A Coarse to Fine Thin Cloud Removal Network with Pyramid Non-local Attention

GUAN Wang[1], *TIAN Zhenkai*[2], *MA Tao*[3], *ZHAO Lingyuan*[4], *XIE Shizhe*[5], *YAN Jin*[6], *DU Yang*[1*], *ZOU Yunkun*[4*]

1. College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, P. R. China;
2. The Fourth Topographic Surveying Brigade of Ministry of Natural Resources, Harbin 150025, P. R. China;
3. The Second Institute of Geographic Information and Mapping of the Ministry of Natural Resources, Harbin 150025, P. R. China;
4. Huantian Wisdom Technology Co., Ltd., Meishan 620500, P. R. China;
5. China Energy Digital Intelligence Technology Development (Beijing) Co., Ltd., Beijing 100011, P. R. China;
6. National Earthquake Response Support Service, Beijing 100049, P. R. China

(Received 10 March 2025; revised 15 June 2025; accepted 1 September 2025)

Abstract: In remote sensing imagery, approximately 67% of the data are affected by cloud cover, significantly increasing the difficulty of image classification, recognition, and other downstream interpretation tasks. To effectively address the randomness of cloud distribution and the non-uniformity of cloud thickness, we propose a coarse-to-fine thin cloud removal architecture based on the observations of the random distribution and uneven thickness of cloud. In the coarse-level declouding network, we innovatively introduce a multi-scale attention mechanism, i.e., pyramid non-local attention (PNA). By integrating global context with local detail information, it specifically addresses image quality degradation caused by the uncertainty in cloud distribution. During the fine-level declouding stage, we focus on the impact of cloud thickness on declouding results (primarily manifested as insufficient detail information). Through a carefully designed residual dense module, we significantly enhance the extraction and utilization of feature details. Thus, our approach precisely restores lost local texture features on top of coarse-level results, achieving a substantial leap in declouding quality. To evaluate the effectiveness of our cloud removal technology and attention mechanism, we conducted comprehensive analyses on publicly available datasets. Results demonstrate that our method achieves state-of-the-art performance across a wide range of techniques.

Key words: channel attention; thin cloud removal network; pyramid non-local attention (PNA); remote sensing image; residual dense connection

CLC number: TN925 **Document code:** A **Article ID:** 1005-1120(2025)05-0589-12

0 Introduction

Reaching the satellite sensor, the signal is susceptible to degradation due to the absorption and scattering of atmospheric particles like mid-altitude clouds. These atmospheric conditions result in blurred image details and missing content, significantly reducing the quality of remote sensing image (RSI). As per statistics, approximately 67% of the land surface is covered by clouds[1]. Based on their transmittance capabilities, clouds are typically classified as "thin clouds" and "thick clouds". For thick cloud removal, existing algorithms relying solely on single images exhibit low content credibility, while traditional methods involving multi-source images for atmospheric correction and pixel registration are highly complex. Unlike the irreversible information loss caused by thick clouds, thin clouds cover a broader area and occur with greater frequency. Removing clouds from a single image better preserves the image's detailed information and original features, thereby achieving higher fidelity. Consequent-

*Corresponding authors, E-mail addresses: dudujia@nefu.edu.cn; zouyk@htwisdom.com.

How to cite this article: GUAN Wang, TIAN Zhenkai, MA Tao, et al. A coarse to fine thin cloud removal network with pyramid non-local attention[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025,42(5):589-600.
http://dx.doi.org/10.16356/j.1005-1120.2025.05.003

ly, the development of thin cloud removal technology for remote sensing images has garnered significant attention. However, removing thin clouds based on a single image remains a challenging and highly ill-conditioned task, including thin clouds[2-4]. Numerous methods have been developed for the removal of thin cloud from RSI. These methods can be broadly classified into two main categories: Traditional thin cloud removal algorithms and declouding algorithms based on neural network.

Traditional image-processing-based thin cloud removal methods rely on simplified models or priors. Through data sampling, He et al.[5] discovered the dark channel prior and conducted research on image declouding using the atmospheric scattering model, which yielded positive results. However, the restoration effect on the sky area was found to be inadequate. Gao et al.[6] addressed the issues of overly smooth and missing details in Ref.[5] method by incorporating the morphological reconstruction method. Zhu et al.[7] proposed a dehazing method that estimates the transmittance by minimizing the energy function, effectively resolving the drawbacks of the dark channel prior method. These traditional model- or prior-based methods rely heavily on handcrafted features, and their restoration results have low accuracy and robustness for remote sensing images with various ground cover conditions and complex textures electromagnetic interference or power consumption.

