Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8292612, 8 pages
https://doi.org/10.1155/2021/8292612

Research Article

SACTNet: Spatial Attention Context Transformation Network for Cloud Removal

Linlin Liu[1] **and Shaohui Hu**[2]

[1]*School of Information and Engineering, China Jiliang University, Hangzhou, 310000 Zhejiang, China*
[2]*School of Computer Science and Technology, Hangzhou Normal University, Hangzhou, 310000 Zhejiang, China*

Correspondence should be addressed to Shaohui Hu; 2012110014@stu.hznu.edu.cn

Received 15 September 2021; Revised 7 October 2021; Accepted 16 October 2021; Published 26 October 2021

Academic Editor: Deepak Kumar Jain

Optical remote sensing image has the advantages of fast information acquisition, short update cycle, and dynamic monitoring. It plays an important role in many earth observation activities, such as ocean monitoring, meteorological observation, land planning, and crop yield investigation. However, in the process of image acquisition, an optical remote sensing system is often disturbed by clouds, resulting in low image clarity or even loss of ground information, affecting the acquisition of feature information and subsequent applications. We propose a spatial attention recurrent neural network model combined with a context transformation network to overcome the challenge of cloud occlusion. This model can obtain the core information in remote sensing images and consider the remote dependencies in the network. Furthermore, the network proposed in this paper has achieved excellent performance on the RICE1 and RICE2 datasets.

1. Introduction

An optical remote sensing system has been applied to many fields in recent years, such as earth resource survey, vegetation classification, and land use. If the remote sensing image is polluted by cloud shadow, the ground information will become sparse or even completely blurred, resulting in the loss of information. Therefore, using auxiliary means to remove cloud shadow from remote sensing images is a very meaningful research direction. According to the degree of cloud cover, it can be divided into thin cloud cover and thick cloud cover.

Lin et al. [1] provided a dataset of remote sensing images called RICE and studied the cloud removal of remote sensing images. This dataset contains two subdatasets RICE1 and RICE2. The RICE1 dataset is collected based on the cloud display function of Google Earth, with a resolution of 512×512. The RICE2 dataset is collected based on Landsat images, mainly using natural colour images and quality images.

Thin cloud removal can obtain a priori knowledge by using typical mathematical models or by using the differences of spectra in different bands [1–5]. Thick cloud is difficult to remove, because it will seriously block the image and cause the overall loss of ground information. [6–10] used generative adversarial networks GAN and RNN, inpainting methods, etc. to scan the pixels near the cloud shadow and then remove the cloud-contaminated area to generate a declouded image [11–19]. Although these methods are complementary, the effect is usually limited, especially for large-area thick cloud removal and complex scene reconstruction. How to remove the thick cloud is still a very valuable and meaningful problem.

There are two main challenges: first, it is difficult to remove thick clouds and preserve the details of the original image because of the serious occlusion of thick clouds, which leads to the hiding of a lot of ground knowledge, and the ground occlusion information below it is completely invalid. Second, the existing network does not have the ability to consider the dependence of the remote relationship. This is because in the CNN model, the number of operations required to calculate the association increases with the distance between the two positions through convolution.

To address these problems, we propose a novel spatial attention context transformation network for image cloud removal named SACTNet. Specifically, it is divided into two basic networks: one is the backbone network used to remove cloud and the other is the transformer network composed of texture content extractors and correlation embedding modules.

Generally speaking, it is very difficult to obtain the information of thick cloud occlusion by using the backbone network, mainly because the network will add all functional nodes to the holistic dependence, thus missing the local key features. In order to gain the spatial domain features of the core in the remote sensing images with thick cloud, we proposed a basic network based on the spatial attention mechanism in order to transform the spatial domain features into an image. The attention mechanism performs spatial transformation through the spatial domain information in the image, so as to gain the key information. Second, in the existing deep learning image-to-cloud algorithms, the relationship of remote dependencies in the network is rarely considered, and these learning remote dependencies are the key challenge. A key factor affecting the ability to learn this dependence is the path length that the forward and backward signals must traverse in the network. By shortening the path of position combination in any input and output sequence, the learning of distance dependence can be made easier. For this reason, we designed the transformer network, which contains contexture extractors and correlation embedding modules to increase the dependency of the network, and the result is shown in Figure 1. In general, the main contributions of this paper are as follows:

(1) An innovative contexture transformer network for image cloud removal named SACTNet is designed to achieve cloud removal in two types of cloud-blocking datasets with thin and thick clouds

(2) SACTNet is the first to apply transformer structure to remote sensing cloud removal. In particular, we design a contexture transformer network with three related modules

(3) The method in this paper has achieved excellent results on the RICE dataset and is better than existing state-of-the-arts

2. Methodology

The structure of SACTNet is shown in Figure 2. SACTNet consists of a contexture transformer and a backbone network. (1) The contexture transformer network is inspired by [20], and it consists of a content extractor and a correlation embedding and soft attention module. The content extractor is introduced to get content feature of ground truth and cloudy image. The correlation embedding module is designed to calculate the correlation between ground truth and input. The distance between the real image and the cloud occluded image can be ensured. Joint feature learning can discover depth features, which can correspond to provide accurate texture features. (2) First of all, from the backbone network, the feature information is obtained by standard convolution and following bottlenet. Finally, it passes spatial attention recurrent neural network modules (SARNN) and concat the output of the transformer. The SARNN module is used to improve the network's attention to special cloud parts. Its principle is to search the focus mapping from the mapping of input elements. The attention map is a two-dimensional matrix; in particular, the value of the element is uninterrupted and represents how much attention needs to be assigned to pixels.

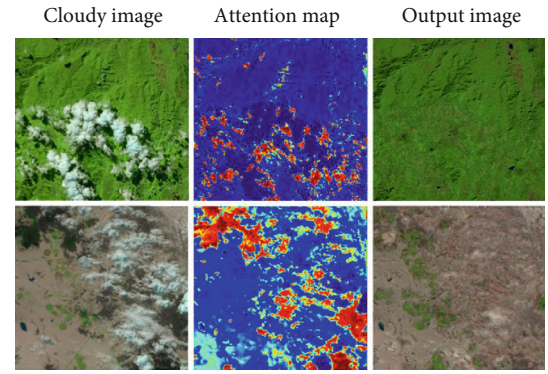

Figure 1: SACTNet results on RICE1 and RICE2.

In the backbone, GAN trains a generator and a discriminator, and the generator is used to generate the discriminator which cannot distinguish whether the data comes from the training sample or the generation. The discriminator is dedicated to learning to distinguish between real data and fake data, or it can be considered two parallel networks in this article. The role of the transformer network is to give more detailed information guidance when the backbone is well trained.

2.1. Contexture Transformer Network

(1) Content extractor: the content extractor of real images is critical in the task of image cloud removal. Declouded images can be produced with the aid of appropriate content texture details. As a learnable content extractor, this article uses the VGG19 pretrained classification model to extract semantic properties, with parameters changed during end-to-end training. This enables the network to perform joint feature learning on real images and cloudy images and can capture more accurate texture features

In this paper, the Pytorch pretraining model VGG19 is used to extract image features and get image embedding. If we pretrain the natural scene ourselves, it will be time-consuming and computational resources, so we can use the pretrained model to apply it to our own tasks. And the pretraining model of VGG19 on the ImageNet dataset has very strong generalization, which can be suitable for pictures in many situations. Simply put, the function of VGG19 is only

