SI-SA GAN: A Generative Adversarial Network Combined With Spatial Information and Self-Attention for Removing Thin Cloud in Optical Remote Sensing Images

JUNTAO LIU1, WEIMIN HOU1, XIN LUO2,3, JIA SU1, YANLI HOU1, AND ZHENZHOU WANG1

1School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
2Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Huzhou 313001, China
3School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Weimin Hou (hwm@hebust.edu.cn)

This work was supported in part by the Science and Technology Program of Hebei under Grant 21355901D, Grant 20355901D, and Grant 19255901D; and in part by the Science and Technology Program of Sichuan under Grant 2020YFG0240 and Grant 2020YFG0055.

ABSTRACT In agricultural remote sensing monitoring, climate often affects the quality of optical remote sensing image data acquisition. The acquired satellite imagery results usually contain cloud information, leading to a lack of ground data information. Unlike thick clouds, the semi-transparent nature of thin clouds prevents thin clouds from completely obscuring the ground scene. In order to remove thin clouds in the cultivated land and restore the actual ground information as much as possible, we proposed a cloud removal method of spatial information fusion self-attention generative adversarial network (SI-SA GAN) based on multi-directional perceptual attention and self-attention mechanism. The proposed method identifies and focuses on cloud regions using spatial attention, channel attention, and self-attention mechanism, which can enhance image information. The modules of the discriminator utilize residual networks and self-attention non-local neural networks to guide image information output. The generative adversarial network (GAN) is applied to remove clouds and restore the corresponding irregular occlusion area according to the depth characteristics of the input information. A gradient penalty is applied to improve the robustness of the generative network. In this paper, we compared the evaluation indexes of other advanced models. The qualitative and quantitative results of Sentinel-2A and public RICE datasets confirmed that the proposed method could enhance image quality effectively after cloud removal. The model has excellent thin cloud removal performance with small-scale training data.

INDEX TERMS Remote sensing, thin cloud removal, GAN, self-attention, spatial multi-directional perception.

I. INTRODUCTION

Nowadays, remote sensing significantly contributes to monitoring and acquiring information in various fields, and remote sensing imaging technology can periodically obtain objects on the ground in real time. However, the acquired remote sensing images often contain useless interference information, such as clouds caused by climate and weather.

It is difficult to discern the feature information with the interference of the clouds, which is not conducive to the subsequent analysis and processing of remote sensing image data. Thick clouds completely cover the content of the image, and thin clouds can display part of the surface feature information to restore the original information from the image. The interference of these clouds will cause the lack of ground feature information, which in turn will cause inconvenience to later earth resource observation tasks [1], geological disaster investigation [2], and agricultural production [3], [4].
In quantitative remote sensing monitoring of agriculture, various vegetation indices (VI) and leaf area indices (LAI) need to be quantitatively calculated using spectral data in the visible and near-infrared wavelengths. These two parameters are used to study crops’ spatial and temporal variation and compare the correlation between agronomic parameters and VI so that remote sensing inversion models can be established. The crop growth and yield are monitored quantitatively by remote sensing models. The vegetation index is not only related to the surface condition but also the different atmospheric conditions and sensor imaging conditions during the transit of remote sensing satellites can lead to the deviation of vegetation index values in the same area. Thin clouds seriously affect the accuracy of vegetation index calculation. Therefore, cloud-free images are essential for crop growth monitoring and yield prediction [5], [6], [7], [8], [9].

The current methods for thin cloud removal in single images are classified into frequency characteristic-based [10], spatial characteristic-based [11], [12], spectral characteristic-based [13], and image transformation-based methods [14], [15]. The cloud removal method uses the idea of detection and denoising. In the cloud detection from the arable area of the GF-1 satellite image, Xiong et al. [16] combined multiple spectral features with dynamic thresholds to detect the cloud information of the cropland. Haze Optimized Transformation (HOT) method [17] and the dark channel prior (DCP) method [18] both use a priori assumptions to remove clouds and restore images easily and effectively. However, prior knowledge obtained from observational statistics lacks generalizability for remote sensing images. The difference between haze removal and cloud removal is that the distribution of haze interference to information in the image is uniform. In contrast, the distribution of cloud interference to information is uneven and random. The thickness of clouds [19] with different distributions is different, so the process of cloud removal is more complex.

Zhang and Yuan et al. [20] used a deep convolutional neural network (STSCNN) combined with spatial information, time, and spectrum for removing clouds. It showed significant advantages compared with the traditional model. At the same time, they also proposed [21] gap filling by multiple spatiotemporal images to solve the problem of thick cloud coverage in large-scale areas. Compared with the methods of spatial characteristic-based and spectral characteristic-based, the method based on multi-temporal shows significant advantages and strong generalization ability. However, the multi-temporal phase requires multiple groups of historical images and cannot remove the cloud from a single unknown image. This approach will not be suitable if the image changes significantly within a short period or if clean and unobstructed reference data are unavailable within adjacent time.