In recent years, the neural network algorithm represented by data support has shown advanced performance in the field of image thin cloud removal. Jing et al.[8] proposed a multi-scale residual convolutional neural network for thin cloud removal of remote sensing images and it took the synthesized thin cloud patches as input and outputs the corresponding transmission value. Li et al.[9] designed a multi-input and output dehazing network based on the band characteristics of remote sensing images. However, they may not lead to significant performance improvements by simply increasing the number of layers or using wider layers. Ma et al.[10] obtained cloud information from the perspective of linear mixing of image overlay and successfully restored the surface information of thin cloud areas. This method relies too much on prior knowledge and performs well in scenes with light cloud cover or uniform cloud distribution, but performs poorly in the face of thick clouds or irregularly distributed cloud layers[11]. Liu et al.[12] used a two-stage supervised network to stratify and remove clouds, enhancing the clarity and contrast of the image and providing a reliable data source for subsequent small object detection. Cai et al.[13] proposed DehazeNet, where they employed a neural network model to learn the direct mapping relationship between foggy images and transmittance maps. Similarly, Ren et al.[14] also predicted transmission map with convolutional neural network (CNN). But the atmospheric scattering model is a simplified approximation of the thin cloud effect, making it challenging to achieve a clear reconstruction effect solely based on the transmittance map. As a result, the two-stage dehazing network, which progresses from coarse to fine, has gained significant attention. Zhao et al.[15] introduced a two-stage weak supervision framework called RefineDNet. This framework first utilizes the dark channel prior to restore visibility and then employs GAN to enhance authenticity. Tran et al.[16] considered different fog densities and adopted an efficient pooling mechanism to replace the traditional SA module, significantly reducing the computational demand. Zhang et al.[17] estimated the atmospheric light A and transmittance T using two CNN network models. They then combined these estimates with the GAN network to obtain the final clear picture. Li et al.[18] designed a two-stage dehazing network architecture to address the problem of missing detailed information during the dehazing process, reconstructing the firststage dehazing features with multi-scale detail adjustment. Du et al.[19] and Li et al.[20] had also proposed a two-stage repair network that follows a physical model to remove fog and heavy rain scenarios, respectively. Although the aforementioned methods achieve satisfactory results in fog removal, atmospheric particles in complex satellite scenes are susceptible to the coupled effects of wind speed, humidity, illumination direction, and terrain occlusion. Consequently, fog concentration exhibits random and non-uniform spatial variations[21]. This leads to content estimation errors in

traditional defogging algorithms at fog boundaries and transition zones. These errors can cause color loss, halo artifacts, and detail loss. Therefore, explicitly modeling the randomness and non-uniformity of fog distribution within the defogging framework is an effective solution.

Furthermore, multi-scale representations and recursive reasoning have demonstrated powerful contextual modeling capabilities across numerous low-level computer vision tasks. Multiscale strategies capture both broad fog concentration trends across different spatial resolutions by concurrently or sequentially extracting features from "global semantics" to "local textures". Recursive mechanisms progressively refine reconstructions from coarse to fine scales, feeding high-level semantic priors back into low-level detail recovery to enhance edge consistency. Theoretically, this organic integration of both approaches holds promise to overcome the expressive limitations of traditional "single-scale-single-step" defogging frameworks for non-uniform fog distributions, offering a novel research paradigm for image defogging in complex scenes. In the current research, thin cloud scenarios and fog scenarios are used interchangeably, and this paper uniformly uses thin cloud for description[22-24].

Based on this, this paper revisits the degradation mechanisms of non-uniform thin cloud images and proposes a coarse-to-fine thin cloud removal method with a refined network. During the coarse declouding stage, we introduce a plug-and-play pyramid non-local attention mechanism module to enhance the fusion of global and local features. This mechanism captures semantic information and details at different levels, thereby improving the network's ability to restore areas with non-uniform thin cloud concentration. During the refinement declouding stage, we focus on mitigating the impact of cloud thickness on declouding results. Through a cleverly designed residual dense module, we significantly reduce the loss of feature details. Consequently, this approach precisely restores lost local texture features on top of coarse declouding results, achieving a further leap in declouding quality.

1 Methodology

As illustrated in Fig.1, we propose a two-stage remote sensing image thin cloud network that progressively refines reconstructions from coarse to fine scales. This approach feeds high-level semantic priors back into low-level detail recovery to enhance edge consistency. In the first stage, we concatenate the obtained multi-scale feature maps and project them onto the feature map matrices K_T and V_T, then

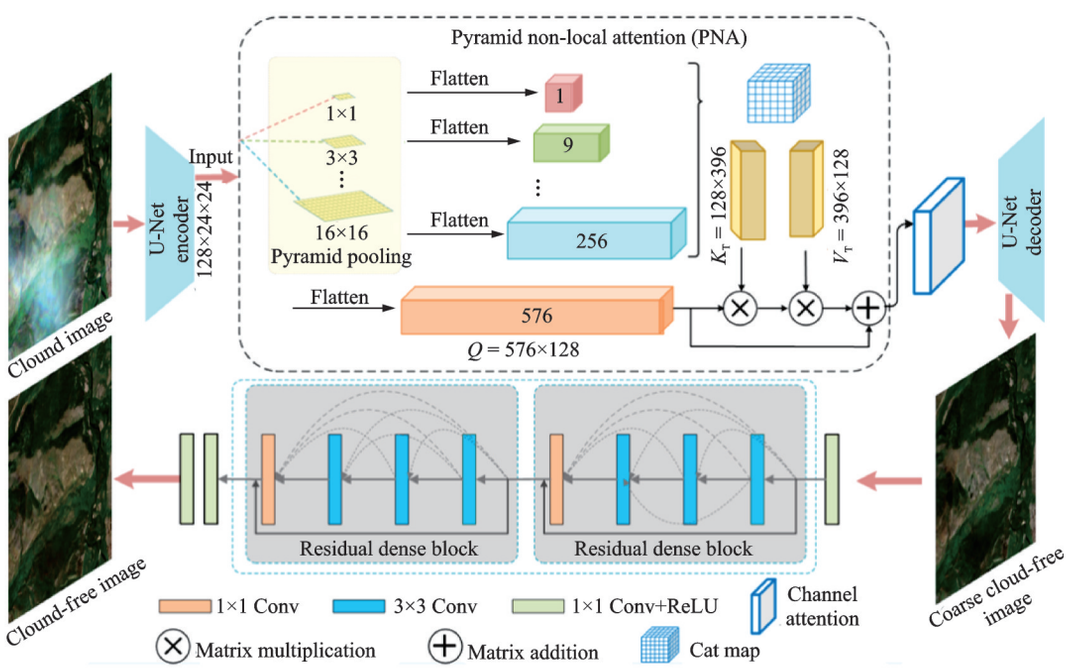


Fig.1 Proposed cloud removal network

we combine two attention mechanisms by simultaneously considering spatial distribution and channel importance to enhance thin cloud region restoration. In the second stage, we develop a recursive block to further improve declouding performance without significantly increasing training parameters.