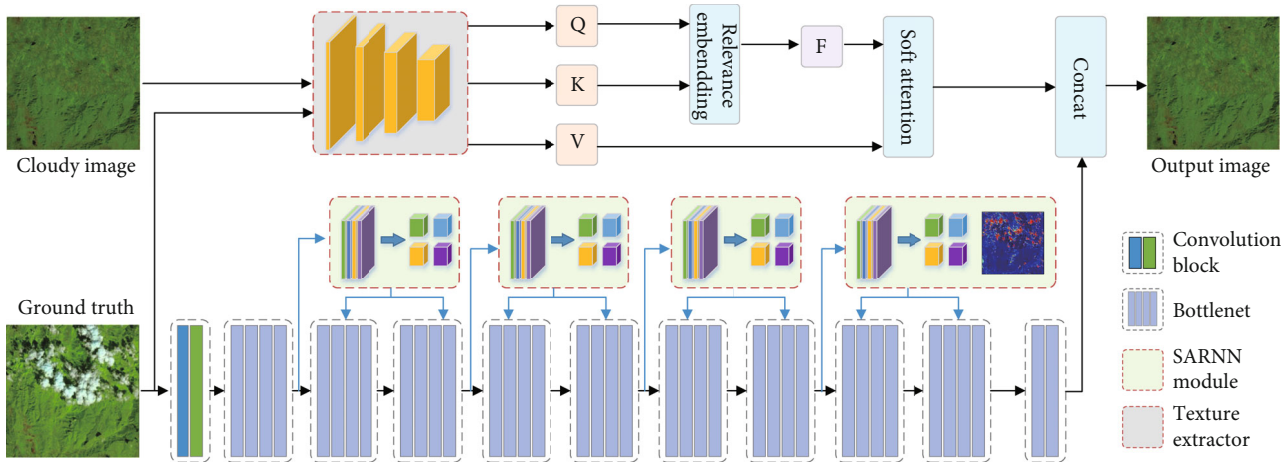

Figure 2: The overview of the proposed SACTNet for image cloud removal.

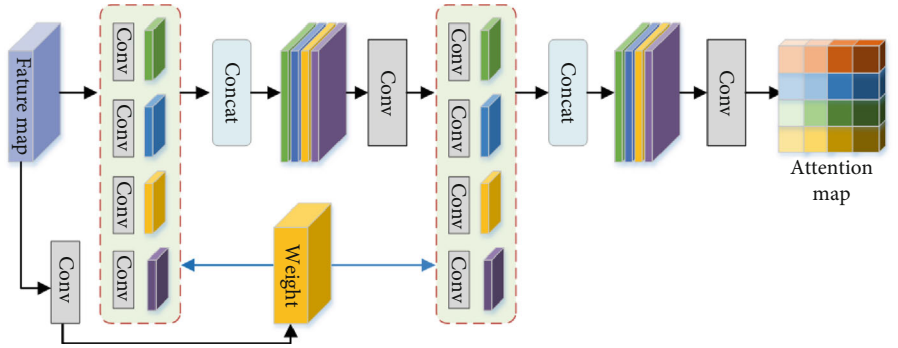

Figure 3: Scheme of the SARNN module.

to extract the feature points of remote sensing images, which can be considered a tool. First of all, we choose the mean shift clustering algorithm as an iterative algorithm, taking the input as the starting point; then, we split the pretrained VGG19 model into three modules, each of which is composed of layers in the pretrained model, and finally, the contexture extractor method can be summarized as follows:

$$\begin{aligned} K &= \mathrm{CE}(\mathrm{GT}), \\ Q &= \mathrm{CE}(GT), \\ V &= \mathrm{CE}(\mathrm{Input}), \end{aligned} \tag{1}$$

where CE() represents the outcome of the content extractor. V, K, and Q indicate two basic units of the attention structure existing in a transformer and will continue to be used in the next relevance embedding module. Q, K, and V are just a way to construct this potential energy function.

(2) Correlation embedding: correlation embedding is to embed the correlation between the real and the cloud image by estimating the similitude between Q and K. The modules expand both Q and K to patches, represented as q_i and k_j. The correlation $R_{i,j}$ between the two parts is calculated by standardized inner product:

$$R_{i,j} = \left\langle \frac{q_i}{\|q_i\|}, \frac{k_j}{\|k_j\|} \right\rangle. \tag{2}$$

(3) Soft attention: this paper presents a soft attention module that uses real images to synthesize texture elements. The related texture quality will be improved throughout the synthesis process. We compute soft attention feature $S_{i,j}$ in the position

Table 1: Parameter setting of model training.

Parameter	Value
Earning rate	0.0004
Small batch	1
Period	200
Optimizer	Adam
β value	0.999

Table 2: Environment configuration of model training.