With the improvement and development of GAN [22], [23], the generative adversarial model is a better method to restore missing information based on learning distributions than other network models. The multi-scale network model GridDehazeNet [24] based on the attention mechanism can alleviate the insufficient recovery of missing information in traditional methods.

GAN to remove the cloud has been widely applied in the cloud removal problem. Singh et al. [25] described using Cycle-Consistent adversarial networks [26] to eliminate thin clouds. Gao [27] used GAN to remove clouds from high-resolution optical and SAR images. Wen et al. [28] used YUV color space combined with the method of GAN for cloud removal processing; and achieved excellent results for thin cloud removal. Overall, it is worthwhile to continue exploring the possibilities of this idea by using GAN to remove thin clouds.

Pan [29] first used the SPA GAN model for cloud removal of remote sensing images, which was first proposed by Wang [30] to remove rain stripes. The SPA GAN model [31] uses a spatial attention network in the generator’s design to better simulate the human visual mechanism and obtain high-quality cloud-free images. Because the model is weak in the generalization ability of actual data, it is necessary to strengthen the cloud removal ability of the model, further improve the accuracy of detail reconstruction and scene recognition, and obtain a stable cloud removal model. The self-attention mechanism is a non-local neural network calculated by the response at a particular location as a weighted sum of features at all locations in the input feature mapping. The classical method assumes a uniform haze distribution in a region, but thin clouds in remote sensing images are a non-uniform medium. The difference from the previous description is that we use multilayer fusion attention combined with GAN networks to capture the probability distribution of cloud thickness, and self-attention is utilized to obtain feature associations between different cloud and non-cloud pixels.

We proposed a spatial information fusion self-attention generative adversarial network (SI-SA GAN). There are three major contributions as described below:

1) We improved the GAN model by enhancing useful feature acquisition with channel attention and spatial four-directional attention and applying self-attention in the discriminator to enable the generative network to focus better on the structural information of the cloud area.

2) We combined the method of the WGAN-GP model [32] to introduce gradient penalty in training and used random interpolation between true and false samples to ensure the stability of the gradient descent during the training process and optimize the overall training process.

3) To make the generator reconstruct the actual ground information, the global attention and local attention information are combined to constitute the mixed loss function. The pixel spatial feature information and self-attention information are used to supervise the model training.

Section II details the proposed method of the SI-SA GAN for cloud removal. Section III provided the experimental results, analysis, and discussion. Section IV discusses the possibility of thick cloud removal. Section V concludes this paper.
II. METHODS

The discriminator in GAN can learn which part of the image is more worthy of attention. In contrast, the generative network does not have this ability because its learning process only maps randomly sampled values to the generated images and cannot identify which regions to focus on. In general, the difference between the two parts of the network is always such that the discriminator is stronger than the generator in the adversarial training process. Spatial information fusion self-attention uses the self-attention mechanism in the discriminator to compute the correlation between feature vectors, which exhibits a better balance between the ability to model long-range dependencies and computational and statistical efficiency. The spatial attention feature map is obtained by image spatial information-aware multidirectional feature extraction, and the fine details of each pixel position in the reconstructed image are fully coordinated with the fine details at the far end of the image. The features of each pixel in the four directions of up, down, left, and right can be fully captured to finally generate a globally consistent scene, which ensures that the generator and the discriminator have the same detail extraction ability and solve the GAN unbalanced adversarial training.

We proposed a spatial information fusion self-attention generative adversarial network (SI-SA GAN) for removing thin clouds. The generative network of this GAN achieves cloud removal by a local-to-global approach and reconstructs the land feature information of cloud-free images. During the training process, the simulation of clouds is computed by thresholding the difference between the cloudy image and the actual cloud-free image. The label combination of the pseudo cloud-free image generated by the generative network and the actual cloud-free image is input to the discriminative network. The discriminative network not only acquires the probability distribution of cloud thickness through a self-attention network to target the cloud areas that affect the image quality; but also achieves global reference of the image to ensure the restoration of surrounding feature information. The spatial attention map generated by the generative network is also used to guide the model to focus on more detailed information of ground objects and reconstruct samples with similar distribution to the actual image data as much as possible. In the condition of the stronger capability of the discriminative network, the generative network is still better to deceive the discriminative network. This process makes the whole network conform to the distribution characteristics of the data and achieves the cloud removal of optical images. The complete structure of SI-SA GAN is shown in Fig. 1.

A. IMPROVED GENERATOR STRUCTURE OF SI-SA GAN

The generative network of SI-SA GAN fuses spatial attention network (SPANet) with channel attention to extract attention feature maps in a multilayer end-to-end manner. It is an extended model of GAN and uses horizontal/vertical neighborhood information in space to model missing information. The SPANet of spatial attention network is shown in Fig. 2. It is a two-round four-direction recurrent neural network with ReLU and identity recurrent neural networks (IRNN) [33]. The network obtains globally perceptive attention feature...
maps from the contextual information of image pixel locations. The attention weight map is divided into four parts, represented as $W_{\text{left}}$, $W_{\text{right}}$, $W_{\text{down}}$, and $W_{\text{up}}$, respectively, and the direction-perception weight map is shared in both rounds.

