Remote Sensing of Environment 271 (2022) 112902

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Attention mechanism-based generative adversarial networks for cloud removal in Landsat images

Meng Xu [a,b], Furong Deng [b], Sen Jia [a,b,*], Xiuping Jia [c], Antonio J. Plaza [d]

[a] Key Laboratory for Geo-Environmental Monitoring of Coastal Zone of the Ministry of Natural Resources & Guangdong Key Laboratory of Urban Informatics, Shenzhen University, Shenzhen 518060, China
[b] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[c] School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia
[d] Hyperspectral Computing Laboratory (HyperComp), Department of Computer Technology and Communications, Escuela Politécnica de Cáceres, University of Extremadura, Cáceres E-10003, Spain

ARTICLE INFO

Editor: Jing M. Chen

Keywords:
Cloud removal
Landsat images
Attention mechanism
Generative adversarial networks (GANs)

ABSTRACT

The existence of clouds affects the quality of optical remote sensing images. Cloud removal is an important preprocessing procedure to effectively improve the utilization of optical remote sensing images. Thin clouds partly obscure the land surfaces beneath them, making it possible to correct the cloudy scenes according to the available information. In this research, we introduce the attention mechanism-based generative adversarial networks for cloud removal (AMGAN-CR) method for Landsat images. First, attention maps of the input cloudy images are generated to extract the cloud distributions and features through an attentive recurrent network. Second, clouds are removed by an attentive residual network under the guidance of the attention maps. Finally, the generated feature maps are fed to a reconstruction network to restore the final cloud-free images. The networks are trained by cloudy and cloud-free Landsat image pairs, and the cloudy images are tested to validate the effectiveness of AMGAN-CR. Both simulated and real cloud experimental results show that the proposed method is more outstanding than the other five state-of-the-art traditional and deep learning methods in removing cloud.

1. Introduction

Because of the progress of remote sensing technology and the upgrading of hardware equipment, optical remote sensing images with higher spatial and spectral resolution are now available. The sensors onboard satellites, airplanes or other airborne systems collect electromagnetic radiation signals of ground objects. These data can be used in Earth observation applications, such as resource detection, wetland resource monitoring, vegetation management, atmospheric environment monitoring and disaster monitoring (Kennedy et al., 2007; Mueller et al., 2016; Inglada et al., 2017). Unfortunately, affected by external factors such as climate and environment, the acquired remote sensing images are often obscured by clouds, which affects the interpretation accuracy of remote sensing images and the interpretation of target features. According to statistics, 35% of the Earth's surface is covered by clouds in a year (Ju and Roy, 2008). When solar radiation passes through the atmosphere, it will be affected by scattering, reflection, and absorption of the atmosphere or clouds. The electromagnetic wave received by the remote sensing satellite sensor will have a certain degree of loss. In this case, the information contaminated by clouds is not conducive to subsequent image processing and interpretation, and cloud contamination greatly affects the availability of images for further research.

To solve the problem that large amount of remote sensing images are covered by clouds, scholars have proposed different methods. These methods can be classified into two categories: traditional methods and deep learning methods. Traditional cloud removal methods take advantage of the spatial or spectral features of clouds to correct or reconstruct cloudy images. Before the rise of deep learning techniques, traditional single-layer methods were mainly developed to address the cloud coverage problem in the last few decades. Traditional methods can be categorized into multi-spectral and multi-temporal algorithms

* Corresponding author at: Key Laboratory for Geo-Environmental Monitoring of Coastal Zone of the Ministry of Natural Resources & Guangdong Key Laboratory of Urban Informatics, Shenzhen University, Shenzhen 518060, China.
E-mail address: senjia@szu.edu.cn (S. Jia).

https://doi.org/10.1016/j.rse.2022.112902
Received 30 August 2021; Received in revised form 9 December 2021; Accepted 8 January 2022
Available online 22 January 2022

Table 1
Summary of advantages and disadvantages of cloud removal methods.

Category	Method/Characteristic	Advantage	Disadvantage/Limitation	Reference
Multispectral	Inpainting	Synthesizes the cloudy areas by propagating the surrounding pixels	Can not fill the large cloudy areas	Maalouf et al. (2009); Cheng et al. (2014); Li et al. (2019b)
	Signal-to-noise ratio	Do not require manual intervention	Is influenced by bright objects and cloud shadow	Xu et al. (2019)
	Linear correlation	Preserves the spectral characteristics of cloudless areas.	Assumes a linear relationship among visible bands covered by thin clouds	Zhang et al. (2002); Lv et al. (2016); Chen et al. (2016); Hong and Zhang, 2018
	Spetral unmixing	Does not require meteorological data and reference images	Treats cloud as an endmember and select the number of endmembers	Xu et al. (2015)
Multitemporal	Information cloning	Recovers the large and heterogeneous landscapes covered by clouds	Can not reconstruct accurately when the land cover changes significantly in a short period of time	Lin et al. (2012, 2013)
	Sparsity decomposition	Is effective for different sensors, numbers of spectral bands and temporal acquisition	Parameter setting and time consuming	Chen et al. (2019a)
	Matrix completion	Recover heavily cloud-contaminated regions	The temporal differences between multitemporal images can not be large	Wang et al. (2016)
Deep learning	Multiscale feature fusion	Deep feature extraction of images	Massive GPU memory consumption and time consuming	Qin et al. (2018); Meraner et al. (2020); Gao et al. (2020)
	Cascade convolutional neural network	Is directly extended to multi or hyper-spectral, medium- or low-resolution remote sensing images	Highly dependent on the performance of cloud detection	Ji et al. (2020)
	Convolution neural network	Suitable for thick clouds, thin clouds, and cloud shadows	Can not be applied to multitemporal images with significant land cover changes	Zhang et al. (2018)
	Encoder-decoder network	Converges fast and attain a higher-quality local optimum	Does not consider perception information	Mao et al. (2016)
	Residual learning	Preserves the original information, speed up the training and boost the denoising performance	Fails to provide a detailed and fully accurate reconstruction when the scene is complex and the cloud cover is thick	Zhang et al. (2017); Li et al. (2019a)
	Generative adversarial network	No need for paired cloudy/cloud-free training images	Completely fails to recover images with too much cloud coverage	Singh and Komodakis (2018); Li et al. (2020)

depending on how each method uses cloud-free reference images. Among them, the multispectral method is used to process thin clouds that do not completely cover ground objects, and the multitemporal method is generally used to restore the image covered by thick clouds. Zhang et al. (2002) created a haze optimized transformation (HOT) approach to detect and describe the spatial distribution of haze or clouds in Landsat images. Several other researchers (Chen et al., 2016; Hong and Zhang, 2018) have followed and extended the HOT method to execute thin cloud or haze removal. Xu et al. (2015) introduced a novel approach based on signal transmission and spectral unmixing, which considers the reflection, transmission, and absorption of thin clouds. Furthermore, Xu et al. (2019) established an excellent thin cloud removal approach using a noise-adjusted principal components transform model. Lv et al. (2016) proposed an algorithm based on empirical analysis and radiation transmission model (RTM) to eliminate thin clouds in the visible bands. The histogram of the image corrected by the algorithm completely overlaps the histogram curve of the reference image, which confirms the effectiveness of the algorithm. Zhou and Wang (2019) improved the method on the basis of Lv et al. (2016) by using band 9 as supplementary information and combining band 9 with RTM-based algorithm to remove thin clouds. Tedlek et al. (2018) used the K-means clustering method to divide cloud-free images into several homogeneous regions and employed the level set method to determine the cloud thickness level in each pixel. Subsequently, the authors reconstructed the areas under thin clouds. Several multispectral methods synthesize contaminated cloudy areas and can be categorized as inpainting methods. For example, the geometric flow curves of distinct regions of the image were found in Maalouf et al. (2009) by employing the bandelet transform with multi-scale grouping. After accurately representing this geometric shape, the information contained within the cloud contaminated area was synthesized by propagating the geometrical flow curves in this area. Some other similar multispectral methods to remove clouds can be found in Cheng et al. (2014) and Li et al. (2019b).