Environment	Value
Memory	128.00 GB
CPU	Intel(R) Xeon(R) Bronze 3204 CPU @ 1.90 GHz (6-core)
Graphics card	NVIDIA GeForce RTX 3090 (24 G)
System	Windows 10
Environment	Pytorch-GPU 1.8.0
Configuration	Python 3.8.8 + Cuda 11.1 + Cudnn 8.0.5

$r_{i,j}$, which represents the trust of the transferred texture function at each location:

$$s_i = \max_j r_{i,j}. \tag{3}$$

The soft attention map S multiplies these fusion features on the element and adds them back to F to get the texture converter's final performance. This operation can be written as follows:

$$H_{\text{out}} = H1 + \text{Conv}(V) \odot S, \tag{4}$$

where H_{out} represents the output characteristics, $H1$ represents the output of the backbone network, and $\odot$ operators represent element-wise multiplication between feature maps.

2.2. Spatial Attention Mechanism and SARNN. SARNN divides the RNN into two layers. The RNN's first layer will basically spread data throughout the whole picture for aggregation, resulting in more coherent semantic data. This module can form semantic consistency and obtain the context information of distance eigenvalues at the same time. The module would gather local background features in order to obtain the overall perceived eigenvalue, which are critical in cloud image elimination. The technique may also be used to track cloud. In order to create the global perceptual feature map, the RNN's second round accumulates extra nonlocal background knowledge. Direction is the key information in looking for significant clues between shadow/nonshadow parts. This two-round four-way RNN architecture is used to detect shadow regions. The module calculates feature $H_{i,j}$ in the pixel point (i, j). The four directions of the recurrent neural networks are summarized.

$$H_{(i,j)} \longleftarrow \max\left(\alpha_{\text{dir}} H_{(i,j-1)} + H_{(i,j)}, 0\right). \tag{5}$$

The spatial attention model is built based on the two-round and four-direction recurrent neural network structure. The recurrent neural network performs descending projection in four key directions. In Figure 3, the spatial context information is obtained by another branch network to highlight the expected shadow feature information. SARNN can efficiently detect areas impacted by clouds based on the imported cloud image. First, three standard residual blocks are used to obtain features, and then, the feature information assists the following three attention residual blocks to learn negative residuals to eliminate shadows. Finally, multiple residual block units are used to reestablish the generated feature map into a cloud-removed image.

Table 3: Comparison of cloud removal effects on RICE1 and RICE2 images.

	RICE1			RICE2		
	PSNR	SSIM	RMSE	PSNR	SSIM	RMSE
Cloud-GAN [23]	27.656	0.935	4.124	26.5385	0.8758	4.363
CGAN [25]	26.547	0.903	5.323	25.384	0.811	5.213
CycleGAN [26]	25.88	0.893	5.121	23.91	0.793	6.091
SpA-GAN [27]	30.232	0.954	4.765	28.368	0.906	5.857
RSCNet [28]	28.112	0.951	3.212	26.168	0.865	5.451
GCANet [29]	29.568	0.961	3.583	27.322	0.916	6.201
MLA-GAN [30]	31.858	0.971	4.031	31.124	0.938	4.391
Our results	33.251	0.989	5.335	32.851	0.958	6.338

3. Loss Function

In this section, five items are introduced to mainly compose the loss function, which are presented by L_{CGAN}, L_1, L_{perctual}, L_{style}, and $L_{\text{attention}}$, respectively. To start with, the L_{CGAN} is defined:

$$L_{\text{CGAN}}(H, D) = E_{a,b\sim p_{dda}(a,b)}[\log D(a, b)] + E_{a\sim p_{\text{data}}(a), z\sim p_c(c)} \cdot [\log(1 - D(a, H(a, c)))]. \tag{6}$$

With the aim of measuring the accuracy of each reconstructed pixel, the standard L_1 is provided below.

$$L_1(G) = \frac{1}{4HW}\sum_{c=1}^{C}\sum_{v=1}^{H}\sum_{u=1}^{W}\theta_c\left|M_{\text{output}}^{(u,v,c)} - \delta\left(M_{\text{input}}\right)^{(u,v,c)}\right|_1. \tag{7}$$

In this definition, C, H, and W are used to represent the number of channels, the height of the image, and the width of the image. θ_C is the weight of each channel, δ is the calculated value of this model, M_{input} is the input value, and M_{output} is the output value.