1.1 Coarse thin cloud removal stage

Since the attention mechanism can significantly improve the performance of the network with minimal cost, it is widely used. In this paper, we continuously use the decoder-encoder structure of U-Net network and consider designing multi-scale attention mechanism from both spatial and channel dimensions to improve the performance of the coarse cloud removal network spatial features. In order to balance the computational efficiency and performance of the network, we provide a pyramid non-locale attention (PNA). PNA can integrate different levels of features through progressive pyramid pooling, which can not only enhance the network's feature extraction capability for the different distribution of cloud layers, but also effectively improve the network's integration effect of global and local semantic information. In order to reduce the impact of PNA on the computational efficiency of the network, no additional convolution operation is added in the pyramid pooling but adaptive pooling is used directly. The reason why K and V are forced to be consistent here is also considered for efficiency. In addition, in terms of feature channels, we additionally use channel attention to make up for the lack of PNA in cross-channel information integration, so that the model can adaptively re-weight each channel, thus strengthening effective, suppressing redundant responses, and further improving the compactness and cloud layer discrimination of the representation.

Feature maps with 128 channels and a pixel size of 24×24 will undergo adaptive pooling at five different scales (1, 3, 7, 9, 16). We validate the appropriateness of the scale in our ablation experiments to obtain features with different dimensions, and the final PNA can be defined as

$$\begin{aligned}&\mathbf{PHA}_{\text{out}}=\text{reshape}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{X}\\&\mathbf{PHA}_{\text{out}}\in\boldsymbol{R}^{c\times h\times w},c=128,h=w=24\end{aligned}\tag{1}$$

where $\boldsymbol{R}$ is the feature space, and c the number of channels; h and w represent the length and width of features.

After obtaining the feature map with the initial encoding, we perform multi-scale processing: $\boldsymbol{X}_1\in\boldsymbol{R}^{c\times h\times w}$, $c=128$, $h=w=1$; $\boldsymbol{X}_3\in\boldsymbol{R}^{c\times h\times w}$, $c=128$, $h=w=3$, $\boldsymbol{X}_7\in\boldsymbol{R}^{c\times h\times w}$, $c=128$, $h=w=7$, $\boldsymbol{X}_9\in\boldsymbol{R}^{c\times h\times w}$, $c=128$, $h=w=9$, $\boldsymbol{X}_{16}\in\boldsymbol{R}^{c\times h\times w}$, $c=128$, $h=w=16$. Then all these values will be flattened into a new 1D vector: $\boldsymbol{Q}\in\boldsymbol{R}^{n\times c}$, $n=576$, $c=128$.

1.2 Fine thin cloud removal stage

The uneven distribution of thin cloud poses a challenge to the network's performance. While increasing the depth of the network can be beneficial for tasks like semantic segmentation, classification, and target detection (where pixels are classified into specific categories), it is not as effective for thin cloud removal[25]. Therefore, we propose a recursive block using a residual dense block (RDB) to reuse features and enhance image detail restoration by superimposing detail information on advanced information. This ensures uninterrupted information flow between network layers and effectively enhances the thin cloud removal effect.

Specifically, in RDB, the original input features can access network layers of different depths one by one through dense connections. This allows for a continuous state transmission and the conveyance of information that needs to be preserved. RDB is composed of four convolution layers, with the first three layers expanding the dimension of feature maps and the last layer fusing these feature maps. The input channel size and growth rate in each RDB are set to 16. The output result of RDB can be expressed as

$$\text{RDB}_{\text{out}}=\text{ReLU}(\text{Conv}(F_1,F_2,F_3)+\text{Conv}(F_1))\tag{2}$$

where ReLU represents the activation function, and Conv the convolution operation; F_1, F_2 and F_3 represent feature maps after Conv and ReLU.

1.3 Loss function

To achieve an end-to-end training network, we

used the average absolute error loss L_1 and perceptual loss L_p for joint optimization, and the loss calculation process is as follows

$$\begin{cases} L_{coarse} = L_1(J_{coarse} - I_{gt}) + L_p(\phi(J_{coarse}) - \phi(I_{gt})) \\ L_{refine} = L_1(J_{refine} - I_{gt}) + L_p(\phi(J_{refine}) - \phi(I_{gt})) \\ L_{finall} = L_{coarse} + L_{refine} \end{cases} \quad (3)$$

where J_{coarse} and J_{refine} represent the dehazed image optimized by coarse and fine dehazing stage; I_{gt} represents the clear image of the real scene; ϕ corresponds to the output of the 14th layer of the VGG19; and L_{finall} represents the final loss function output.

2 Experiments

2.1 Dataset introduction

WHUS2-CR dataset[26]: It is derived from the Sentinel-2 satellite and consists of cloud images captured in various areas between 2016 and 2021. The dataset covers different types of land such as vegetation, water, cities, bare land, and snow/ice. It contains a total of 24 450 images, with resolutions of 10, 20, and 60 m. For training and testing purposes, we chose 384 pixel×384 pixel size data with 10 m resolution for training.

RICE dataset[27]: It comes from the Landsat8 satellite and consists of 500 sets of thin cloud data and 736 sets of thick cloud data. Each image has a size of 384 pixel× 384 pixel. Subsequently, the dataset was divided into a 4∶1 ratio, with 80% used as a training set and 20% as a test set.

