, which consists of five convolutional blocks of different scales. Each block contains convolution operation, channel attention, and deconvolution operation.

3.2 Transformer-Based Encoder

The encoder works as the recognition model to infer the proposal distribution $q_\phi(z|X)$. To allow the network can be trained using the gradient descent algorithm, z is drawn with the reparameterization trick, which is written as (4):

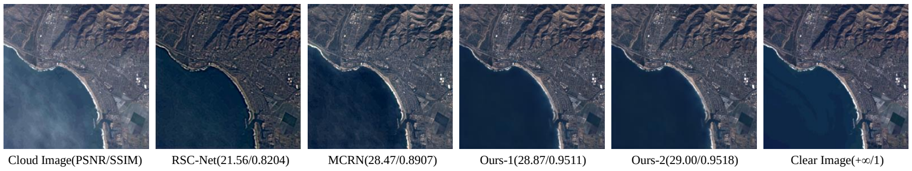
$$z = \mu + \sigma \cdot \varepsilon \ . \tag{4}$$

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.

In designing the network architecture, we exploit the self-attention mechanism to learn the mapping between degeneration factors and standard normal distribution. Different from the original ViT [27], we design a more lightweight model, reducing the number of stacked layers of the transformer encoder, the embedding dimension, and the number of heads of the multi-head attention mechanism. Let $\mathcal{D} = \{X_i, Y_i^{ref}\}_{i=1}^N$ be the training dataset, where X_i denotes the cloud image, Y_i^{ref} denotes the corresponding cloud-free image. The goal of the encoder network is as follows:

$$\phi^* = \arg\min_{\phi} D_{KL}(q(z|X;\phi)||p(z|X;\theta)) \ . \tag{5}$$

The whole pipeline of the encoder network during training and testing is illustrated in Figure 2.



Cloud Image(PSNR/SSIM) RSC-Net(21.56/0.8204) MCRN(28.47/0.8907) Ours-1(28.87/0.9511) Ours-2(29.00/0.9518) Clear Image(+∞/1)

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

3.3 Multi-Scale Prediction Decoder

With the latent variables z obtained by the inference network and the original input cloud image, we develop a decoder network to reconstruct the cloud-free image. In the decoder, the variable z is first passed through a generative module $p_\theta(W|z)$ (See Figure 2), which contains a MLP with two hidden layers and generates the image manifold of degeneration factors W_i. The latent variables z allow for modeling multiple modes, making the decoder network suitable for modeling one-to-many mapping. As shown in Figure 3, the clear images recovered from different sampling results are different in PSNR and SSIM scores.

We downsample the input cloud image X to the same size as W_i. After that, we concatenate these two variables and feed them to the first layer of the decoder network. The decoder network consists of five convolutional layers. Each layer focuses on extracting features and learning the underlying principles of de-clouding at different scales. The input to each layer is composed of the original cloud image as conditional variable and feature maps from the previous layer. We utilize the channel attention mechanism [28] to invest more learning attention on valuable matters.

The decoder network combines the latent variables z_i obtained by a single sampling and the original cloud image to produce a clear image. It can be written with the formula as $Y_i = f_D(X, z_i)$, where $f_D(\cdot)$ is the decoder network. We sample multiple variables z_i to produce multiple clear images corresponding to the original cloud image. We take the expectation of these multiple clear images as the final clear image. This procedure can be written as:

$$Y = \frac{1}{M}\sum_{i=1}^{M} Y_i \ . \tag{6}$$

where we set $M = 10$ and the goal is to maximize the posterior probability:

$$\theta^* = \arg\max_{\theta} \log p_\theta(Y|X, z) \ . \tag{7}$$

3.4 Loss Function

The loss of the overall CVAE framework consists of two parts: the inference network and the decoder network. Combining (5) and (7), it can be formulated as follow:

$$\mathcal{L}_{CVAE} = D_{KL}(q_\phi(z|X)||p_\theta(z|X)) + \frac{1}{M}\sum_{m=1}^{M} -\log p_\theta(Y|X, z^{(m)}) \ . \tag{8}$$

where M is the number of samples. The first term in the loss function is the KL divergence. We assume that $q_\phi(z|X)$ obeys a normal distribution with parameters μ, $\sigma^2 I$, and $p_\theta(z|X)$ obeys the standard normal distribution, then this loss can directly calculate the closed-form solution:

$$D_{KL}(q(z|X;\phi)||p(z|X;\theta)) = \frac{1}{2}(tr(\sigma^2 I) + \mu^T\mu - d - \log(\left|\sigma^2 I\right|)) \ . \tag{9}$$

where the $tr(\cdot)$, $|\cdot|$ represent the trace and determinant of the matrix, respectively, d is the dimensions of the distribution.