Attention loss $L_{\text{attention}}$ is the third item of the whole loss, which can be obtained by the attention map module. The

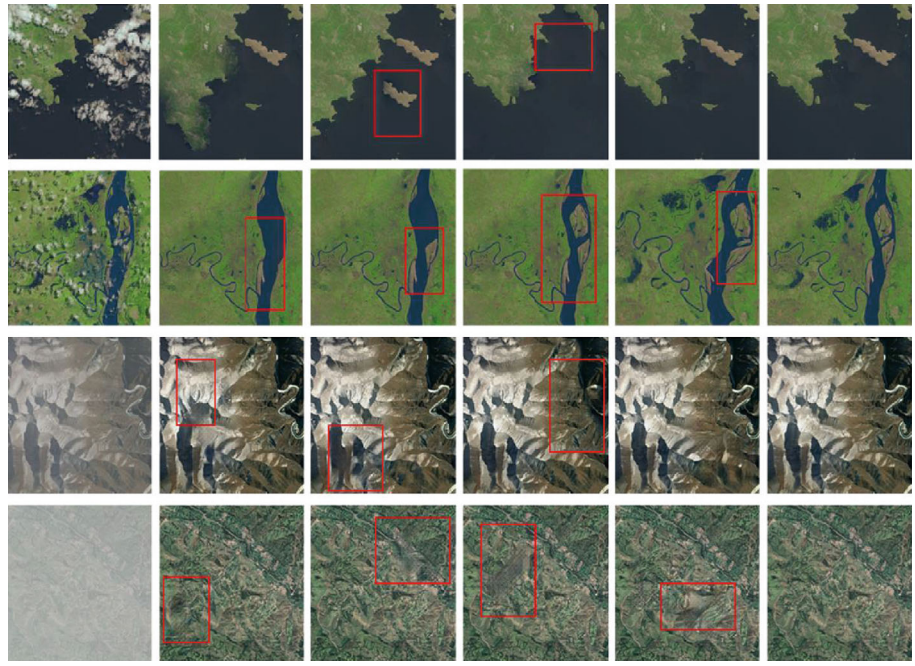

Figure 4: Quantitative comparison of our result with other networks on the RICE dataset.

binary image of the cloud part is processed by the soft attention module to generate matrix A. Matrix B is composed of cloudy images and noncloud images, which is presented in (7).

$$L_{\text{Attention}} = \|A - B\|_2^2. \tag{8}$$

Furthermore, to promote the performance of this novel network, we added the perceptual loss L_{perctual} [21] and the style loss L_{style} [22], which are defined as follows.

$$L_{\text{perctual}} = \mathbb{E}\left[\sum_i \frac{1}{N_i}\left\|\delta_i(Y) - \delta_i\left(\widehat{Y}\right)\right\|_1\right], \tag{9}$$

$$L_{\text{style}} = \mathbb{E}\left[\left\|G^{\delta}\left(\widehat{Y}\right) - G^{\delta}(Y)\right\|_1\right]. \tag{10}$$

In the former equation, δ_i is the feature map of the pretrained model i-th layer. In this research, δ_i is the first layer's feature information of each ReLu module in the ImageNet pretrained VGGNet-19 model. G^{δ} is the reestablished gram metric in the style loss on the foundation of activation map δ_j. Finally, L_{CGAN}, L_1, L_{perctual}, L_{style}, and $L_{\text{attention}}$ are weighted and combined into the total loss by the following formulation.

$$L_{\text{total}} = \theta_{\text{dis}}L_1 + \theta_{\text{adv}}L_{\text{CGAN}} + \theta_{\text{att}}L_{\text{Attention}} + \theta_p L_{\text{perctual}} + \theta_s L_{\text{style}}, \tag{11}$$

where $\theta_{\text{dis}} = 1$, $\theta_{\text{adv}} = 1$, $\theta_p = 0.1$, $\theta_s = 250$, and $\theta_{\text{att}} = 1$.