2.2 Exprimental details

In the training process of declouding network, the Adam optimizer is used, and parameters are set to default. The model is trained with batch size of 5 and a termination iteration of around 300 epochs. In terms of evaluation indicators, the widely used structural similarity (SSIM), peak signal-to-noise ratio (PSNR) and mean squared error (MSE) are adopted. We validated the effectiveness of our method by utilizing seven approaches from 2020 to 2025, based on the WHUS2-CR and RICE datasets. In the ablation experiment, we used the same training method and parameter settings to validate the attention network.

2.2.1 Comparison of dehazing methods

MSBDN[27]: Ref.[27] proposed a multi-scale enhanced dehazing network with dense feature fusion based on boosting and error feedback principles.

LapDehazeNet[28]: Ref.[28] introduced the principle of infinite approximation of Taylor's theorem with the Laplace pyramid pattern to build a dehazing model, in which low-order polynomials reconstructed the low-frequency information of the image, and high-order polynomials regressed the high-frequency information of the image.

Refusion[29]: Ref.[29] proposed a latent space diffusion model based on U-Net, which can diffuse in the low-resolution latent space while retaining the high-resolution information of the original input for decoding.

FCTF-Net[25]: Ref.[25] designed a first-coarse-then-fine two-stage dehazing neural network, with DensUnet as the baseline and extra channel attention.

CR4S2[9]: Ref.[9] designed a multi-input and output dehazing network based on the band characteristics of remote sensing images.

TFFDNet[18]: Ref.[18] designed a two-stage dehazing network architecture to address the problem of missing detailed information during the dehazing process, reconstructing the firststage dehazing features with multi-scale detail adjustment.

SENet[30]: Ref.[30] designed a novel fast dehazing framework based on the saturation algorithm, which used a new feature extraction convolution faster and more performant than the common 3×3 convolution, and reduced the information redundancy between the channels of the feature map, while significantly improving the computing efficiency.

2.2.2 Comparison of attention mechanisms

CBAM[31]: Ref.[31] proposed an attention mechanism module that integrated space and channel, which was extensively employed in the design of neural networks.

ECA-Net[32]: It primarily enhanced the SENet module by introducing a local cross-channel interaction strategy known as the ECA module. This approach does not involve dimensionality reduction

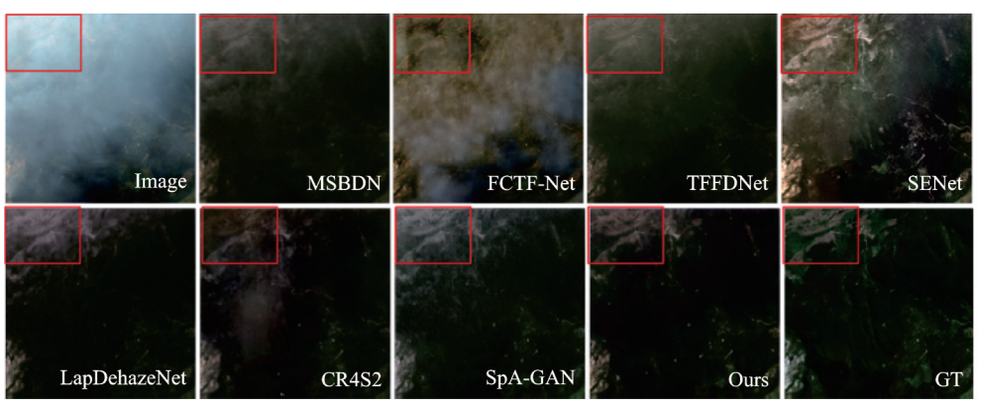
and incorporates a method for adaptively selecting the size of a one-dimensional convolution kernel.

As the ECA module, this approach does not involve dimensionality reduction and also incorporates a method for adaptively selecting the size of a one-dimensional convolution kernel.

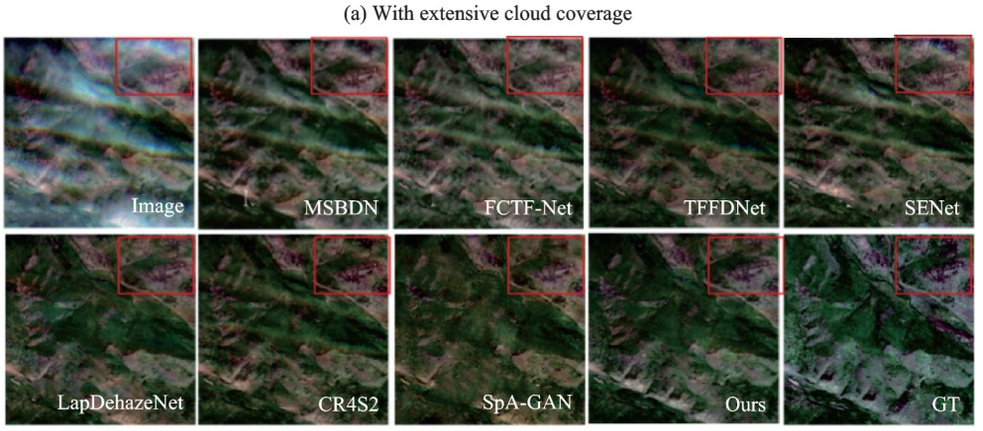
SA-Net[33]: Ref.[33] proposed a shuffle attention (SA) module. Initially, the shuffle unit was employed to depict the interdependence of features in the spatial domain and channel dimension of the channel split data. Subsequently, all the features were consolidated through channel shuffle.