Multitemporal methods require additional clear images as auxiliary data to reconstruct the cloud-covered ground. Lin et al. (2012) detected clouds and cloud shadows in the input image using a semi-automatic cloud detection method. The assessed image quality was based on the structural similarity index, and finally, information cloning was used to fill the cloudy areas. Lin et al. (2013) has made improvements on the basis of Lin et al. (2012), adding processing procedures for image intensity normalization, multitemporal image segmentation and seam determination. Chen et al. (2019a) decomposed cloud-contaminated images into low-rank clean image components and sparse components and detected cloudy and shadowy regions by the sparse components. Finally, the cloud and shadow detection results were used to guide the information compensation of the target image. Xu et al. (2016) presented a multitemporal dictionary learning-based cloud removal methodology. The coefficients in the reference image were combined with the dictionary learned in the target image to execute the removal procedure. Wang et al. (2016) used the temporally contiguous robust matrix completion to restore the missing scenes from clean regions.

In recent years, deep learning techniques have demonstrated significant benefits in computer vision and image processing and have been applied in noise removal, target identification, and image classification fields. Accordingly, cloud removal by deep learning models has been expected and investigated. Mao et al. (2016) developed a network that is composed of multiple layers of convolution and deconvolution operators, and skip connection is added to increase the efficiency of image restoration. In Zhang et al. (2017), residual blocks are utilized to better eliminate the distortions induced by additive noise. The approach provided in Qin et al. (2018) depicted thin clouds as haze covers in each band and utilized a multiscale dehazing convolutional neural network (CNN) to remove clouds. Li et al. (2019a) presented an end-to-end residual symmetrical concatenation network (RSC-Net) for thin cloud removal that estimates the cloudless result straight from the cloud

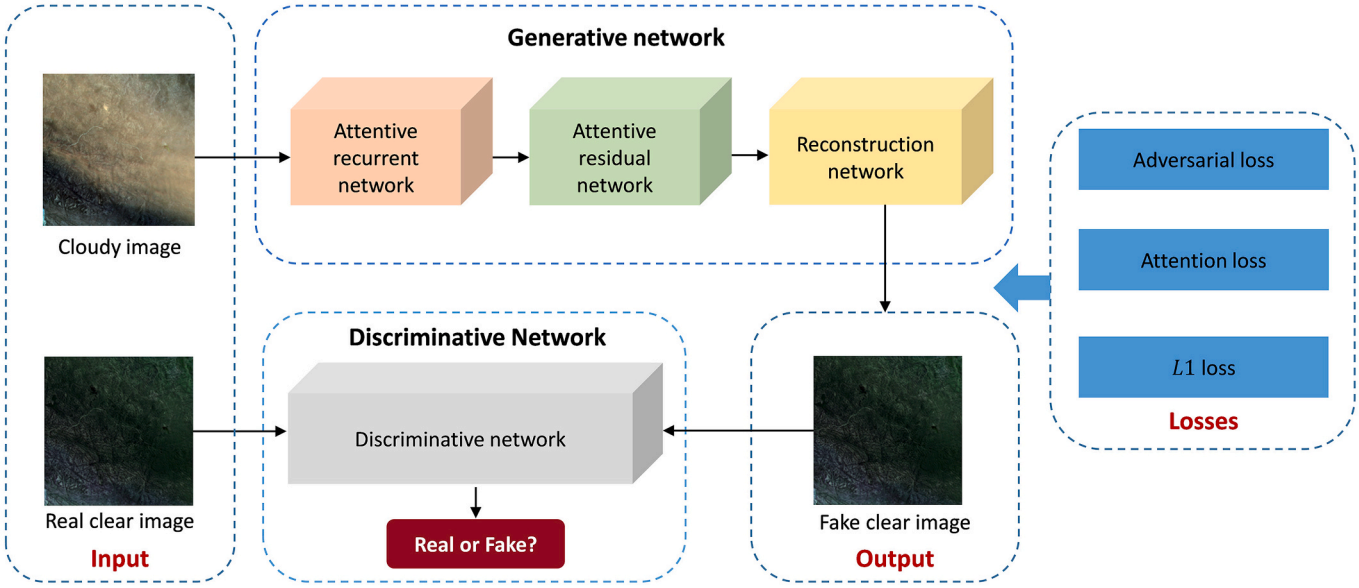

Fig. 1. The network structure of the attention mechanism-based generative adversarial networks for cloud removal (AMGAN-CR) method.

contaminated image. A self-trained multiscale full convolutional network (FCN) for cloud removal from bitemporal images was designed in Ji et al. (2020). Simultaneously, a deep residual neural network architecture was built in Meraner et al. (2020) to remove clouds from multispectral Sentinel-2 images. Multisensor data are also used as additional data in many methods. SAR-optical data fusion is used to guide image reconstruction by using synergistic characteristics. Chen et al. (2019b) suggested a CNN architecture for detecting thick clouds. In addition, cloudy ZY-3 satellite images were removed using the content generation network, texture generation network, and spectrum generation network. Li et al. (2019b) proposed a convolutional-mapping-deconvolutional network for cloud removal with optical and SAR data. Convolutional layers were used for encoding, mapping layers for feature transfer, and deconvolutional layers for decoding. Adding spectral information to the reconstruction of missing data provides another solution. In Gao et al. (2020), SAR images were converted into simulated optical images using a special convolutional neural network. The missing areas were subsequently reconstructed using SAR data, simulated optical images, and actual optical images affected by clouds, yielding a cloud-free output with correct spectral accuracy and high-frequency texture. Table 1 summarizes the advantages and limitations of these techniques briefly.

Recently, generative adversarial networks (GANs) (Goodfellow et al., 2014) have been discovered to be effective in various fields, including removing clouds from remote sensing images. For example, Enomoto et al. (2017) exploited visible light to remove clouds from cloudy images by applying conditional generative adversarial networks (cGANs) (Mirza and Osindero, 2014) to multispectral images. Subsequently, Singh and Komodakis (2018) proposed a Cloud-GAN to map the relations between cloudy images and cloud-free images. A semisupervised technique for thin cloud removal using unpaired images from various areas based on GANs and a physical model of cloud distortion was presented in Li et al. (2020). In Pan (2020), the authors combined the generative adversarial networks and the spatial attention mechanism, which can improve the ability of recovering cloudy areas and generating cloud-free images with higher qualities.