The second term is the reconstruction loss. In supervised training, the declouding performance can be quantified by counting the differences between the encoder network output Y with its corresponding reference clear image Y^{ref} under some proper loss L, *e.g.* the L_1 norm and Mean Square Error (MSE). In our method, we choose SmoothL1Loss as the criterion to optimize the parameters and the reconstruction loss can be expressed as:

$$\mathcal{L}_{rec} = \frac{1}{CP}\sum_{c=1}^{C}\sum_{p=1}^{P} F_s((Y_c(p) - Y_c^{ref}(p)); \beta) \ . \tag{10}$$

where

$$F_s(e;\beta) = \begin{cases} 0.5e^2/\beta \ \ if\,|e| < \beta, \\ |e| - \beta/2 \ otherwise. \end{cases} \tag{11}$$

C represents the channel number and P denotes the total number of pixels.

In addition, to guide the network to focus on details, we introduce an edge loss function to preserve the edges of the output image. The edge loss is defined as:

$$\mathcal{L}_{edge} = \ \left\|\nabla^2 Y - \nabla^2 Y^{ref}\right\| \ . \tag{12}$$

where ∇^2 is the Laplace operator for image edge detection.

Based on the above consideration, the total loss function for the proposed CVAE cloud removal network is as follows:

$$\mathcal{L} = \lambda_1 D_{KL} + \lambda_2\mathcal{L}_{rec} + \lambda_3\mathcal{L}_{edge} \ . \tag{13}$$

where λ_1, λ_2, and λ_3 represent the weight parameters.

4 Experiments

To demonstrate the superiority of the proposed algorithm, we compare it with several state-of-the-art methods including the prior-based method (DCP [9]), the data-driven methods (RSC-Net [11], SPA-GAN [29], MSAR-DefogNet [12], PCFAN [30], Pix2Pix [31]), and the methods combining deep learning with physical models (MCRN [32], Qin *et al.* [16], Zheng *et al.* [17]). For fair comparisons, we train these networks on the same dataset with the same learning rate and the number of training epochs. We use full-reference image quality evaluation metrics PSNR and SSIM [33] and no-reference evaluation metrics NIQE [34] and BRISQUE [35] for quantitative evaluation.

4.1 Dataset

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 counterpart separated by one satellite re-entry period (16 days). We select the image pairs which has similar lighting conditions and crop them into 256 x 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.

There are three main characteristics in the proposed dataset: (1) T-CLOUD is a large-scale natural benchmark for remote sensing image thin cloud removal while the previous datasets are only composed of synthetic data; (2) T-CLOUD includes many different ground scenarios such as cities, mountains, and coasts; (3) The proposed dataset is much more challenging because the cloud is non-homogeneous and the texture details of the image are more complex.

Note that these cloud and cloud-free image pairs are captured by the same satellite sensor at different times, the illumination noises are unavoidable due to the change of ambient light. Although we try to select images with the same lighting conditions as possible, achieving outstanding results on this dataset is still a challenging task.

4.2 Implementation Details

The proposed algorithm is implemented with the PyTorch framework. The hardware facilities of the computing platform include an Intel Gold 6252 CPU and an NVIDIA A100 GPU. To optimize the proposed network, we use the Adam optimizer [36] with parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. The batch size and training epochs are set to 1 and 300, respectively. The initial learning rate is set to 0.0001, which decreases by 10 times after training for half the number of epochs. We set the size of the interference factor image manifold to 16 x 16. The values of the parameters λ_1, λ_2, and λ_3 are empirically set to 0.01, 1.0 and 0.18, respectively.

4.3 Results on Real Dataset

Table 1 shows the quantitative results in terms of PSNR, SSIM, NIQE, and BRISQUE on our dataset. It can be seen that our method achieves the best performance on both the full-reference metrics PSNR and SSIM. Also in the

Table 1. Quantitative Evaluations on the real-world dataset. Where red text and blue text indicate the best and second-best performance, respectively. ↑: The larger the better. ↓: The smaller the better.