Fig. 3 shows that the overall generative network is based on the architecture of the residual network. The Squeeze-and-Excitation (SE) block [34] is widely used for denoising and removing rain direction. In our generative network, three residual networks with SE modules are used to add an attention mechanism that automatically acquires the significance of each feature channel, perceives the meaningful features, and discards the useless features according to the degree of importance. The image input to this network is stretched by global pooling and FC layers and then multiplied with the original image to assign weights to each channel.

The spatial attention network generates direction-aware attention features of the cloud information part from the input feature maps. The value of each element of the attention map represents how much cloud information is assigned to the pixel. It indicates the spatial distribution of clouds and is helpful for the subsequent steps of thin cloud information removal. The spatial attention residual block is shown in Fig. 3. Its structural mechanism uses the form of a residual network, which is guided by the attention map for cloud removal processing.

**B. IMPROVED DISCRIMINATOR STRUCTURE OF SI-SA GAN**

At present, self-attention mechanisms are widely used in natural language processing to learn text representations [35]. It was used in this research to capture the internal correlation between cloud information and surrounding ground features.

As shown in Fig. 4, the network passes through three branches of convolution architecture, $f(x)$, $g(x)$, and $h(x)$, respectively; $f(x)$ and $g(x)$ are used to match each other and determine the strength of the connection between each pair of pixels, and the output of $h(x)$ keeps the number of channels constant. The initial attention map $D_{i,j}$ is obtained by Softmax normalization; the $D_{i,j}$ is multiplied with the output of $h(x)$ to reshape the output into a new feature map $s_j$.

$$m_{i,j} = f(x_i)^T g(x_j)$$

$$D_{i,j} = \frac{\exp(m_{i,j})}{\sum_{i=1}^{N} \exp(m_{i,j})}$$

$$s_{i,j} = \sum_{j=1}^{N} D_{i,j} h(x_j)$$

where the $D_{i,j}$ indicates the correlation of the model with region $i$ when judging region $j$. The final output of this
The proposed generator loss function in the SI-SA GAN is as follows:

\[ \gamma_t = \gamma s_t + x_t \]  

(4)

where \( \gamma \) is an adaptive scalar with an initial value of 0.

In Fig. 5, the discriminative network in SI-SA GAN model is composed of ordinary CNN and a self-attention network. The structure of the residual neural network is used in the discriminator to enable the generator and discriminator to optimize each other efficiently. Firstly, the features are extracted through three layers of the convolution network with batch normalization. Then, the self-attention feature map and the correlation matrix \( D_{i,j} \) are output by superimposing two layers of the self-attention network and the convolutional block.

The self-attention mechanism makes the model better capture the global information and generate details from the clues from all feature locations. The input of the discriminator is cloud labels and real-cloud free or false cloud-free images output by the generator. The discriminator can achieve constraints on the image structure with the self-attention mechanism and guide the generative network to focus on the cloud details through the loss function so that the detailed feature characteristics are consistent and coordinated with each other.

**C. DESIGN OF COMBINED LOSS FUNCTION**

The proposed generator loss function in the SI-SA GAN model is designed by combining the \( L_1 \) loss function [29], the similarity loss function [36], and the attention loss function [30] based on the adversarial loss function to further optimize the model training process. The total loss function of the proposed generator is listed as follows:

\[ L_G = L_{GAN}(G, D) + L_{L1}(G) + L_{SSIM} + L_{Att} \]

(5)

In order to generate and reconstruct images practically identical to the original image, we first introduced an adversarial loss \( L_{GAN}(G, D) \), which takes the form of a binary cross-entropy to implement the discriminator and generator game. The equation of the function is as follows:

\[ L_{GAN}(G, D) = -E_{x \sim P_{data}(x)}[\log(1 - D(x, G(x)))] \]

(6)

where \( x \) is the cloudy image, \( E \) is the mathematical expectation, \( D(x, G(x)) \) is the probability of discriminative network judging that the image reconstructed by the generative network after removing the cloud layer information is the reference label image. \( G(x) \) is the prediction result of removing the cloud by the generator, \( p_{data}(x) \) is the actual data \( x \) probability distribution.

The minimum absolute value deviation between the factual cloud-free label and the image reconstructed by the generator is used to measure the accuracy of the model in reconstructing the ground information for each pixel. The formula is defined as follows:

\[ L_{SSIM} = 1 - SSIM(G(x), R) \]

(8)

To better remove thin clouds and make the generator enhance the attention of clouds, the model introduces an attention \( L_{Att} \) loss function is introduced, which is defined as follows:

\[ L_{Att} = \|G_{Att} - M\|_2^2 \]

(9)

The matrix \( G_{Att} \) is a two-dimensional contextual direction-aware feature generated by the spatial attention network (SPAN), and the matrix \( M \) is the cloud binarization label information.