Multispectral methods design specific feature extractors based on different cloud thicknesses or uses spectral relationships between different bands for inference. They have limited generalization abilities. For the multitemporal methods they assume the landcover changes are small between the multitemporal images processed. The difficulty with traditional methods is that it is vital to select which features are important for each individual task. Contrastingly, cloud removal methods based on deep learning can extract multilevel and multiscale features of clouds and are more robust, and provide the idea of end-to-end learning, in which the machine is fed by massive amounts of data labeled with high quality. End-to-end learning is a type of deep learning process where all the parameters are trained together rather than one by one. In end-to-end learning, we can use a single machine learning algorithm rather than several individual components to achieve more effective performance. It has been demonstrated that deep learning methods perform better than traditional algorithms when it comes to complex problems, yet with trade-offs with regard to computing requirements and time. However, the existing deep learning methods for cloud removal mainly deal with RGB images and do not make full use of multispectral and multitemporal characteristics of cloudy images. Most of them require accurate cloud masks as the basis so that cloud detection becomes the significant preprocessing step and has a great impact on cloud removal results.

Considering the merits and demerits of the traditional and deep learning methods and inspired by the good performance of GANs in other computer vision tasks, in this paper, we develop attention mechanism-based GANs for cloud removal (AMGAN-CR) capable of addressing severe cloud cover problems in optical remote sensing images and enormously increase the accessibility of useful data. First, an attention module for cloud removal is created and embedded in a generative adversarial network. We find that attention mechanism can capture the possible distribution of cloud thickness. The attention map of cloud cover is generated to feed into the generative networks so that the generative network can pay greater attention to the structural information of the cloudy areas and the surrounding areas. The attention loss is used to calculate the degree of similarity between the attention map and the cloud mask, which improves the generalization ability of the networks. Second, the training and testing data are derived from Landsat 8 Operational Land Imager (OLI). The training process is accomplished in paired cloudy-clear images that are selected from the nearest date to avoid significant surface changes between them. The AMGAN-CR method is compared with five baseline advanced methods. Finally, we design ablation experiments to assess the parameters of the networks and the effects of the cloud mask on removal. Ablation experiment intends to assess how one variable affects the model performance while holding the others unchanged. The results of both

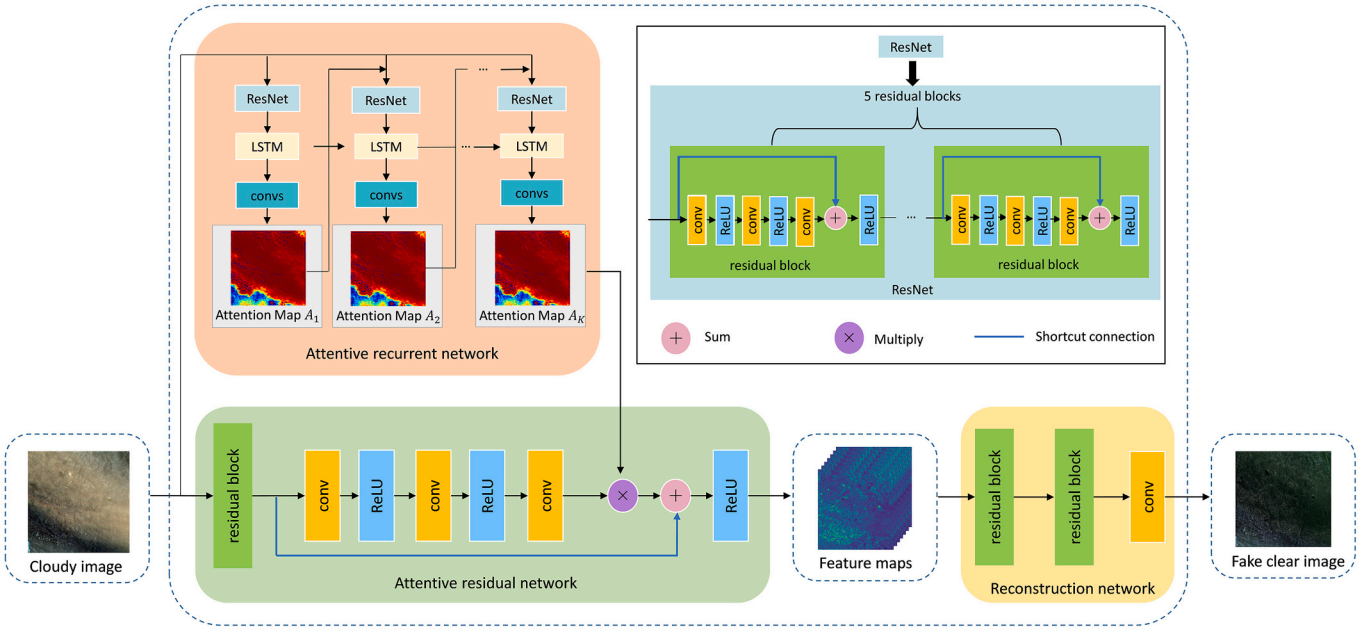

Fig. 2. The generative network of AMGAN-CR.

simulated and real experiments demonstrate the superior performance of AMGAN-CR quantitatively and qualitatively. AMGAN-CR does not rely on the performance of cloud detection as an attention module is embedded to generate the spatial attention map showing cloud distributions. Seven multispectral bands are taken as the input to the network for cloud removal simultaneously, which can fully exploit the cloud features of multichannel remotely sensed images.

The rest of this paper is arranged as follows: Section 2 introduces the proposed AMGAN-CR method, which mainly includes generator, discriminator and loss function. Landsat data preprocessing, dataset preparation, and experimental evaluation matrices are covered under Section 3. Experiments on simulated and real datasets, ablation experiments and outcomes of AMGAN-CR and five baseline cloud removal approaches are shown in Section 4. At last, Section 5 and 6 describe the discussion and conclusion, respectively.

2. Method

2.1. The GAN architecture for cloud removal

The whole architecture of our proposed AMGAN-CR is built on GANs, which are made up of two major components: the generative network (generator) and the discriminative network (discriminator). Suppose the generator is denoted as P and the discriminator is denoted as Q. The GANs can be described as a minimax issue in the following way:

$$\min_{P}\max_{Q} V(Q,P) = \mathbb{E}_{\boldsymbol{u}\sim d(\boldsymbol{u})}[log Q(\boldsymbol{u})] + \mathbb{E}_{\boldsymbol{j}\sim d(\boldsymbol{j})}[log(1 - Q(P(\boldsymbol{j})))], \quad (1)$$

where $\boldsymbol{u}$ represents the given input cloudy image, $d(\boldsymbol{u})$ is denoted as distribution of $\boldsymbol{u}$, $\boldsymbol{j}$ refers to the random noise data, $d(\boldsymbol{j})$ is denoted as distribution of $\boldsymbol{j}$, and the result of the discriminator is represented by $Q(\boldsymbol{u})$, which indicates the probability of $\boldsymbol{u}$ being a true sample. $P(\boldsymbol{j})$ is trained to minimize $\log(1 - Q(P(\boldsymbol{j})))$ or, alternatively, maximize $Q(P(\boldsymbol{j}))$. Additionally, for each $\boldsymbol{u}$, we wish $Q(\boldsymbol{u})$ to be maximized. In this way, the function $V(Q,P)$ plays a minimax game by training P and Q simultaneously.