Method	PSNR↑	SSIM↑	NIQE↓	BRISQUE↓
DCP	19.94	0.6646	3.300	53.60
Qin *et al.*	26.35	0.8035	3.306	47.39
RSC-Net	23.98	0.7596	3.407	50.24
SPA-GAN	15.36	0.5280	2.860	56.90
Zheng *et al.*	23.71	0.7630	2.969	51.01
MSAR-DefogNet	28.84	0.8432	2.786	51.07
PCFAN	28.27	0.8342	2.800	50.69
MCRN	26.60	0.8091	2.888	50.73
Pix2Pix	28.77	0.8476	2.677	50.77
Ours	30.14	0.8600	2.762	49.61

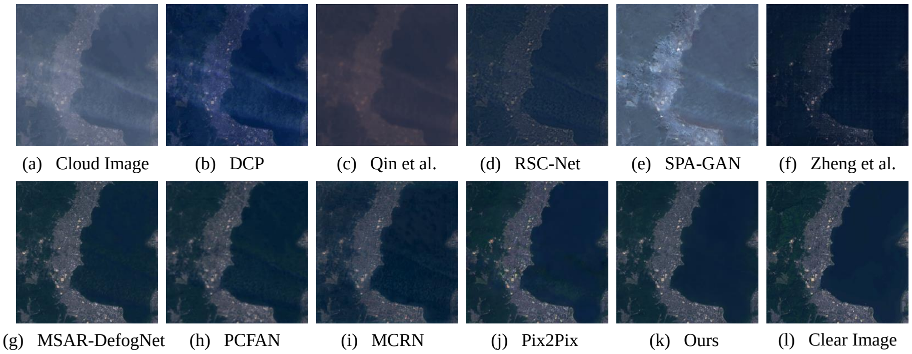
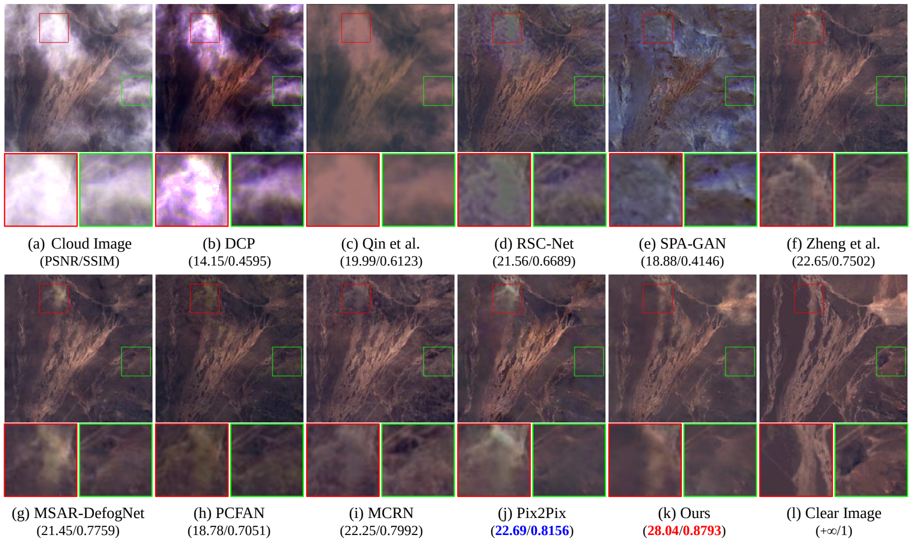

(a) Cloud Image (b) DCP (c) Qin et al. (d) RSC-Net (e) SPA-GAN (f) Zheng et al.
(g) MSAR-DefogNet (h) PCFAN (i) MCRN (j) Pix2Pix (k) Ours (l) Clear Image

Fig. 4. Thin cloud removal results of the real-world cloud image. Zoom in for a better view. More examples can be found in supplementary material.

no-reference metrics, excellent results have been achieved.

For evaluating the visual effect of each method, we compare the qualitative results (see Figure 4). It can be observed that DCP tends to over-enhance the cloud image and is unable to remove the dense cloud. The CNN-based methods achieve competitive results, however, some of them are still poor in visual quality. For example, restoration results of Qin *et al.* and RSC-Net miss a lot of detailed information. SPA-GAN fails to remove the cloud. The method of Zheng *et al.* has some obvious grid artifacts. The de-clouding results by MSAR-DefogNet, PCFAN, MCRN, and Pix2Pix have different degrees of color distortion in the sea area (Figure 4(g)∼4(j)). The reason for these low-quality results may be their methods are mostly designed based on the principle of synthetic images so that the underlying cloud removal laws cannot be well learned on real-world cloud images. In contrast, the result of our method shown in Figure 4(k) has