In order to better guide the training of the generator, penalty terms and self-attention loss function are applied to the model in this paper. The definition of the generated adversarial loss function is expressed as:

\[ L_{GAN}_2(G, D) = E_{x \sim P_{data}(x)}[\log(1 - D(x, G(x)))] - E_{x \sim P_{data}(x), R \sim P_{data}(R)}[\log D(x, R)] \]

(10)

To ensure the stability of the training process of GAN. According to the method of WGAN-GP [32], a penalty term...
The penalty term $L_{gp}$ can stabilize the gradient and ensure that $G(z)$ is close to $x$, and the process $D(G(z))$ of the approach will not exceed $D(x)$.

$$L_{gp} = \lambda_{gp}(\|\nabla_{\hat{x}} D_w(\hat{x})\|^2_2 - 1)^2 + \frac{1}{N} \sum_i \log \frac{1}{D_{i,j}(G(x))}$$  \hspace{1cm} (11)

where $\lambda_{gp}$ is the gradient penalty coefficient. In order to solve the spatial sampling problem, the formula $\hat{x} = \varepsilon[x, R] + (1 - \varepsilon)G(x)$ means to interpolate on the actual cloud-free ground image and the pseudo cloudless image data generated by the generative network. $\varepsilon$ is a random number, $\varepsilon \sim U[0, 1]$. $N$ is the amount of data in training. This penalty constraint can ensure that $L_{GAN2}(G,D)$ has Lipschitz continuity, which makes the training more stable. The proposed self-attention loss function applied to this model is as follows:

$$L_{S-Att} = \|D_{SA}(G(x)) - D_{SA}(R)\|^2_2$$  \hspace{1cm} (12)

The $D_{SA}$ represents the final output of the feature map from the self-attended layers. The model focuses not only on the parts with clouds but also on the ground information of other crucial features in the image. These combined loss functions can change the instability of the GAN’s alternating transformations and optimize the generative network’s supervised training by differentiating between attention maps and constraint terms. It uses gradient information to guide the generative distribution toward the actual distribution, so the gradient is not too small, effectively speeding up the convergence speed.

### III. EXPERIMENTS AND RESULTS

Experiments are performed with different datasets to evaluate the proposed method and verify the feasibility and practicality of SI-SA GAN for thin cloud removal. We used a remote sensing image dataset (RICE) proposed by Lin [37] et al. for cloud removal. The RICE dataset contains both types of data: the thick and thin cloud datasets. At the same time, we selected the data acquired from Sentinel-2A as the training object to verify the reliability and generalization effect of the method.

### A. DESCRIPTION OF DATASETS

The RICE1 public thin cloud dataset is derived from the Google Earth platform. It contains 500 thin cloud data samples at $512 \times 512$ resolution. In addition, we choose the
TABLE 1. Quantitative evaluation of reconstructed images from the RICE1 test sets.

| Method                  | PSNR(dB) | mSSIM   | mRMSE  |
|-------------------------|----------|---------|--------|
| Frequency characteristic-based | 15.301   | 0.7507  | 0.01079|
| Grid Dehaze Net         | 26.175   | 0.9204  | 0.00467|
| SPA GAN                 | 30.232   | 0.9546  | 0.00352|
| SI-SA GAN               | 31.391   | 0.9732  | 0.00323|

FIGURE 7. Evaluation index of training results for RICE1: (a) results of FID, (b) results of inception score.

The results in Figure 6 show that the frequency-based characteristic of thin cloud removal is significantly ineffective. Because the ground information obscured by thin clouds is not entirely lost, the frequency-based method can recover some ground information. However, it cannot effectively remove thin clouds in places with more thin clouds.

FIGURE 8. Results of attention map in different area for RICE1: (a) results of area 1, (b) results of area 2, (c) results of area 3.

B. EXPERIMENTAL ENVIRONMENT SETTINGS

The experiments were performed on the Ubuntu 18.04 system with an NVIDIA Tesla K80 GPU and 8 GB of RAM. In the training process of 200 epochs, the hyperparameters of the generator and discriminator were set as $\beta_1 = 0.999$ and $\beta_2 = 0.9$, respectively. We adopted the Adam algorithm to optimize the proposed model. The learning rate is set as 0.0004. The ratio of the training set and validation set is set to 8:2.

C. THE RESULTS OF CLOUD REMOVAL FOR RICE2

The results of the classical Frequency characteristic-based cloud removal method [10], [38], GridDehazeNet [23], SPA GAN, and SI-SA GAN on the RICE1 dataset are shown in Fig. 6. From left to right in Fig. 6 a-f, each column represents the cloud image, the actual cloudless image, the Frequency characteristic-based, the GridDehazeNet result, the SPA GAN result, and the SI-SA GAN result of this research.