4. Experiments

4.1. Implementation Details and Configurations. Lin [23] provides an open source dataset of remote sensing images, including RICE1 and RICE2, which can be used for the research of cloud removal of remote sensing images. In the model training stage, the main experimental parameter settings are shown in Table 1, and the hardware environment configuration is shown in Table 2.

4.2. Evaluation Metrics. We used PSNR and SSIM to evaluate the results. Although the effect diagram of cloud removal can intuitively show the effect of the model, it can better carry out quantitative analysis by using indicators. PSNR and SSIM are defined as follows:

$$\text{PSNR} = 10 * \log_{10}\left(\frac{(2^n - 1)^2}{\text{MSE}}\right), \tag{12}$$

$$\text{SSIM}(x, y) = \frac{\left(2\alpha_x\alpha_y + c_1\right)\left(2\beta_{xy} + c_2\right)}{\left(\alpha_x^2 + \alpha_y^2 + c_1\right)\left(\beta_x^2 + \beta_y^2 + c_2\right)}, \tag{13}$$

where α_x and α_y are the means of x and y and β_x, β_y, and β_{xy} represent the variance and covariance of x and y, respectively. c_1 and c_2 are used to maintain stable constants.

4.3. Results and Comparison. For quantitative analysis of cloud removal performance, SSIM, PSNR, and RMSE can be used. Through the above three parameters and the generated map, the cloud removal effect of the image can be evaluated.

Table 4: Ablation experiments of a contexture transformer on RICE1.

CE	Correlation embedding	Soft attention	SSIM	PSNR
			0.921	27.655
	√		0.948	30.248
		√	0.956	30.851
√			0.974	31.322
√	√	√	0.989	33.251

We compare SACTNet with existing research results including Cloud-GAN [24], CGAN [25], Cycle-GAN [26], SpA-GAN [27], RSCNet [28], GCANet [29], and MLA-GAN [30]. In order to make a reasonable comparison with the existing research results, we mainly evaluated the RICE dataset, and the results are shown in Table 3. The performance of SACTNet is better than the existing research, and the images generated by the network have better consistency and balance. In particular, Figure 4 is an intuitive comparison of various network-generated images. The image generated by our proposed network is closer to the ground image than the existing network.

4.4. Ablation Study. In order to further study the performance of the context network, we conducted ablation experiments. The contexture transformer contains three parts: a content extractor (CE) that can be learned, correlation embedding module, and soft attention module. The ablation results are shown in Table 4. We progressively add the content extractor and soft attention once the backbone network is complete. For comparison, we use the VGG network rather than the CE network. PSNR output is effectively increased after soft attention is added, important texture features are enhanced, and less relevant texture features are weakened. PSNR improved to 33.251 after adding the content extractor, proving that embedding joint functionality in the content extractor is superior. We also visually assess the RICE results' quantitative capabilities. For complicated remote sensing images, the combination of these parts will quickly and naturally restore information and achieve practical results.

4.5. Quantitative Study. We use the related research of the article [31] to set the parameters of perception loss and style loss, so the point is on L_{att}, L_{adv}, and L_{dis} and their corresponding parameters. In order to further study the influence of these parameters, this paper makes a quantitative analysis of parameters. Results of analysis are shown in Table 5. Through comparison and analysis, it can be seen that when the values of θ_{att}, θ_{adv}, and θ_{dis} are, respectively, 0.1, 1, and 1, the corresponding PSNR and SSIM results are the highest. It shows that detailed information such as rivers and mountains in the remote sensing image is more complete, and the effect of cloud removal is optimal. At the same time, we conducted an ablation study on the three loss functions. It can be found that after adding L_{adv}, L_{dis}, and L_{att}, the PSNR is increased by 0.02, 0.05, and 0.027, respectively, and the SSIM is increased by 0.01, 0.01, and 0.05, as shown in Figure 5. It is found that just a single loss function does not greatly improve the effect, and there will still be some problems about missing information.

Table 5: Quantitative analysis on parameters.