2.3 Evaluation of results

As shown in Fig.2, the experimental results of various methods on WHUS-CR datasets are presented. GT represents the ground truth. The red rectangle box denotes the area of focused attention. For the cloud mist scattering in Fig.2(a), the image has a large amount of cloud mist diffusion phenomenon, which poses a high demand on thehazing algorithm for such a complex scenario. The experimental results show that the FCTF-Net, TFFDNet, and SENet methods all failed to effectively remove thin cloud, and there are still obvious traces of cloud mist in the image. Although MSBDN, LapDehazeNet, CR4S2, and SpA-GAN can remove some of the cloud mist to a certain extent. But there are still problems of incomplete cloud removal in some areas, and the details of the terrain are not completely clear.



(a) With extensive cloud coverage



(b) Featuring concurrent cloud and halo phenomena

Fig.2 Comparison of WHUS2-CR data

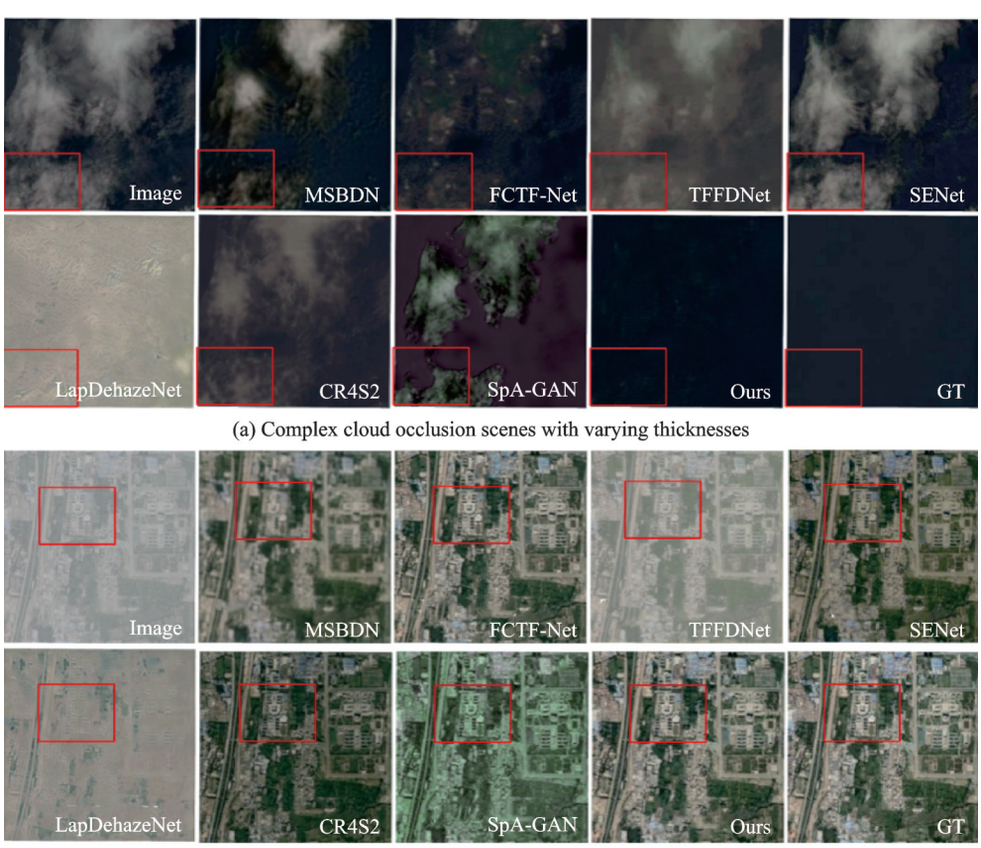
Our proposed PANet method performs exceptionally well, completely removing the thin cloud occlusion and restoring clear terrain details, which is significantly better than other methods. In Fig.2(b), the image has a significant halo phenomenon, which poses a higher challenge for the anti-interference ability and detail recovery ability of the dehazing algorithm. The experimental results show the MSBDN, FCTF-Net, TFFDNet, and SENet methods are not satisfactory in removing the halo, and there are still obvious artifacts in image, affecting the visual effect and the accuracy of subsequent

applications. However, although LapDehazeNet, CR4S2, and SpA-GAN can correct and repair color cast to some extent, the effect is not ideal, and the problem of color distortion is still more prominent. In contrast, our PANet method performs excellently in dealing the halo phenomenon, effectively removing the artifacts and precisely correcting and repairing the color cast, restoring more natural and true image colors. Table 1 presents the quantitative evaluation results of the WHUS-CR dataset. By examining the data in the table, it is evident that our PNANet all other methods across the board. Specifically, PNANet achieves the highest score in SSIM, PSNR, and the lowest in MSE. Compared to other methods, PNAN outperforms them by an average of 5.801% in SSIM, 1.72 dB in PSNR, and a reduction of 8.582 in MSE. These improvements highlight the superior performance of PNANet in preserving image structures, enhancing signal quality, and minimizing errors, thereby demonstrating its effectiveness in the context of the WHUS-CR dataset.

Table 1 Comparison of experimental data evaluation indexes

Dataset	Metric	MSBDN	SpA-GAN	FCTF-Net	TFFDNet	LapDehazeNet	CR4S2	SENet	PNANet
WHUS-CR	SSIM	0.680 0	0.731 1	0.813 0	0.784 0	0.805 5	**0.813 1**	0.781 9	**0.826 5**
	PSNR/dB	24.266	21.650	26.222	24.582	25.842	**26.344**	25.194	**26.593**
	MSE	75.224	84.471	65.507	75.491	68.002	**64.563**	69.212	**62.971**

Fig.3 gives the comparison of RICE data. GT represents the ground truth and the red rectangle box denotes the area of focused attention. Similarly, for the scene with uneven cloud distribution in Fig.3(a), the situation that is more complicated. Most of the other methods cannot effectively remove the cover when processing such images, resulting in the information of the ground objects still being obscured. Especially, the LapDehazeNet method not only fails to successfully remove the thin cloud, but also