(a) Cloud Image (PSNR/SSIM)
(b) DCP (14.15/0.4595)
(c) Qin et al. (19.99/0.6123)
(d) RSC-Net (21.56/0.6689)
(e) SPA-GAN (18.88/0.4146)
(f) Zheng et al. (22.65/0.7502)
(g) MSAR-DefogNet (21.45/0.7759)
(h) PCFAN (18.78/0.7051)
(i) MCRN (22.25/0.7992)
(j) Pix2Pix (**22.69**/**0.8156**)
(k) Ours (**28.04**/**0.8793**)
(l) Clear Image (+∞/1)

Fig. 5. Thin cloud removal results of the hard examples. Zoom in for a better view.

high color fidelity and is more effective in texture detail preservation.

4.4 Results on Hard Examples

One of the most distinctive characteristics of clouds is that they are heterogeneous. We further select some images with thick clouds and evaluate our method on this non-homogeneous dataset. The uneven cloud images are very common in remote sensing images. Therefore, comparing the performance of different methods on these challenging data can verify their effectiveness in practical applications.

The visual results are illustrated in Figure 6. The result produced by RSC-Net suffers from severe color distortion. The GAN-based method SPA-GAN fails to keep the semantic information consistent. Zheng *et al.*, MSAR-DefogNet, PC-FAN, MCRN, and Pix2Pix yield unnatural results with varying degrees of color distortion. Recovering details under thick clouds is highly uncertain due to the heavy loss of scenario information. Our method keeps the semantic information as consistent as possible through probability estimation. It can be observed that our algorithm can effectively mitigate color distortion compared to other methods. The quantitative results also demonstrate that our method has overwhelming advantages. It is much higher than other methods in terms of PSNR and SSIM. Our PSNR is higher than the Pix2Pix with 5.35dB showing that the CVAE framework is more robust in such scenarios.

Table 2. Quantitative Evaluations on the clear images.

	DCP	Qin *et al.*	RSC-Net	SPA-GAN	Zheng *et al.*	MSAR-DefogNet	PCFAN	MCRN	Pix2-Pix	Ours
PSNR↑	21.46	22.87	25.99	15.77	28.21	27.78	25.43	26.14	28.01	28.76
SSIM↑	0.8585	0.7018	0.9079	0.6237	0.9268	0.9174	0.8948	0.8999	0.9224	0.9448

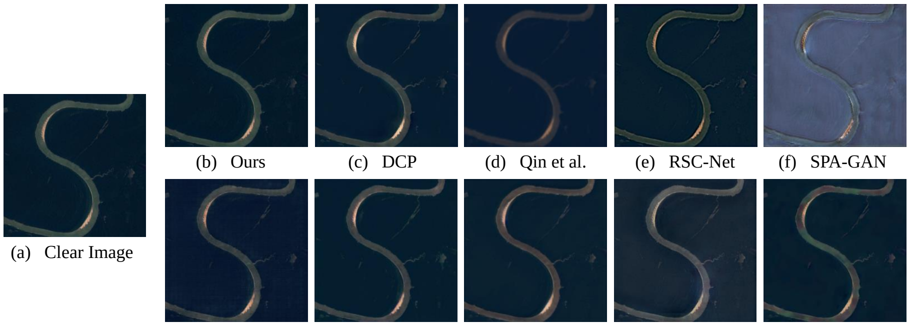

(a) Clear Image (b) Ours (c) DCP (d) Qin et al. (e) RSC-Net (f) SPA-GAN (g) Zheng et al. (h) MSAR-DefogNet (i) PCFAN (j) MCRN (k) Pix2Pix

Fig. 6. The visual comparison of image fidelity on the clear image. Zoom in for a better view.

4.5 Results on Clear Dataset

To further verify the fidelity of our algorithm, we additionally compare the results of various de-clouding methods on a clear image dataset. The visual comparisons are shown in Figure 5. It can be observed that several methods (DCP, Qin *et al.*, MSAR-DefogNet, PCFAN) tend to over-darken or over-enhance the clear image and are inconsistent in the color space. The checkerboard artifacts can be observed in Zheng *et al.* and SPA-GAN changes the original scene radiance information. Our result does not produce color distortion and is very close to the original clear image, which is necessary for practical applications because not all image patches obtained by satellite sensors are occluded by clouds. Table 2 shows the quantitative results. The PSNR and SSIM values also indicate that our algorithm surpasses other methods in terms of image fidelity.