The idea of a two-player minimax game inspires us to take the GAN architecture into consideration to remove clouds and propose the AMGAN-CR method. As seen in Fig. 1, the generative network contains three networks: an attentive recurrent network, an attentive residual network, and a reconstruction network. The details of these three networks will be described in the next section. The generative network can learn the data distribution from a cloudy image, and the discriminator is used to evaluate the possibility that an image is a real clear image rather than a generated clear image from the generative network. The generator-discriminator game eventually approaches Nash equilibrium (Nash, 1950), which is a solution concept in game theory. When each player has chosen a strategy and no player can do better by changing their strategy. The goal of the generator is to yield fake clear images that the discriminator cannot recognize so that the generated images are as realistic as possible; the goal of the discriminator is to distinguish between actual and fake clear images as accurately as possible. The loss used in the networks is the total of the adversarial loss $\mathscr{L}_{\text{cGAN}}$, attention loss $\mathscr{L}_{\text{Att}}$ and $\mathscr{L}_{\text{L1}}$ loss, which will be given in detail in Section 2.4.

2.2. Generative network

The generative network is used to generate fake clear images by learning the features of cloudy areas. As shown in Fig. 2, the first component of the generator, i.e., the attentive recurrent network, tries to detect the cloudy regions, and the attention maps are generated in this process. With the guidance of the attention maps and the input cloudy images, the attentive residual network can remove the clouds from the cloudy images via the learned negative residuals and the feature maps. Finally, the reconstruction network, which contains two residual blocks and one convolutional layer, can reconstruct a cloud-free image using the generated feature maps from the attentive residual network.

2.2.1. Attentive recurrent network

By choosing the certain inputs, the attention mechanism can enable neural networks to focus on a subset of their inputs. According to the specific task objectives, the direction of attention and the weighting model are adjusted. Attention mechanisms combined with deep learning models have become increasingly popular in recent years. Several methods have been presented in fine-grained object classification (Zhao et al., 2017), speech recognition (Chorowski et al., 2015), video captioning (Yan et al., 2019) and other areas. In a similar way, attention models are considered in our networks.

We exploit an attention mechanism to locate the cloudy areas in an image and capture the characteristics of cloud cover. As shown in Fig. 2,

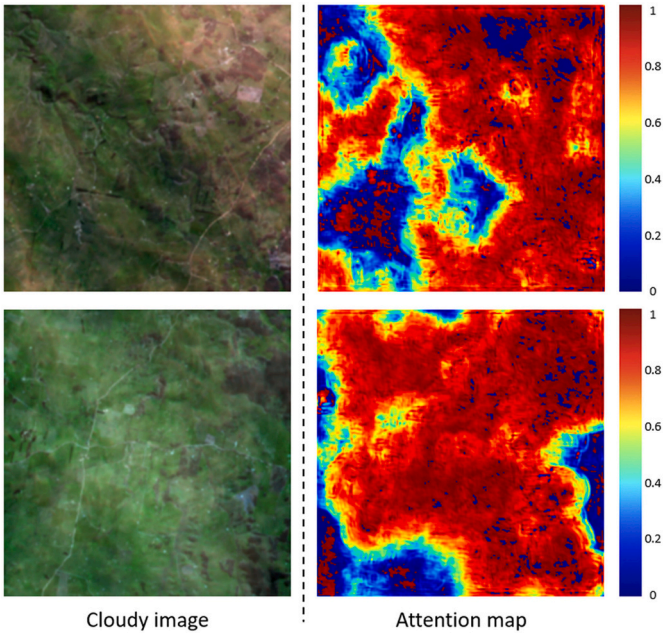

Fig. 3. Examples of cloudy images and their attention maps learned by the attentive recurrent network of AMGAN-CR. The attention map is a nonbinary graph that represents the increase in attention from cloudless pixels to cloudy pixels, with values between 0 and 1.

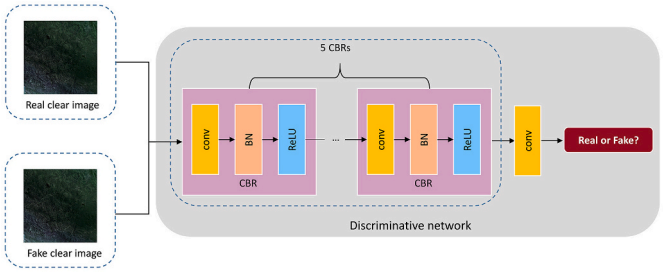

Fig. 4. The discriminative network of AMGAN-CR.

the attentive recurrent network contains K blocks in total. Each block includes five residual blocks (ResNets), a convolutional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) unit, and convolutional layers. The cloudy images are first input to the ResNets, which are used to extract the features of clouds. Then, the LSTM and convolutional layers can generate the attention maps $\boldsymbol{A_i}$, where i is denoted as the i-th attention map learned from the i-th block. $\boldsymbol{A_i}$, accompanied by the input images, is fed into the $(i+1)$-th block.

Unlike a binary mask, the attention map is a nonbinary graph that represents the increase in attention from the cloudless pixels to the cloudy pixels, with values between 0 and 1. The attention map shows the spatial distribution of clouds. A larger value represents greater attention in the map. Two cloudy images and their attention maps are shown in Fig. 3 for a better intuitive understanding. Fig. 3 shows the real-color cloudy images and their corresponding attention maps produced by the attentive recurrent network. We can clearly see that the attention values of surfaces covered by clouds are larger than those with no cloud cover. Increasing values represent increasing thickness of clouds, which can help us locate and remove the cloud cover in cloudy images.

The detailed structure of each component in the block is described as follows:

- ResNet: As shown in Fig. 2, ResNet consists of five residual blocks, in which there are three convolutions with stride = 1, which is the step length of the kernel movement, three rectified linear unit (ReLU) activation functions, an addition operation, and a shortcut connection. ResNet helps extract features from the input image and makes use of a shortcut connection to skip some layers in the forward step of an input, functioning as a direct channel for information flow. Starting from the second block, the input is a combination of the cloudy images and the attention map learned from the previous block.
- Long short-term memory (LSTM): LSTM is a common recurrent neural network design and crucial in the process of learning attention maps. Each LSTM unit is made up of a storage unit $\mathbf{c}$, a hidden state $\mathbf{h}$ and three different types of gates: input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$. These units are used to control storage devices that read and write data. The three gates determine how much to forget, remember, and acquire, which can be automatically learned by the error backpropagation algorithm during the training process. In the meanwhile, LSTM also contains the update of memory unit $\mathbf{c}$ and the update of hidden state $\mathbf{h}$:

1) Update of storage unit $\mathbf{c}$: $\mathbf{c}_{t-1}$ is updated to $\mathbf{c}_t$. $\mathbf{c}_{t-1}$ is first multiplied by the output value of the forget gate of $\mathbf{f}_t$, and only part of the historical memory is kept; The current state of $\widetilde{C}_t$ to be newly memorized is the one that should be recorded, and the current $\widetilde{C}_t$ to be memorized to the memory unit is produced by $\mathbf{h}_{t-1}$ and $\mathbf{x}_t$.
2) Update of the hidden state $\mathbf{h}$: $\mathbf{h}_{t-1}$ is updated to $\mathbf{h}_t$, and $\mathbf{h}_t$ takes the value from the current memory unit $\mathbf{c}_t$. After applying the hyperbolic function, *tanh*, the memory unit is weakened by the input gate $\mathbf{o}_t$, that is, $\mathbf{h}_t$ takes the value from $\mathbf{c}_t$ in a certain proportion.