We will validate and discuss the ability of the SI-SA GAN to remove thick clouds with the RICE2 dataset in Section IV. The RICE2 dataset contains 736 thick cloud data samples, with the exact resolution as RICE1. These data were derived from Landsat 8, synthesized from three bands (bands 6, 5, and 4).
results generated by GridDehazeNet are relatively poor in definition and brightness, and this method is relatively weak in reconstructing small ground objects. The results produced by the SPA GAN model have deviations from the ground truth in the spectral information. However, the proposed method dramatically retains the ground feature information when reconstructing and generating the cloud-free image, and the ground feature information is not seriously lost. From a visual point of view, each reconstructed image, after cloud removal, retains the land’s correct feature structure and spatial information. Therefore, the SI-SA GAN shows more significant advantages than other algorithms in restoring brightness color information.

The quantitative evaluation results are listed in Table 1, which indicates the average PSNR, SSIM, and RMSE values of the original cloud-free images recovered in the RICE1 test set in different algorithm models. From the results shown in Table 1, the proposed method attained 31.391 dB (PSNR), 0.9732 (mSSIM), and 0.00323 (mRMSE). The results of the SPA GAN and the SI-SA GAN achieved the best performance. In contrast, we can see that the proposed method with a self-attention strategy achieved 31.545 dB (PSNR) and 0.9748 (mSSIM) better results than the SPA GAN method. Meanwhile, we compare the quality of the images generated by SI-SA GAN and SPA GAN by using the evaluation metrics of FID and inception score. It is obvious from Fig. 7 that the SI-SA GAN has a relative advantage in the quality of the generated images after the cloud removal.

Fig. 8 a-c shows the attention maps generated by SI-SA GAN in the process of cloud removal in different regions. In addition to focusing on cloud information, the model also focuses on feature information with apparent features in the original image. The proposed method demonstrates excellence in spectral restoration and physical location construction.
D. THE RESULTS OF CLOUD REMOVAL FOR REAL EXPERIMENT

To verify the applicability of SI-SA GAN to real situations, we obtain images with thin clouds (B4, B3, B2, natural color) from Sentinel-2A. We used different model methods to predict Sentinel-2A data with the training weight of RICE1 in the conditions of unknown actual ground information and without training Sentinel-2A data. We chose the wheat and
corn growing area to test the performance difference between different methods. Fig. 9-b-d represents the global reconstruction results after cloud removal for the GridDehazeNet model, SPA GAN and SI-SA GAN.

We demonstrated the effectiveness of training with multiple attention by comparing the results of the four methods, shown in Fig. 9, where column (a) is the image covered by a thin cloud, column (b) is the result of the Frequency characteristic-based method, column (c) is the results of the GridDehazeNet, column (d) is the results of the SPA GAN, and column (e) is the results of the SI-SA GAN model. The area selected by the red box represents the difference in details of different methods.

In addition, the proposed method still has good cloud removal ability for the new datasets. It not only removes the thin cloud but also restores the feature information of cultivated land to a great extent. Although the traditional Frequency characteristic-based method can compress the brightness range and enhance the contrast of the land image around the cloud, the cloud in the coverage area cannot be removed. GridDehazeNet and SPA GAN are weak in reconstructing the textural detail information of the land. However, SI-SA GAN shows its superiority in texture generation, which means that our method could process the more complicated interference information in remote sensing images. The SPA GAN model and our method could reconstruct the obscured parts. However, GridDehazeNet only generates a result with fuzzy surface features. The SPA GAN model ignores the local cloud removal and cannot completely remove the thin clouds from the new data. The cloud layer is not entirely removed for the GridDehazeNet model and SPA GAN model, and the original cultivated land feature information is seriously lost.

As an example of an agricultural field covered by vegetation, as shown in the second row of Fig.9, the GridDehazeNet model removes the clouds, but the image undergoes global distortion. Although SPA GAN removes the slightly thicker clouds in the region, it produces partial distortions. The distortion level in the images of the proposed method is very low, and the information is almost recovered. We quantitatively evaluated the spectral similarity between cloudy images, actual cloud-free images, and images after model cloud removal with the spectral angle mapper [39]. It uses the calculation of the inverse cosine of the normalized dot product of the spectra to measure the spectral angle between the images after different methods of cloud removal and the actual cloud-free images (reference), where the smaller the angle, the higher the similarity. It can be seen from Fig. 10 that the SI-SA GAN is closest to the actual cloud-free reference image in terms of reflectance ratio index, followed by the SPA GAN. Although the GridDehazeNet model can remove the cloud information, it also leads to severe spectral distortion.