5 Ablation Study

To further demonstrate the effectiveness of the proposed algorithm, we conduct ablation experiments to verify whether the network structure is effective. The ablation experiments are divided into the following parts: 1) the structure of the encoder network; and 2) the choice of the reconstruction loss function.

Table 3. Ablation studies on the structure of encoder network.

Encoder-Network	PSNR↑	SSIM↑	Size(Mb)
ResNet34	28.42	0.8512	111.95
VGG-19	28.37	0.8495	567.71
ViT-Base	28.44	0.8501	244.31
ViT-Large	28.56	0.8518	798.11
ViT-Small(Ours)	30.14	0.8600	58.86

5.1 Evaluations on The Encoder Network

In our method, the encoder network is used to infer the distribution of degeneration factors. First, we conducted ablations on the structures of encoder architecture. The configurations of variant models including: 1) VGG-19; 2) ResNet34; 3) ViT-base; 4) ViT-Large; 5) a ViT variant model proposed by us: ViT-Small. The original ViT stacks multiple layers of transformer encoders and embeds the input image patches into a high-dimensional vector space, which makes its parameters very large. To make the network more lightweight, we design the ViT-Small, which simplifies the original transformer structure. We reduce the stacking layers of the encoder from 12 (ViT-Base) or 24 (ViT-Large) to 4, the embedding dimension of the image patches from 768 (ViT-Base) or 1024 (ViT-Large) to 512, and set the number of heads of the multi-head attention mechanism to 8.

The quantitative results are shown in Table 3. The CNN-based architectures achieve similar performance. However, they are worse than the three ViT structures, which indicates that the self-attention mechanism of ViT can better learn the distribution of degeneration factors in the latent space. Meanwhile, we also noticed that the parameter amount of ViT-Small is much lower than the other four structures. The lightweight ViT not only did not weaken the powerful representation ability but also made the model better. This indicates that this lightweight variant of vision transformer in the inference network is effective.

5.2 Evaluations on The Reconstruction Loss Function

We also explore the influence of different reconstruction loss functions in the optimization process. The reconstruction loss directly measures the quality of the cloud removal effect. Therefore, choosing a good measurement criterion plays a crucial role in the training process.

We compare the impact of MSE and SmoothL1 (beta=0.5) loss function and verify the boosting effect of the edge loss. For a quick comparative experiment, the models with different loss functions are trained for 150 epochs. The experimental results are shown in Figure 7. It can be seen that the SmoothL1 loss function achieves better optimization results in PSNR and SSIM during the training process. The SSIM shows that the optimization results are significantly improved after adding the edge loss, which demonstrates the edge loss can promote the optimization of the network.

Table 4. Ablation studies on the reconstruction loss function.

	SmoothL1	MSE	Edge Loss	PSNR↑	SSIM↑
1	-	✓	-	27.14	0.8293
2	✓	-	-	27.57	0.8302
3	-	✓	✓	27.70	0.8417
4	✓	-	✓	27.83	0.8421

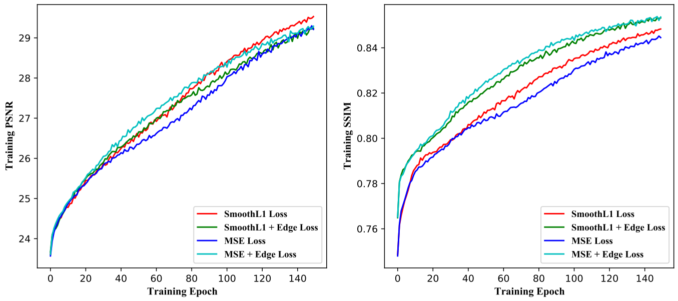


Fig. 7. Graph of PSNR and SSIM with different objective function during training process.

We also tested the generalization performance to verify that achieving superior performance on the training set is not due to overfitting. It can be seen from Table 4 that the model trained with SmoothL1 joint edge loss on the test set also has better PSNR and SSIM scores. At the same time, the higher SSIM shows that the edge loss can make the model pay more attention to the detailed textures.

6 Conclusion

In this paper, we propose a thin cloud removal network based on CVAE and tackle single image de-clouding from the perspective of uncertainty analysis. The novelty of the proposed algorithm is that it can achieve one-to-many mapping, and can generate multiple clear and reasonable images corresponding to a single cloud image. Moreover, we construct a large-scale dataset from the real world to overcome the shortcomings of synthetic datasets that cannot fully represent real scenes. Both quantitative and qualitative experimental results show the superiority of our proposed method and demonstrate that considering uncertainty has great potential to improve the thin cloud removal algorithm.