Among them, t stands for time step, $\mathbf{c}_t$ serves as an accumulator of state information, which will be propagated to the next LSTM unit. $\mathbf{h}_t$ is the LSTM unit's output feature, which will be passed to the convolutional layer to generate the initial attention map.

- Convolutions: A 32-dimensional feature map can be learned from the LSTM. In this step, convolutions are used to translate the feature map into the desired dimension. Convolutional layers with stride of 1 and kernel size of 3×3 are used to generate the 2D attention maps.

2.2.2. *Attentive residual network*

As shown in Fig. 2, the attentive residual network includes a residual block, three convolutions, three ReLU activation functions, a multiplication operation, an addition operation, and a shortcut connection. The cloudy images are the input of the residual block, which is the same as the block in ResNet shown in the upper right of Fig. 2. A feature map is output through this residual block and then fed to a convolutional layer with 1×1 kernels. Convolution kernels generally feature an odd number of lines and columns within the shape of a square. A ReLU activation is connected after the convolutional layer following the 2-*nd* convolution with 3×3 kernels, ReLU, and a 3-*rd* convolution with 1×1 kernels. Then, the attention map learned from the attentive recurrent network multiplies the output to generate a 32-dimensional feature map. The removal of clouds is achieved through the learned negative residuals. That is, before the third ReLU activation function, the feature map is added to the output of the residual block.

2.2.3. *Reconstruction network*

The reconstruction network includes two residual blocks that are the same as described above and a final convolution with a kernel of size 3×3 as illustrated in Fig. 2. It is worth noting that in the residual block, the first and third convolutions use kernel size $= 1 \times 1$, and the second convolution uses kernel size $= 3 \times 3$, stride $= 1$, and padding $= 2$. Padding refers to padding around the edges in order to preserve the edge information of an image, and we use zero-padding to preserve the spatial dimensionality. The dimension of the output feature maps learned from

Table 2
Landsat 8 OLI bands.

Band	Band name	Wavelength (μm)	Resolution (m)
1	Coastal Aerosol	0.43–0.45	30
2	Blue	0.45–0.51	30
3	Green	0.53–0.59	30
4	Red	0.64–0.67	30
5	Near Infrared	0.85–0.88	30
6	SWIR1	1.57–1.65	30
7	SWIR2	2.11–2.29	30
8	Panchromatic	0.50–0.68	15
9	Cirrus	1.36–1.38	30

the attentive residual network is 256 × 256 × 32. The final convolution reconstructs the spectral dimensions to match the input, that is, converts the 32 dimensions to the number of spectral bands of the cloudy image.

2.3. Discriminative network

The real clear image and fake clear image are input the discriminative network, which is used to evaluate the possibility that an image is a real clear image rather than a generated clear image from the generative network. As shown in the architecture in Fig. 4, the discriminator consists of five CBRs and a 3 × 3 convolution, where C represents convolutions with 3 × 3 kernels and stride = 1, B is denoted as the batch normalization, and R refers to the activation function ReLU. The convolution in the first CBR transforms the input 7-dimensional data into 32-dimensional features, followed feature dimension of each CBR was 64, 128, 256, 512 from left to right, all using 4 × 4 kernels and stride 2. Finally, convolution is used to convert the 512 dimensions to 1 dimension, with 3 × 3 kernels and stride 1.

2.4. Loss function of AMGAN-CR

We name the loss function of the proposed AMGAN-CR method $\mathscr{L}_{\mathrm{AMGAN-CR}}$, and it is employed in the optimization of parameters in generative and discriminative networks. To reduce the gap between the generated clear image and the ground-truth image, $\mathscr{L}_{\mathrm{AMGAN-CR}}$ is formulated as:

$$\mathscr{L}_{\mathrm{AMGAN-CR}} = \underset{P}{\mathrm{argmin}}\underset{Q}{\mathrm{max}}\theta_1 \mathscr{L}_{\mathrm{cGAN}}(P,Q) + \theta_2 \mathscr{L}_{\mathrm{Att}} + \theta_3 \mathscr{L}_{\mathrm{L1}}, \tag{2}$$

where $\mathscr{L}_{\mathrm{cGAN}}(P,Q)$ is the conditional GAN loss function, $\mathscr{L}_{\mathrm{Att}}$ is the attention loss, and $\mathscr{L}_{\mathrm{L1}}$ is an improved $\mathscr{L}1$ loss function used to calculate the precision of each reconstructed pixel. θ_1, θ_2 and θ_3 are the balance factors of these losses. Specifically, $\mathscr{L}_{\mathrm{cGAN}}(P,Q)$ is expressed as

$$\mathscr{L}_{\mathrm{cGAN}}(P,Q) = E_{u,v\sim p_{data}(u,v)}[logQ(u,v)] + E_{u\sim p_{data}(u),j\sim p_j(j)}[log(1-Q(u,P(u,j)))], \tag{3}$$

where u represents the cloudy image, $p_{data}(u)$ represents the distribution of u, v represents the real cloud-free image, j is denoted as random noise data, $p_j(j)$ is the noise distribution, $P(u,j)$ refers to the cloud-free image generated by u with the aid of j, and $Q(u,v)$ is the result of the discriminator, representing the possibility that v is close to the real clear image.

The attention map produced by the attentive recurrent network is used to calculate the attention loss $\mathscr{L}_{\mathrm{Att}}$:

$$\mathscr{L}_{\mathrm{Att}} = \|\mathbf{A} - \mathbf{M}\|^2, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{W\times H}$ is a matrix that represents the attention map and $\mathbf{M} \in \mathbb{R}^{W\times H}$ is the so-called cloud mask. W and H represent the width and height of $\mathbf{A}$ and $\mathbf{M}$, respectively. In Eq. (4), $\mathbf{M}$ is calculated by using the cloudy image to subtract the real clear image, and then an interval $[0,1]$ is selected to clip the values smaller than 0 and larger than 1. Values outside the interval are clipped to the interval edges. Afterwards, a binary mask can be generated, which can be regarded as a so-called cloud mask. The values 1 and 0 in $\mathbf{M}$ represent cloudy and clear pixels, respectively. $\mathscr{L}_{\mathrm{Att}}$ optimizes the network by matching the attention map to the subtraction of cloudy and clear images.

The third component of $\mathscr{L}_{\mathrm{AMGAN-CR}}$ is utilized to determine the accuracy of each reconstructed pixel, which is denoted as:

$$\mathscr{L}_{\mathrm{L1}} = \frac{1}{NWH}\sum_{r=1}^{N}\sum_{w=1}^{W}\sum_{h=1}^{H}\lambda_r \left\| I_{\mathrm{output}}^{(w,h,r)} - f\left(I_{\mathrm{input}}^{(w,h,r)}\right)\right\|_1, \tag{5}$$

where N represents the number of bands of the input cloudy image, $W * H$ represents the size of the image, λ_r refers to the weight of the r-th band, I_{input} is denoted as the input image of the generative network, the output result is I_{output}, $f(I_{\mathrm{input}})$ refers to the predicted image of the generative network, and (w,h,r) represents a pixel at location (w,h) in the r-th band.

3. Data preprocessing and evaluation

3.1. Landsat data preprocessing

The Landsat project is continually collecting global land cover data, which provides us with suitable data to implement the experiments.