Parameters	PSMR	SSIM
Baseline	27.655	0.921
Only L_{adv}	27.677	0.922
Only L_{dis}	27.66	0.922
Only L_{att}	27.681	0.926
$\theta_{\text{adv}} = 0.5$	27.911	0.903
$\theta_{\text{adv}} = 1.0$	28.042	0.91
$\theta_{\text{adv}} = 1.5$	27.856	0.907
$\theta_{\text{adv}} = 2.0$	27.924	0.904
$\theta_{\text{dis}} = 0.5$	28.023	0.904
$\theta_{\text{dis}} = 1.0$	28.246	0.909
$\theta_{\text{dis}} = 1.5$	28.011	0.893
$\theta_{\text{dis}} = 2.0$	28.023	0.889
$\theta_{\text{att}} = 0.1$	28.372	0.915
$\theta_{\text{att}} = 0.5$	28.445	0.91
$\theta_{\text{att}} = 1.0$	28.041	0.893
$\theta_{\text{att}} = 2.0$	27.992	0.899

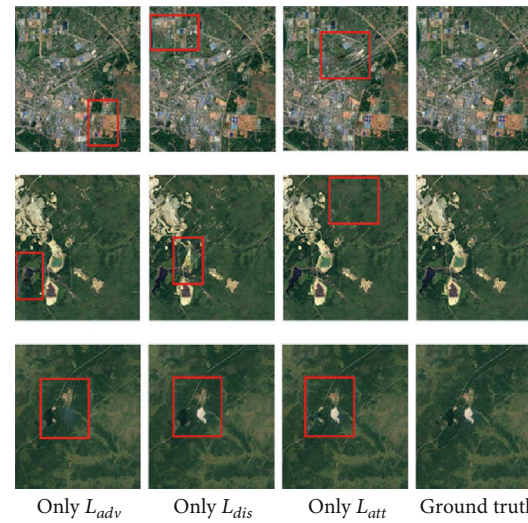

Figure 5: Parameter elimination experiment on the RICE1 dataset.

4.6. Discussion and Future Work. Although our method has achieved excellent results, the network can still be improved regarding subsequent research, such as improving the performance of network cloud removal. By adjusting the structure of the soft attention module, more important texture features can be obtained, and the information of complex remote sensing images can be quickly restored.

At the same time, there are other low-vision tasks similar to cloud occlusion in remote sensing images, such as dust occlusion, snow occlusion, and aerosol occlusion. Therefore,

the generalization effect of the model can be continuously enhanced in the future. By using the same network to achieve the task of removing the cover of multiple objects in remote sensing images, practical performance can be improved and related costs can be reduced. The model proposed in this paper can not only be used for cloud removal of remote sensing images but also be applied to many scenes such as image highlight removal, artifact removal, blur removal, and reflection removal. For example, the model can detect and remove the highlight of the natural image, remove the reflection problem in the flickering image, remove the fuzzy area of the surveillance camera, etc.

5. Conclusions

This paper proposes a spatial attention context transformation network for image cloud removal named SACTNet. We introduce the spatial mechanism in the generative adversarial network, which allows the encoder to learn the prior information on the real images and enhances the reasoning ability of the encoder. In addition, we designed the transformer network, which contains content extractors and correlation embedding modules to increase the dependency of the network.

By shortening the path of position combination in any input and output sequence, the learning of distance dependence can be made easier. In the experimental part, a large number of comparative tests and ablation tests have been completed in the experimental part of this paper. The quantitative and qualitative comparison with the excellent algorithms at the present stage proves the superiority of the method proposed in this paper in handling cloud removal tasks.

By combining L_{adv}, L_{dis}, and L_{att} loss function to form a new loss function, our model has achieved good cloud removal effect. In the quantitative comparison with the existing research results, our model has achieved the best results in PSNR, SSIM, RMSE, and other indicators. Furthermore, our proposed SACTNet achieves excellent results and is superior to the existing state-of-the-arts.

Data Availability

The RICE datasets that support the findings of this study are openly available in GitHub at https://github.com/BUPTLdy/RICE_DATASET [14].