The images of the planting area covered by thin clouds at the ear stage of corn from July to August were selected for testing. The area covered by thin clouds was the corn planting area of Fuyu County, Heilongjiang Province, China. As shown in Fig. 11, column (a) is the cloudy image, column (b) is the VI based on the combination of bands (B8 - B4)/(B8 + B4) corresponding to the cloudy image, and column (c) is the result of the proposed method after cloud removal, column (d) is the VI image after cloud removal.

From the plot of the vegetation index, the green area represents the corn vegetation region at the ear stage, and the blue area represents the non-corn crop region. Because clouds have larger visible reflectance than near-IR reflectance, thin clouds indirectly affect the analysis results by influencing the NDVI values, making it impossible to see crop vegetation growth clearly in precision agriculture and analyze crop information. After removing the cloud by SI-SA GAN, the thin cloud in the corn planting area is completely removed, the vegetation index of the cloudless part is similar to the value after the cloud removal, and the ground feature information of the corn field is restored. The proposed method of removing thin clouds from cultivated land has yielded excellent results.

Further, we compare the results of cloud removal and the result of generating attention maps between the SPA GAN model and the proposed method in real situations. As shown in Fig. 12, column (a) is cloud images. Columns (b1) and (b2) are images generated by SPA GAN, where (b1) is the image after removing the cloud, and (b2) is the attention map generated by SPA GAN. Columns (c1) and (c2) are images generated by the proposed method, where (c1) is the image after removing the cloud, and (c2) is the attention map generated by SI-SA GAN. In Fig. 12, the yellow box is the selected detail part. From the perspective of removing the thin cloud details, SPA GAN mixes cloud layer information with significant ground object background information, which leads to the judgment that it is similar to surrounding ground objects, so that thin clouds cannot be completely removed. The discriminator in SI-SA GAN encourages the generated image to be similar to the cloudless image in data distribution through a self-attention mechanism. Therefore, the reconstructed image looks more real than the image by SPA GAN.

### IV. DISCUSSION

The proposed method for the model structure and training algorithm is validated in the previous section as effective in removing thin translucent clouds in the RICE1 dataset and in the data obtained from Sentinel-2A. In order to further verify the ability of SI-SA GAN to remove thick clouds, we tried...
to process the thick cloud dataset. The experiment results of RICE2 are displayed in Fig. 13.

The RICE2 datasets contain a large number of thick cloud images, and the feature information in the area covered by the cloud is completely lost. The reconstruction of these land feature pixel information cannot require learning from a large amount of similar data. The traditional frequency characteristic-based method is impossible to remove thick clouds. The proposed method can retain more details and consistency for a few sparse thick clouds, making the result visually closest to the ground truth. However, it is hard for any model to recognize the actual ground features obscured by thick clouds in a single optical image without additional information. Therefore, in Fig. 13, the model restores the area covered by thick clouds and the area covered by shadows as blocks of similar color in the adjacent space to recover the actual ground.

Because the details of the ground features in the places occluded by dense, thick clouds cannot be recovered well, most methods cannot recover the occluded information in a single image. Therefore, it is necessary to use multi-temporal combined with prior knowledge to remove thick clouds and restore the original feature information. Table 2 lists the quantitative evaluation results of reconstructed cloud-free images, including PSNR, mSSIM, and RMSE, from the RICE2 test set. Compared with other methods, the SI-SA GAN has obvious superiority in removing thin and a few thick clouds.

![FIGURE 13. Thick cloud removal results of RICE2 with different methods: (a) original real cloudy images, (b) ground truth of the optical, (c) results of the Frequency characteristic-based method, (d) results of the GridDehazeNet, (e) results of the SPA GAN, (f) results of the SI-SA GAN.](image)

V. CONCLUSION

We proposed a method for thin cloud removal based on spatial information fusion self-attention generative adversarial networks (SI-SA GAN), taking advantage of the SPA GAN model and self-attention networks as the basic structure. The generative network identifies cloud areas using contextual feature information of spatial pixels in a multi-attentive manner from local to global. Complex correlations between channels are fitted by Squeeze-and-Excitation to maximize the utilization of detailed features. In addition, the discriminator evaluates the generated images globally and locally with the self-attention mechanism, directing the spatial attention map to focus on cloud information and recovering the actual ground information. The generator retains the original ground truth image information and removes clouds in the attention map update and guidance. A gradient penalty mechanism is added to improve the generalization ability and robustness of the model. The SI-SA GAN validates the effectiveness of the public benchmark for thin cloud removal RICE1 data set and eliminates the thin cloud interference of actual cultivated land. For thin cloud removal, removing the cloud information and compounding the critical information of ground targets provides the possibility to optimize satellite images for future applications of remote sensing images.

In our proposed method, the removal of thin clouds from images and recovering obscured ground scenes are achieved. Overall, the proposed method is superior compared to other methods in quantitative indexes and subjective visualization. We will consider the advantages of multi-temporal prediction and cloud variability to improve thick cloud removal capability. Because of the different types of land use, the generation confrontation network reconstructs the image according to the image color information. For the same image, different land types may cause color mismatch problems (the water
body’s color becomes the land’s color). It needs to be con-
structed for the specific land use type. Furthermore, training
data sets further realize the model’s application value.
The corresponding cloud removal processing is implemented for
images of the same land type in different bands. We have
further optimized the network details to improve generaliza-
tion performance.