Table 3
Summary of the study sites used in the simulated and real cloud datasets. The percentage of cloud over is acquired from the USGS website (https://earthexplorer.usgs.gov).

Dataset	Pair	Condition	ID	Percentage of Cloud Cover	Acquisition Date
Simulated Dataset	1	Cloud Band	LC08_L1TP_090084_20140421_20170423_01_T1	42.69%	2014/04/21
		Cloud-Free	LC08_L1TP_090084_20140115_20170426_01_T1	0.28%	2014/01/15
	2	Cloud Band	LC08_L1TP_091084_20191104_20191115_01_T1	37.96%	2019/11/04
		Cloud-Free	LC08_L1TP_091084_20191019_20191029_01_T1	0.01%	2019/10/19
Real Dataset	1	Cloudy	LC08_L1TP_090084_20140507_20170422_01_T1	9.73%	2014/05/07
		Cloud-Free	LC08_L1TP_090084_20140421_20170423_01_T1	42.69%	2014/04/21
	2	Cloudy	LC08_L1TP_091084_20191104_20191115_01_T1	37.96%	2019/11/04
		Cloud-Free	LC08_L1TP_091084_20191019_20191029_01_T1	0.01%	2019/10/19
	3	Cloudy	LC08_L1TP_090084_20131230_20170427_01_T1	5.24%	2013/12/30
		Cloud-Free	LC08_L1TP_090084_20140115_20170426_01_T1	0.28%	2014/01/15
	4	Cloudy	LC08_L1TP_089084_20131020_20170429_01_T1	27.15%	2013/10/20
		Cloud-Free	LC08_L1TP_089084_20130817_20170502_01_T1	0.03%	2013/08/17
	5	Cloudy	LC08_L1TP_090085_20130824_20170502_01_T1	5.12%	2013/08/24
		Cloud-Free	LC08_L1TP_090085_20131011_20170429_01_T1	0.10%	2013/10/11
	6	Cloudy	LC08_L1TP_096072_20140126_20170426_01_T1	31.12%	2014/01/26
		Cloud-Free	LC08_L1TP_096072_20140315_20170425_01_T1	1.63%	2014/03/15
	7	Cloudy	LC08_L1TP_090084_20160715_20170323_01_T1	9.85%	2016/07/15
		Cloud-Free	LC08_L1TP_090084_20160222_20170329_01_T1	0.02%	2016/02/22

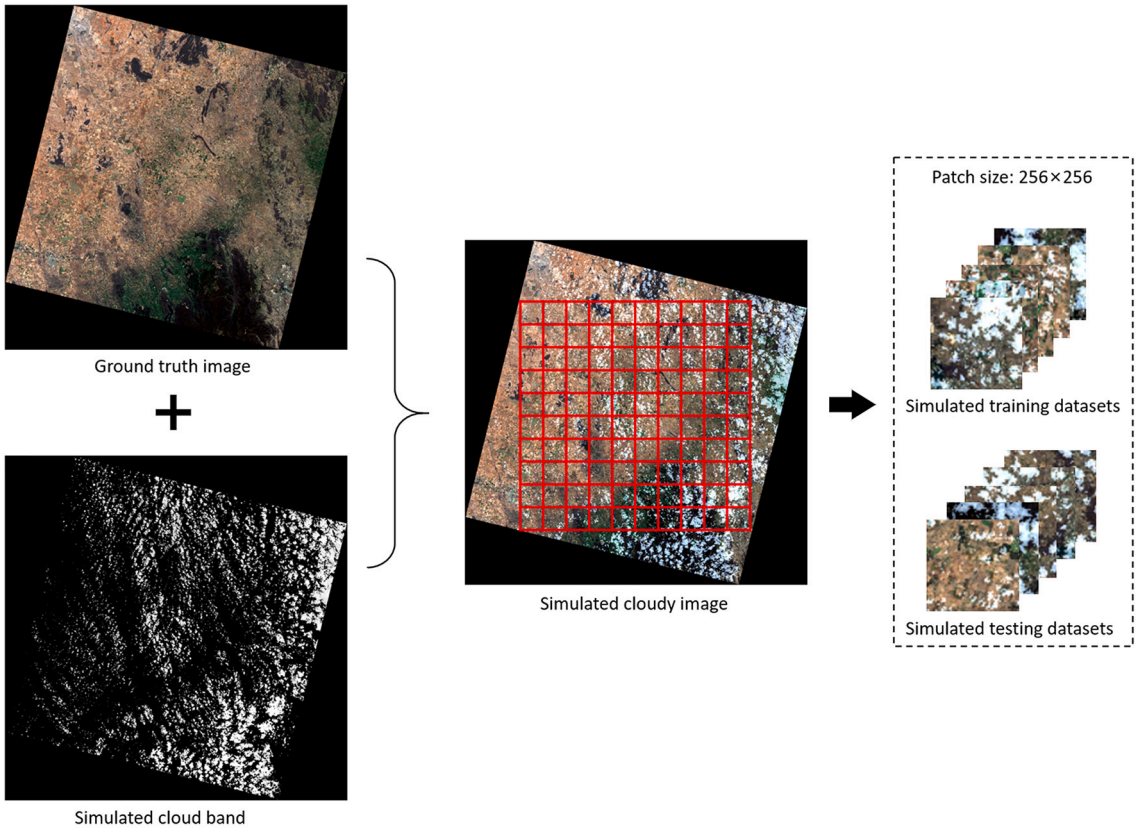

Fig. 5. Diagram of synthesizing a simulated cloudy image by Eq. (6). The simulated cloudy image is generated by a ground-truth image and a simulated cloud band. The thickness factor α_i is set as 1. The entire Landsat scene is partitioned into a set of small 256×256 nonoverlapping patches and separated into simulated training and testing datasets.

Table 4
Training, validation and test datasets for the simulated cloud dataset and real cloud dataset.

	Images (256×256)		
	Train 64%	Validate 16%	Test 20%
Simulated cloud dataset	384	96	120
Real cloud dataset	578	144	181

Table 5
Average PSNR, SSIM, RMSE values on the simulated testing dataset corresponding to different θ_1 values when $\theta_2 = 10$ and $\theta_3 = 10$.

θ_1 value	0.01	0.03	0.05	0.07	0.1
PSNR	**25.256**	25.013	24.901	24.808	24.743
SSIM	**0.900**	0.895	0.888	0.883	0.863
RMSE	**0.060**	0.061	0.062	0.063	0.064

Furthermore, Landsat time series satellite images are frequently utilized for land use and land cover change detection, which is generally necessary to remove cloud cover in the data preprocessing step. From the above considerations, Landsat 8 images were used in the experiments. The OLI imagery has a spatial resolution of 30 m and comprises visible, near-infrared, and shortwave infrared spectral bands, as well as a 15 m panchromatic band. Table 2 shows the wavelengths of Landsat 8 OLI bands used in the experiment. Band 9 is also called the cirrus band, which is a band centered at a wavelength of 1.375 μm in Landsat 8 OLI. It is a strong water vapor absorption band and helps detect cirrus clouds within OLI images. Band 9 is used to help synthesize the simulated cloudy images. We preprocess the images by converting the unsigned

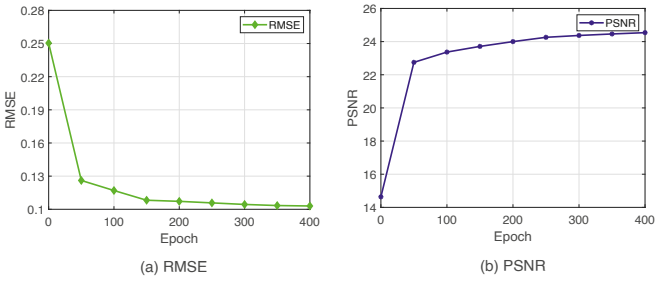

Fig. 6. (a) Plot of RMSE versus the number of epochs during training. (b) Plot of PSNR versus the number of epochs during training.