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] C.-H. Lin, P.-H. Tsai, K.-H. Lai, and J.-Y. Chen, "Cloud removal from multitemporal satellite images using information cloning," *IEEE Transactions on Geoscience Remote Sensing*, vol. 51, pp. 232–241, 2013.

[2] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 33, pp. 2341–2353, 2011.

[3] L. Lorenzi, F. Melgani, and G. Mercier, "Inpainting strategies for reconstruction of missing data in VHR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 5, pp. 914–918, 2011.

[4] A. Maalouf, P. Carré, B. Augereau, and C. Fernandez-Maloigne, "A bandelet-based inpainting technique for clouds removal from remotely sensed images," *IEEE Transactions on Geoscience Remote Sensing*, vol. 47, no. 7, pp. 2363–2371, 2009.

[5] M. Xu, X. Jia, and M. Pickering, "Automatic cloud removal for Landsat 8 OLI images using cirrus band," in *IEEE Geoscience and Remote Sensing Symposium*, Quebec City, QC, Canada, 2014.

[6] P. Rakwatin, W. Takeuchi, and Y. Yasuoka, "Restoration of Aqua MODIS band 6 using histogram matching and local least squares fitting," *IEEE Transactions on Geoscience Remote Sensing*, vol. 47, pp. 613–627, 2009.

[7] X. Li, L. Wang, Q. Cheng, P. Wu, W. Gan, and L. Fang, "Cloud removal in remote sensing images using nonnegative matrix factorization and error correction," *ISPRS Journal Photogrammetry Remote Sensing*, vol. 148, pp. 103–113, 2019.

[8] X. Pan, F. Xie, Z. Jiang, and J. Yin, "Haze removal for a single remote sensing image based on deformed haze imaging model," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1806–1810, 2015.

[9] J. Li, Q. Hu, and M. Ai, "Haze and thin cloud removal via sphere model improved dark channel prior," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, pp. 472–476, 2019.

[10] X. Zhu, F. Gao, D. Liu, and J. Chen, "A modified neighbourhood similar pixel interpolator approach for removing thick clouds in Landsat images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, pp. 521–525, 2011.

[11] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," *European Conference on Computer Vision*, vol. 9906, 2016.

[12] W. Ren, L. Ma, J. Zhang et al., "Gated fusion network for single image dehazing," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3253–3261, Salt Lake City, UT, USA, 2018.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, 2015.

[14] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *European Conference on Computer Vision (ECCV)*, pp. 687–704, 2018.

[15] T. Guo, X. Li, V. Cherukuri, and V. Monga, "Dense scene information estimation network for dehazing," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019.

[16] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: a new estimation principle for unnormalized statistical models," in *International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

[17] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 558–567, Long Beach, CA, USA, 2019.

[18] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-haze: a benchmark for image dehazing with dense-haze and haze-free images," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1014–1018, Taipei, Taiwan, 2019.

[19] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, and R. Timofte, "Ntire 2020 challenge on nonhomogeneous dehazing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 490-491, 2020.

[20] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5791–5800, Seattle, WA, USA, 2020.

[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, pp. 694–711, 2016.

[22] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, Las Vegas, NV, USA, 2016.

[23] D. Lin, "A remote sensing image dataset for cloud removal," 2019, arXiv preprint arXiv:1901.00600.

[24] P. Singh and N. Komodakis, "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, Valencia, Spain, 2018.

[25] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, arXiv preprint arXiv:1411.1784.

[26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle consistent adversarial networks," in *IEEE International Conference on Computer Vision*, pp. 2223–2232, Venice, Italy, 2017.

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

[28] W. Li, Y. Li, D. Chen, and J. C.-W. Chan, "Thin cloud removal with residual symmetrical concatenation network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 153, pp. 137–150, 2019.

[29] D. Chen, M. He, Q. Fan et al., "Gated context aggregation network for image dehazing and deraining," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1375–1383, Waikoloa, HI, USA, 2019.

[30] C. Duan and R. Li, "Multi-head linear attention generative adversarial network for thin cloud removal," 2020, arXiv preprint arXiv:2012.10898.

[31] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: structure guided image inpainting using edge prediction," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), 2019.