REFERENCES

[1] C. Chen, H. Chen, W. Liao, X. Sui, L. Wang, J. Chen, and Y. Chu, “Dynamic
monitoring and analysis of land-use and land-cover change using
landsat multitemporal data in the Zhoushan Archipelago, China,”
IEEE Access, vol. 8, pp. 210360–210369, 2020.

[2] S. Qin, X. Guo, J. Sun, S. Qiao, L. Zhang, J. Yao, Q. Cheng, and Y. Zhang, “Landslide
detection from open satellite imagery using distant domain
transfer learning,” Remote Sens., vol. 13, no. 17, p. 3833, Aug. 2021.

[3] Y. Xie and J. Huang, “Integration of a crop growth model and deep
learning methods to improve satellite-based yield estimation of winter
wheat in Henan Province, China,” Remote Sens., vol. 13, no. 21, p. 4372,
Oct. 2021.

[4] E. I. Mladenova, J. D. Bolten, W. T. Crow, N. Sazib, M. H. Cosh, C. J. Tucker, and C. Reynolds, “Evaluating the operational application of
SMAP for global agricultural drought monitoring,” IEEE J. Sel. Top.-
ics Appl. Earth Observ. Remote Sens., vol. 12, no. 9, pp. 3387–3397,
Jul. 2019.

[5] J. Huang, J. L. Gómez-Dans, H. Huang, H. Ma, Q. Wu, P. E. Lewis, S. Liang, Z. Chen, J.-H. Xue, Y. Wu, F. Zhao, J. Wang, and X. Xie, “Assimilation of remote sensing into crop growth models: Current status and perspectives,” Agricul. Forest Meteorol., vol. 276–277, Oct. 2019,
Art. no. 107669.

[6] J. Huang, H. Ma, F. Sedano, P. Lewis, S. Liang, Q. Wu, W. Su, X. Zhang, and D. Zhu, “Evaluation of regional estimates of winter wheat yield by
assimilating three remotely sensed reflectance datasets into the
coupled WOFOST–PROSAIL model,” Eur. J. Agronomy, vol. 102, pp. 1–13,
Jan. 2019.

[7] J. Huang, F. Sedano, Y. Huang, H. Ma, X. Li, S. Liang, L. Tian, X. Zhang, J. Fan, and W. Wu, “Assimilating a synthetic Kalman filter leaf area index
series into the WOFOST model to improve regional winter wheat yield estimation,” Agric. Forest Meteorol., vol. 216, pp. 188–202, Jan. 2016.

[8] J. Huang, L. Tian, S. Liang, H. Ma, I. Becker-Reshef, Y. Huang, W. Su, X. Zhang, D. Zhu, and W. Wu, “Improving winter wheat yield estimation by
assimilation of the leaf area index from Landsat TM and MODIS data
into the WOFOST model,” Agric. Forest Meteorol., vol. 204, pp. 106–121,
May 2015.

[9] J. Huang, H. Ma, W. Su, X. Zhang, Y. Huang, J. Fan, and W. Wu, “Jointly
assimilating MODIS LAI and ET products into the SWAP model for winter
wheat yield estimation,” IEEE J. Sel. Top. Appl. Earth Observ. Remote
Sens., vol. 8, no. 8, pp. 4060–4071, Mar. 2015.

[10] B. Rasti, J. R. Sveinsson, and M. O. Ulfarsson, “Wavelet-based sparse
reduced-rank regression for hyperspectral image restoration,” IEEE Trans.
Geosci. Remote Sens., vol. 52, no. 10, pp. 6686–6698, Feb. 2014.

[11] R. T. Tan, “Visibility in bad weather from a single image,” in Proc. IEEE
Conf. Comput. Vis. Pattern Recognit., Anchorage, AK, USA, Jun. 2008,
p. 1079–1084.

[12] D. Lin, G. Xu, Y. Wang, X. Sun, and K. Fu, “A remote sensing
model for cloud shadow removal in multifitemporal images,” ISPRS J. Photogramm. Remote Sens., vol. 177, pp. 161–173, Jul. 2021.

[13] Q. Zhang, Q. Yuan, Z. Li, F. Dong, and L. Zhang, “Combined deep prior with
low-rank tensor SVD for thick cloud removal in multitemporal images,”
IEEE Trans. Geosci. Remote Sens., vol. 56, no. 3, pp. 2274–2286,
Feb. 2018.

[14] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, “Thick cloud
and shadow removal in multifitemporal imagery using prospec-
tively spatio–temporal patch group deep learning,” ISPRS J. Photogramm.
Remote Sens., vol. 162, pp. 148–160, Apr. 2020.