(a) Complex cloud occlusion scenes with varying thicknesses

(b) Cloud occlusion with richer detail textures

Fig.3 Comparison of RICE data

leads to the distortion of the ground objects, severely affecting the quality and usability of the image. In this case, only the FCTF-Net and our PNANet method can effectively remove the cloud, and the ground object information is restored to be relatively clear, showing strong robustness and adaptability. For the relatively cloud cover in Fig.3(b), although the overall difficulty is relatively low, it still needs the algorithm to have the good detail recovery ability and the radiation information correction ability. The experimental results show that except for the LapDehazeNet method, which has difficulty in restoring complete ground object information, the other methods can achieve the goal of cloud removal to a certain extent but there are still some problems, such as image blurring, radiation information offset, etc., thus affecting the quality of the image and the accuracy of the follow-up. In comparison, our PANet method has a better performance in restoring ground object details, and the restored details are closer to the real label, which can provide higher quality image for subsequent applications and show its significant advantages in the field of cloud removal in remote sensing images. The quantitative evaluation of the RICE dataset is showcased in Table 2, where it can be seen that our PNANet achieves the best results in all metrics, with an average improvement of 5.625%, 5.363 dB and 24.016 in SSIM, PNSR and MSE, respectively. In Table 3, we demonstrate the performance comparison of different methods in terms of efficiency by evaluating their parameter count (Params), computational complexity (FLOPs), and training duration (t_{train}). Although our method does not achieve the best results, it achieves a good balance between accuracy and efficiency.

Table 2 Comparison of RICE data evaluation indexes

Dataset	Metric	MSBDN	SpA-GAN	FCTF-Net	TFFDNet	LapDehazeNet	CR4S2	SENet	PNANet
RICE	SSIM	0.837 1	0.828 1	0.909 1	0.769 9	0.849 9	**0.911 8**	0.868 5	**0.923 5**
	PSNR/dB	29.670	26.074	**32.773**	21.444	25.349	30.846	27.569	**34.307**
	MSE	50.278	71.248	**37.840**	98.089	77.574	53.191	57.319	**34.010**

Table 3 Contrast experiment efficiency evaluation

Metric	MSBDN	SpA-GAN	FCTF-Net	TFFDNet	LapDehazeNet	CR4S2	SENet	PNANet
FLOPs/10^9	—	55.242	22.607	384.94	104.690	**21.100**	**1.221 1**	46.083
Params/10^6	2.983	28.713	**0.163**	2.653	34.548	1.485	**0.008 4**	2.029
t_{train}/s	308.781	144.562	**94.544**	276.06	185.072	190.603	**16.134**	105.458

2.4 Ablation experiment

In the attention mechanism comparison experiment, we use the U-Net network as the baseline, and add the attention mechanisms of CBAM, ECA-Net, SA-Net and PNA, respectively. Table 4 presents the accuracy after 300 epochs, with our PNA achieving the highest results in terms of both SSIM and PNSR. In terms of efficiency, the proposed method outperforms others in FLOPs, parameter count, and training time. This is because the PNA attention design reduces the number of U-Net channels from 256 to 128. In the ablation experiment, U-Net is used as the baseline, our PNA is the addition of PNA, our L_1 represents the addition of fine thin cloud removal stage on the basis of former, and our L_1-L_p represents the network with added perceptual loss. The experimental results are shown in Table 4. In Fig.4, the attention visualization results of all methods are also shown, and the improvement of the network's attention to the thin cloud coverage area can be clearly observed. In Table 5, we conducted ablation tests on PAN and RDB structures, comparing scenarios where PAN was not used (PAN_0), only scales 1, 3, and 7 were used (PAN_137), PAN_912 (using only scales 9 and 12), RDB_0, RDB_1, RDB_3, and OUR_PAN_RDB. The results show that using RDB generally leads to an upward trend in both SSIM and PSNR metrics compared to not using it. This is because the RDB module's multi-feature re-

use enhances the recovery of fine texture structures. However, as the number of RDBs increases, so does the computational burden of parameter calculations. We observe that when RDB increases to 3, SSIM and PSNR show an negligible improvement. Balancing performance and effectiveness, we therefore employ 2 RDBs for feature recovery. Furthermore, compared to not using PAN, the multi-scale feature attention achieves excellent results in reducing the randomness of thin cloud locations, without significantly increasing parameters or floating-point operations.

Table 4 Ablation evaluation of attention

Metric	U-Net	CBAM	ECA-Net	SA-Net	Our PNA	Our L_1	Our L_1-L_p
SSIM	0.815	0.821	0.821	**0.822**	**0.827**	0.827	**0.841**
PSNR/dB	25.742	26.301	26.239	**26.427**	**26.683 2**	26.722	**27.014**
FLOPs/10^9	**38.882**	38.914	38.934	**38.882**	**38.401**	—	—
Params/10^6	**2.597**	2.602	**2.597**	**2.597**	**1.973**	—	—
t_{train}/s	**62.716**	85.193	72.178	66.675	**58.143 3**	—	—