Acknowledgements 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.

References

1. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018)
2. Zhang, Z., Zhao, L., Liu, Y., Zhang, S., Yang, J.: Unified density-aware image dehazing and object detection in real-world hazy scenes. In: Proceedings of the Asian Conference on Computer Vision (2020)
3. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: End-to-end united video dehazing and detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
4. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. International Journal of computer vision **63**(2), 113–140 (2005)
5. Tarel, J.P., Hautiere, N., Cord, A., Gruyer, D., Halmaoui, H.: Improved visibility of road scene images under heterogeneous fog. In: 2010 IEEE intelligent vehicles symposium. pp. 478–485. IEEE (2010)
6. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (2018)
7. Richter, R.: A spatially adaptive fast atmospheric correction algorithm. International Journal of Remote Sensing **17**(6), 1201–1214 (1996)
8. Vermote, E.F., Tanré, D., Deuze, J.L., Herman, M., Morcette, J.J.: Second simulation of the satellite signal in the solar spectrum, 6s: An overview. IEEE transactions on geoscience and remote sensing **35**(3), 675–686 (1997)
9. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
10. Xu, M., Deng, F., Jia, S., Jia, X., Plaza, A.J.: Attention mechanism-based generative adversarial networks for cloud removal in landsat images. Remote Sensing of Environment **271**, 112902 (2022)
11. Li, W., Li, Y., Chen, D., Chan, J.C.W.: Thin cloud removal with residual symmetrical concatenation network. ISPRS Journal of Photogrammetry and Remote Sensing **153**, 137–150 (2019)
12. Zhou, Y., Jing, W., Wang, J., Chen, G., Scherer, R., Damaševičius, R.: Msardefognet: Lightweight cloud removal network for high resolution remote sensing images based on multi scale convolution. IET Image Processing **16**(3), 659–668 (2022)
13. Zi, Y., Xie, F., Zhang, N., Jiang, Z., Zhu, W., Zhang, H.: Thin cloud removal for multispectral remote sensing images using convolutional neural networks combined with an imaging model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 3811–3823 (2021)
14. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing **25**(11), 5187–5198 (2016)
15. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE international conference on computer vision. pp. 4770–4778 (2017)
16. Qin, M., Xie, F., Li, W., Shi, Z., Zhang, H.: Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture. IEEE journal of selected topics in applied earth observations and remote sensing **11**(5), 1645–1655 (2018)

17. Zheng, J., Liu, X.Y., Wang, X.: Single image cloud removal using u-net and generative adversarial networks. IEEE Transactions on Geoscience and Remote Sensing **59**(8), 6371–6385 (2020)
18. Chavez Jr, P.S.: An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote sensing of environment **24**(3), 459–479 (1988)
19. Fattal, R.: Dehazing using color-lines. ACM transactions on graphics (TOG) **34**(1), 1–14 (2014)
20. Berman, D., Avidan, S., et al.: Non-local image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1674–1682 (2016)
21. Xu, M., Pickering, M., Plaza, A.J., Jia, X.: Thin cloud removal based on signal transmission principles and spectral mixture analysis. IEEE Transactions on Geoscience and Remote Sensing **54**(3), 1659–1669 (2015)
22. Mao, X., Shen, C., Yang, Y.B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Advances in neural information processing systems **29** (2016)
23. Singh, P., Komodakis, N.: Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 1772–1775. IEEE (2018)
24. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.H.: Single image dehazing via multi-scale convolutional neural networks. In: European conference on computer vision. pp. 154–169. Springer (2016)
25. Makarau, A., Richter, R., Müller, R., Reinartz, P.: Haze detection and removal in remotely sensed multispectral imagery. IEEE Transactions on Geoscience and Remote Sensing **52**(9), 5895–5905 (2014)
26. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. International journal of computer vision **48**(3), 233–254 (2002)
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
28. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
29. Pan, H.: Cloud removal for remote sensing imagery via spatial attention generative adversarial network. arXiv preprint arXiv:2009.13015 (2020)
30. Zhang, X., Wang, T., Wang, J., Tang, G., Zhao, L.: Pyramid channel-based feature attention network for image dehazing. Computer Vision and Image Understanding **197**, 103003 (2020)
31. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
32. Yu, W., Zhang, X., Pun, M.O., Liu, M.: A hybrid model-based and data-driven approach for cloud removal in satellite imagery using multi-scale distortion-aware networks. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 7160–7163. IEEE (2021)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
34. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)

35. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)

36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)