[15] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” IEEE
Signal Process. Mag., vol. 35, no. 1, pp. 53–65, Jan. 2018.

[16] J. Dong, R. Yin, X. Sun, Q. Li, Y. Yang, and X. Qin, “Inpainting of remote
sensing SST images with deep convolutional generative adversarial
network,” IEEE Geosci. Remote Sens. Lett., vol. 16, no. 2, pp. 173–177,
Oct. 2018.

[17] X. Liu, Y. Ma, Z. Shi, and J. Chen, “GridDehazeNet: Attention-
based multi-scale network for image dehazing,” in Proc. IEEE/CVF
Int. Conf. Comput. Vis. (ICCV), Long Beach, CA, USA, Oct. 2019,
pp. 7314–7323.

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

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image
translation using cycle-consistent adversarial networks,” in Proc.
IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017,
pp. 2223–2232.

[20] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, “Cloud removal with fusion of
high resolution optical and SAR images using generative adversarial
networks,” Remote Sens., vol. 12, no. 1, p. 191, Jan. 2020.

[21] X. Wen, Z. Pan, Y. Hu, and J. Liu, “Generative adversarial learning in YUV
color space for thin cloud removal on satellite imagery,” Remote
Sens., vol. 13, no. 6, p. 1079, Jun. 2021.

[22] H. Pan, “Cloud removal for remote sensing imagery via spatial attention
generative adversarial network,” 2020, arXiv:2009.13015.

[23] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. H. Lai, “Spatial
attentive single-image deraining with a high quality real rain dataset,”
in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR),
Long Beach, CA, USA, Jun. 2019, pp. 12270–12279.

[24] H. Li, M. M. Aliahmad, D. Dong, and R. B. Chinamm, “SPA-GAN:
Spatial attention GAN for image-to-image translation,” IEEE Trans.
Multimedia, vol. 23, pp. 391–401, Feb. 2020.

[25] I. Guilrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of Wasserstein GANs,” in Proc. Adv. Neural Inf.
Process. Syst., vol. 30, 2017, pp. 1–11.

[26] S. Bell, C. L. Zinnick, K. Bala, and R. Girshick, “Inside-outside net: Detect-
ing objects in context with skip pooling and recurrent neural networks,”
in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 2874–2883.

[27] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc.
IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA,
Jun. 2018, pp. 7132–7141.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,”
in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City,
UT, USA, Jun. 2018, pp. 7294–7303.

[29] B. M. Mathisen, A. Aamodt, K. Bach, and H. Langseth, “Learning simi-
rity measures from data,” Prog. Artif. Intell., vol. 9, no. 2, pp. 129–143,
Oct. 2020.

[30] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, “A remote sensing
image dataset for cloud removal,” 2019, arXiv:1901.00600.

[31] M. Wan and X. Li, “Removing thin cloud on single remote sensing image
based on SWF,” in Proc. IEEE Int. Conf. On-line Anal. Comput. Sci.
(ICONCS), Chongqing, China, May 2016, pp. 397–400.
J. Liu et al.: SI-SA GAN: A GAN Combined With Spatial Information and Self-Attention for Removing Thin Cloud

[39] R. Y. uhas, A. Goetz, and J. Boardman, “Discrimination among semi-arid landscape endmembers using the spectral angle mapper(SAM) algorithm,” in Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop, vol. 1, 1992, pp. 147–149.

JUNTAO LIU received the B.S. degree in communication engineering from the Hebei University of Science and Technology, Shijiazhuang, China, in 2020, where he is currently pursuing the M.S. degree.

His research interests include remote sensing image processing, digital image processing, and deep learning.

WEIMIN HOU received the B.S. degree in electrical engineering from Chongqing University, Chongqing, China, in 1994, and the Ph.D. degree in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2007.

He is currently a Professor with the School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, China. He is the author of a book and more than 20 articles. His research interests include environment remote sensing, image processing and application, array signal processing, wireless communications, and artificial intelligence.

Dr. Hou is a Senior Member of the Chinese Institute of Electronics.

XIN LUO was born in Chengdu, Sichuan, China, in 1977. She received the B.S. degree in electrical technology, the M.S. degree in power electronics and electric drive, and the Ph.D. degree in pattern recognition and intelligent systems from the Northwestern Polytechnical University, Xi’an, China, in 1999, 2004, and 2008, respectively.

From 2008 to 2012, she was a Lecturer at the School of Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. Since 2012, she has been an Associate Professor with the School of Resources and Environment, UESTC. From 2014 to 2015, she was a Visiting Scholar at the Department of Mathematics, University of California, Santa Barbara, Santa Barbara, CA, USA. She is an expert of Sichuan Branch of Chinese National Remote Sensing Center. She is the author of more than 20 articles, more than ten inventions, and holds a patent. Her research interests include remote sensing image processing and application, machine learning, pattern recognition, computer vision, and 3-D modeling.