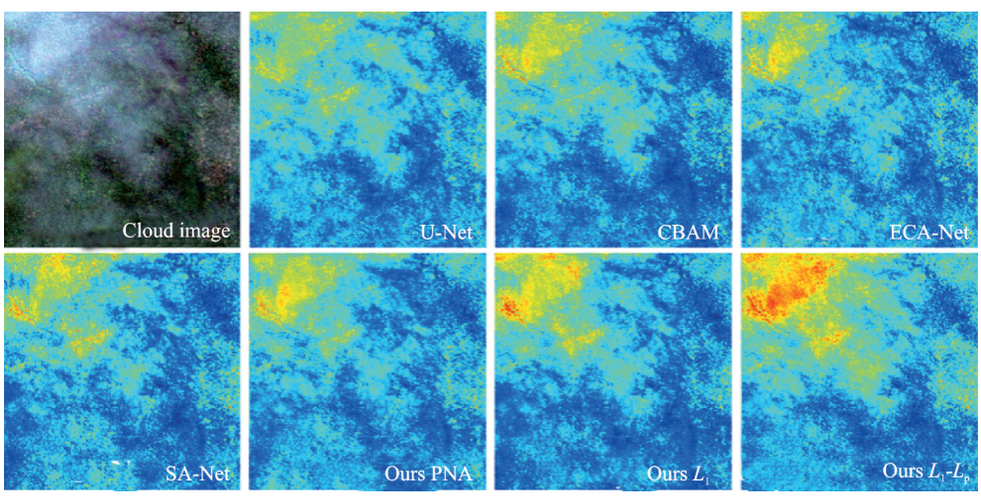


Fig.4 Examples of various attention feature maps

Table 5 Ablation evaluation results of PAN and RDB structure

Parameter	Training time/(s•epoch^{-1})	Params/10^6	FLOPs/10^9	SSIM	PSNR/dB
RDB-0	**42.366**	**1.95**	**16.28**	0.805 25	25.505 7
RDB-1	**48.334**	**2.00**	**18.88**	0.815 37	25.961 7
RDB-3	66.543	2.06	22.08	**0.826 67**	**26.517 3**
PAN_0	53.523	**2.01**	20.18	0.775 84	24.579 9
PAN_137	54.763	2.03	20.48	0.790 62	25.627 6
PAN_916	55.755	2.03	20.48	0.804 68	25.353 3
Ours	56.335	2.03	20.48	**0.828 09**	**26.546 6**

3 Conclusions

In this paper, based on observations of the random distribution and uneven thickness of clouds, we propose a network for removing thin clouds under non-uniform thin cloud conditions. This network adopts a coarse-to-fine thin cloud removal architecture. In the coarse-level de-clouding network, we innovatively incorporate an attention mechanism. By integrating global context with local texture, this mechanism specifically addresses image quality degradation caused by the uncertainty in cloud distribution. During the fine-level de-clouding stage, we focus on the impact of cloud layer thickness on feature detail information. Through a carefully designed residual dense module, we significantly enhance the

extraction and utilization of feature detail information. Through experiments on publicly available datasets, our thin cloud removal network and attention mechanism demonstrate superior performance compared to various existing methods.

References

[1] KING M D, PLATNICK S, MENZEL W P, et al. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites[J]. IEEE Transactions on Geoscience and Remote Sensing, 2013, 51(7): 3826-3852.

[2] LI Y F, CHEN X. A coarse-to-fine two-stage attentive network for haze removal of remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 18(10): 1751-1755.

[3] KANG X D, FEI Z Y, DUAN P H, et al. Fog model-based hyperspectral image defogging[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-12.

[4] LONG C J, LI X H, JING Y H, et al. Bishift networks for thick cloud removal with multitemporal remote sensing images[J]. International Journal of Intelligent Systems, 2023, 2023(1): 9953198.

[5] HE K M, SUN J, TANG X O. Single image haze removal using dark channel prior[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(12): 2341-2353.

[6] GAO Y Y, HU H M, LI B, et al. Detail preserved single image dehazing algorithm based on airlight refinement[J]. IEEE Transactions on Multimedia, 2018, 21(2): 351-362.

[7] ZHU M Z, HE B W, WU Q. Single image dehazing based on dark channel prior and energy minimization[J]. IEEE Signal Processing Letters, 2017, 25(2): 174-178.

[8] JING H, LU N. Multi-scale residual convolutional neural network for haze removal of remote sensing images[J]. Remote Sensing, 2018, 10(6): 945.

[9] LI J, ZHANG Y J, SHENG Q H, et al. Thin cloud removal fusing full spectral and spatial features for Sentinel-2 imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 8759-8775.

[10] MA D Y, WU R Z, XIAO D S, et al. Cloud removal from satellite images using a deep learning model with the cloud-matting method[J]. Remote Sensing, 2023, 15(4): 904.

[11] LIU J, PAN B, SHI Z W. Cascaded memory network for optical remote sensing imagery cloud removal[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-11.

[12] LIU F J, ZHANG F Y, WANG M, et al. Two-level supervised network for small ship target detection in shallow thin cloud-covered optical satellite images[J]. Applied Sciences, 2024, 14(24): 11558.

[13] CAI B L, XU X M, JIA K, et al. DehazeNet: An end-to-end system for single image haze removal[J]. IEEE Transactions on Image Processing, 2016, 25(11): 5187-5198.

[14] REN W Q, PAN J S, ZHANG H, et al. Single image dehazing via multi-scale convolutional neural networks with holistic edges[J]. International Journal of Computer Vision, 2020, 128(1): 240-259.

[15] ZHAO S Y, ZHANG L, SHEN Y, et al. RefineDNet: A weakly supervised refinement framework for single image dehazing[J]. IEEE Transactions on Image Processing, 2021, 30: 3391-3404.

[16] TRAN L A, PARK D C. Distilled pooling transformer encoder for efficient realistic image dehazing[J]. Neural Computing and Applications, 2025, 37(6): 5203-5221.

[17] ZHANG H, PATEL V M. Densely connected pyramid dehazing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2018: 3194-3203.

[18] LI C, YAN W Q, ZHAO H W, et al. TFFD-Net: An effective two-stage mixed feature fusion and detail recovery dehazing network[J]. The Visual Computer, 2025, 41(6): 4001-4016.