16-bit integers to 8-bits. The lower and higher end of 2% data values are saturated to 0 and 255, respectively, in order to present the informative data values better. The raw surface reflectance data can also be directly used in the experiments; nevertheless, a linear 2% stretch is better for visual color display. Finally, we normalize the pixels from 0 to 255 to the range of 0–1 to accelerate the convergence of the training network because normalization does not change the image information and is essential for reducing training times and improving training efficiency.

3.2. Dataset preparation

Because of the lack of ground truth in the real experiments, it is difficult to evaluate the results of cloud removal quantitatively. Therefore, a simulated cloud dataset and a real cloud dataset are both used in the experiment. The simulated experiments more easily demonstrate the effectiveness quantitatively. We use the algorithms mentioned in Xu

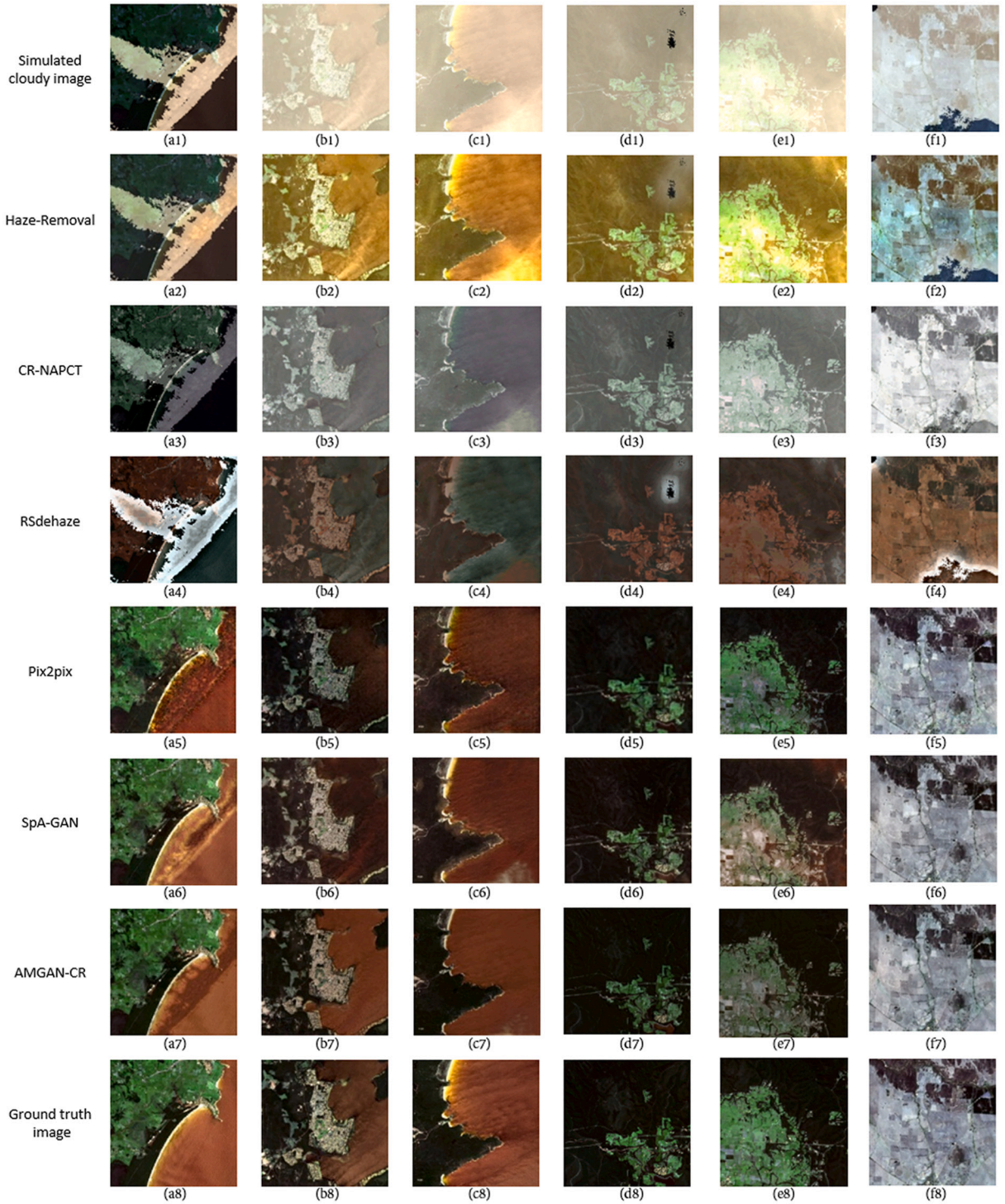

Fig. 7. Experimental results on simulated data displayed in the composition of bands 2, 3, and 4. The first row shows six cloudy images with a size of 256×256 that are subsets of the two paired simulation images. The second to seventh rows show the cloud removal images generated by the Haze-Removal, CR-NAPCT, RSdehaze, Pix2pix, SpA-GAN and AMGAN-CR methods, respectively. The last row shows the corresponding ground-truth scenes.

et al. (2015) to synthesize a simulated cloudy image. First, the cirrus band is converted into a cloud band by manually selecting a threshold value. Additionally, a cloud mask is produced by masking the nonzero values of the cloud band. Second, we add the cloud band onto the seven bands of a cloud-free image. The synthetic formula of the simulated cloud is as follows:

$$\mathbf{I}_i^t = \mathbf{I}_i^g + \alpha_i s^c \mathbf{I}^c, \tag{6}$$

where $\mathbf{I}^c$ represents the cloud band, $\mathbf{I}_i^t$ represents the synthesized cloudy image, and $\mathbf{I}_i^g$ represents the clear image. i represents the number of bands, s^c is the set of cloud spectra that can simulate real cloud effects on different bands, and α_i is the coefficient that can control the thickness of cloud cover. The simulated dataset used in our experiment is presented in Table 3, containing two pairs of simulated images, and the simulation diagram is shown in the left part of Fig. 5, where α_i is set as 1.

The real cloud datasets consist of seven pairs of cloudy and cloudless

Table 6
Average PSNR, SSIM and RMSE values calculated for the scenes in Fig. 7. The best performance in each evaluation metric is marked in bold. RGB denotes the measurement on all three bands.