[19] DU Y, LI J, SHENG Q H, et al. Dehazing network: Asymmetric unet based on physical model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[20] LI R T, CHEONG L F, TAN R T. Heavy rain image restoration: Integrating physics model and conditional adversarial learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2019: 1633-1642.

[21] YU H, LI C Y, LIU Z H, et al. Remote sensing image dehazing algorithm based on adaptive SLIC[J]. National Remote Sensing Bulletin, 2025, 28(12): 3158-3172.

[22] CHI K, YUAN Y, WANG Q. Trinity-Net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond[J]. IEEE Transactions on Geo-

science and Remote Sensing, 2023, 61: 1-14.

[23] LI J, WANG Y H, SHENG Q H, et al. CloudRuler: Rule-based transformer for cloud removal in Landsat images[J]. Remote Sensing of Environment, 2025, 328: 114913.

[24] DU Y, LI J, SHENG Q, et al Dehazing network: Asymmetric U-Net based on physical model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[25] LI J, WU Z C, HU Z W, et al. Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2A imagery[J]. Remote Sensing, 2021, 13(1): 157.

[26] LIN D Y, XU G L, WANG X K, et al. A remote sensing image dataset for cloud removal[EB/OL]. (2025-03-15). https://arxiv.org/abs/1901.00600.

[27] DONG H, PAN J S, XIANG L, et al. Multi-scale boosted dehazing network with dense feature fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2020: 2157-2167.

[28] XIAO B X, ZHENG Z R, ZHUANG Y L, et al. Single UHD image dehazing via interpretable pyramid network[J]. Signal Processing, 2024, 214: 109225.

[29] LUO Z W, GUSFAFSSON F K, ZHAO Z, et al. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2023: 1680-1691.

[30] WANG S C, LIU J M, ZHONG Y L, et al. A fast saturation based dehazing framework with accelerated convolution and attention block[C]//Proceedings of ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 1-5.

[31] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: IEEE, 2018: 3-19.

[32] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2020: 11534-11542.

[33] ZHANG Q L, YANG Y B. SA-Net: Shuffle attention for deep convolutional neural networks[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speach and Signal Processing.[S.l.]: IEEE, 2021.

Acknowledgements This work was supported by the Fundamental Research Funds for the Central Universities (No.2572025BR14), and the China Energy Digital Intelligence Technology Development (Beijing) Co., Ltd. Science and Technology Innovation Project (No.YA2024001500).

Authors

The first author Mr. **GUAN Wang** is currently a Ph.D. candidate at College of Computer and Control Engineering, Northeast Forestry University, and his research interests include smart forestry, remote sensing change detection and its scene application.

The Corresponding authors Dr. **DU Yang** obtained his Ph.D. degree from Nanjing University of Aeronautics and Astronautics in 2024. He currently serves as an associate professor at College of Computer and Control Engineering, Northeast Forestry University. He has published seven papers in journals such as ISPRS and TGRS. His research interests include remote sensing image cloud and fog removal, image super-resolution reconstruction, multi-source information fusion, and target detection. Mr. **ZOU Yunkun** is the Chairman of Huantian Wisdom Technology Co., Ltd. As a core corporate executive and technical expert, he has led the team to complete the first-phase deployment of 10 satellites for the "Tianfu Constellation" network, with extensive industry experience in commercial aerospace, satellite applications, and aeronautics fields.

Author contributions Dr. **DU Yang** and Mr. **ZOU Yunkun** conceived and designed the study and developed the models. Mr. **GUAN Wang** performed the data analysis, interpreted the results, and drafted the manuscript. Mr. **TIAN Zhenkai**, Mr. **MA Tao**, and Mr. **ZHAO Lingyuan** provided data and model components for the study. Mr. **XIE Shizhe** and Mr. **YAN Jin** contributed to the discussion section and contextualization of the study background. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: SUN Jing)

一种结合金字塔非局部注意力的二段式薄云去除网络

管 旺[1],田振凯[2],马 涛[3],赵凌园[4],谢诗哲[5],严 瑾[6],杜 阳[1],邹云坤[4]

(1.东北林业大学计算机与控制工程学院,哈尔滨 150040,中国;
2.自然资源部第四地形测量队,哈尔滨 150025,中国;
3.自然资源部第二地理信息制图院,哈尔滨 150025,中国;
4.环天智慧科技股份有限公司,眉山 620500,中国;
5.国能数智科技开发(北京)有限公司时空信息分公司,北京 100011,中国;
6.中国地震应急搜救中心,北京 100049,中国)

摘要:在遥感影像中,约67%的数据会受到云层覆盖的影响,这显著增加了影像分类、识别等下游解译任务的难度。为有效解决云分布的随机性与云厚度的不均匀性问题,基于对云雾随机分布及厚度不均的观测,本文提出一种由粗略到精细的二段式薄云去除架构。在粗去云网络中,创新性地引入多尺度注意力机制,即金字塔非局部注意力(Pyramid non-local attention, PNA)机制,通过融合全局上下文与局部细节信息,针对性地解决云分布不确定性导致的影像质量退化问题;在精细化去云阶段,重点关注云厚度对去云效果的影响(主要表现为细节信息不足),通过精心设计的残差密集模块,显著增强特征细节的提取与利用能力。因此,本文方法在粗去云结果的基础上,精准恢复了丢失的局部纹理特征,实现了去云质量的大幅提升。为验证所提网络的性能和关键组件,在公开数据集上进行了全面实验分析,结果表明该方法在多种技术指标上均达到当前最优性能。

关键词:通道注意力;薄云去除网络;金字塔非局部注意力;遥感影像;残差密集连接