Metric	Method	Band 2	Band 3	Band 4	RGB
PSNR	Haze-Removal	17.04	10.35	9.62	11.20
	CR-NAPCT	10.97	11.11	10.81	10.80
	RSdehaze	15.49	15.48	14.65	14.65
	Pix2pix	25.97	22.94	21.84	23.11
	SpA-GAN	26.26	24.12	21.31	23.33
	AMGAN-CR	**28.35**	**24.20**	**24.53**	**25.11**
SSIM	Haze-Removal	0.67	0.56	0.54	0.59
	CR-NAPCT	0.50	0.56	0.56	0.54
	RSdehaze	0.56	0.61	0.65	0.61
	Pix2pix	0.85	0.84	0.81	0.83
	SpA-GAN	0.86	0.87	0.84	0.85
	AMGAN-CR	**0.91**	**0.90**	**0.89**	**0.90**
RMSE	Haze-Removal	0.15	0.32	0.35	0.29
	CR-NAPCT	0.29	0.28	0.29	0.29
	RSdehaze	0.20	0.19	0.20	0.20
	Pix2pix	0.05	0.08	0.10	0.08
	SpA-GAN	0.05	0.07	0.10	0.08
	AMGAN-CR	**0.04**	**0.07**	**0.07**	**0.06**

multitemporal images, which are displayed in Table 3. All the study sites are located in Australia. The images with WRS Path/Row 90/85, 90/84, 91/84 and 89/84 are around Canberra and Sydney in southeastern Australia. The image with WRS Path/Row 96/72 is collected near Mareeba in northeastern Australia. The acquisition dates range from 2013 to 2019. The cloud-free images with less than 10% cloud cover are selected as the reference images on the nearest date from cloudy images. There is an exception that the first pair of real datasets contains a cloud-free image with 42.69% cloud cover. It also can be noticed that the cloud-free image in the first pair and the cloudy image in the second pair of the real dataset are also used in the simulated dataset. The reason is that the same image used in both experiments can be better to evaluate the results visually. Since each entire Landsat scene is partitioned into a set of small 256×256 nonoverlapping patches, which is shown in the right part of Fig. 5, we use only the images in the central square without the external images that are filled with dark areas, and the cloudy areas of the cloud-free image are excluded from our datasets. The reference images are treated as part of the training data in the proposed AMGAN-CR method for the purpose of learning the mapping from cloudy to cloudless images and the ground-truth images when evaluating the test results quantitatively and qualitatively. Furthermore, the cloudy and corresponding clear images are coregistered to ensure that the pixels are paired correctly. It is important to note that after cropping the original images, 903 patches in the real cloud dataset and 600 patches in the simulated cloud dataset are obtained. All the cropped patches are divided into three subsets: training, validation, and test datasets. Table 4 shows the proportions of three datasets.

3.3. Evaluation metrics

We utilized three quantitative evaluation metrics to assess the quality of all experimental methods. The first metric is the peak signal-to-noise ratio (PSNR) that is the most frequently used objective measuring tool for evaluating image quality. It is given by:

$$\mathrm{PSNR} = 10 \cdot log_{10}\left(\frac{(2^{\mathrm{B}}-1)^2}{\mathrm{MSE}}\right), \tag{7}$$

where B represents the bit depth. When an image is 8-bit data, $2^{\mathrm{B}}-1$ will be 255. MSE is the mean squared error between the cloud removed image and the ground-truth image. Given a cloud-removed image $\boldsymbol{X}_{\mathrm{removed}} \in \mathbb{R}^{x \times y}$ and the ground-truth image $\widehat{\boldsymbol{X}}_{\mathrm{truth}} \in \mathbb{R}^{x \times y}$, MSE is calculated according to:

$$\mathrm{MSE} = \frac{1}{n} \| \boldsymbol{X}_{\mathrm{removed}} - \widehat{\boldsymbol{X}}_{\mathrm{truth}} \|^2, \tag{8}$$

where $n = x \times y$ denotes the number of pixels in the image. A larger PSNR value indicates a closer relationship between the cloud-removed image and the ground-truth image.

Second, the structural similarity index measurement (SSIM) is a metric used to assess the similarity of two images and determine image quality with the concern of structural information deterioration. SSIM is defined as

$$\mathrm{SSIM} = l(p,q) \cdot c(p,q) \cdot s(p,q). \tag{9}$$

The calculation of SSIM is based on three comparison measures, namely luminance l, contrast c and structure s, between samples p and q: $l(p,q) = \frac{2\mu_p\mu_q + c_1}{\mu_p^2 + \mu_q^2 + c_1}$, $c(p,q) = \frac{2\sigma_p\sigma_q + c_2}{\sigma_p^2 + \sigma_q^2 + c_2}$, $s(p,q) = \frac{\sigma_{pq} + c_3}{\sigma_p\sigma_q + c_3}$. Among them, c_1, c_2, c_3 are constants. μ_p and μ_q are the mean values of p and q. ${\sigma_p}^2$ and ${\sigma_q}^2$ are the variances of p and q. σ_{pq} is the covariance of p and q. The range of SSIM is between 0 and 1. A larger SSIM value indicates a greater similarity between the two images. When the two images are identical, SSIM is equal to 1.

The third evaluation metric is the root-mean-square error (RMSE) between the ground-truth and corrected image: $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$. It is used in the simulation tests.

4. Experiment

To verify the proposed AMGAN-CR method, three traditional methods and two deep learning methods are executed as the baselines for comparison. The traditional methods include haze removal based on the deformed haze imaging model (Haze-Removal) (Pan et al., 2015), haze removal for a single visible remote sensing image (RSdehaze) (Liu et al., 2017) and cloud removal based on the noise-adjusted principal components transform model (CR-NAPCT) (Xu et al., 2019). These methods take only the cloudy image as the input without the requirement of corresponding cloud-free images. Specifically, Pan et al. (2015) proposed a haze imaging model that was used to remove haze from RGB color images: $\mathbf{J}(\mathbf{x}) = \frac{\mathbf{I}(\mathbf{x}) - \mathbf{R}}{\mathbf{t}(\mathbf{x})} + \mathbf{R}$. $\mathbf{J}$ represents the radiance, $\mathbf{I}$ is the target image, $\mathbf{R}$ denotes the global atmospheric light and $\mathbf{t}$ is the proportion of the light not reaching the camera. Since the atmospheric light of the three RGB channels was the same, $\mathbf{R}$ was estimated by calculating the highest pixel values in the image. $\mathbf{J}$ needs to be subtracted by a constant parameter C, which is used to decrease the deviation when utilizing the dark channel prior to estimating the transmittance. C is set to 27 in our comparative experiment. In Liu et al. (2017), haze is treated as additive contamination that can be represented by a haze thickness map (HTM). They first used the total variable regularization of the inpainting method to remove some textures and brighter parts and then used the average value as the upper boundary to suppress ground reflected radiance. Xu et al. (2019) took advantage of the property that a higher local spatial correlation corresponds to a higher the signal-to-noise ratio (S/N) value. Cloud detection was implemented first, and then clouds were removed by inversing the noise-adjusted principal component transform model. Cloud masks are necessary for the CR-NAPCT method in order to preserve the cloud-free area information. Therefore, we obtain the cloud masks by the Fmask (Zhu et al., 2015) method, which detects clouds, cloud shadows, and snow for Landsats 4–8 and Sentinel 2 images.

The two deep learning methods are a pix2pix GAN framework (Pix2pix) (Isola et al., 2017) and spatial attention generative adversarial networks (SpA-GAN) (Pan, 2020), which need paired cloudy images and cloudless images as the training datasets. Isola et al. (2017) learned the mapping and loss function from the input image to the output image. The optimizer for this method is Adam, with a learning rate of 0.0002, and number of epochs equals to 200. Pan (2020) used four spatial attentive blocks (SAB), and each SAB had three spatial attentive residual

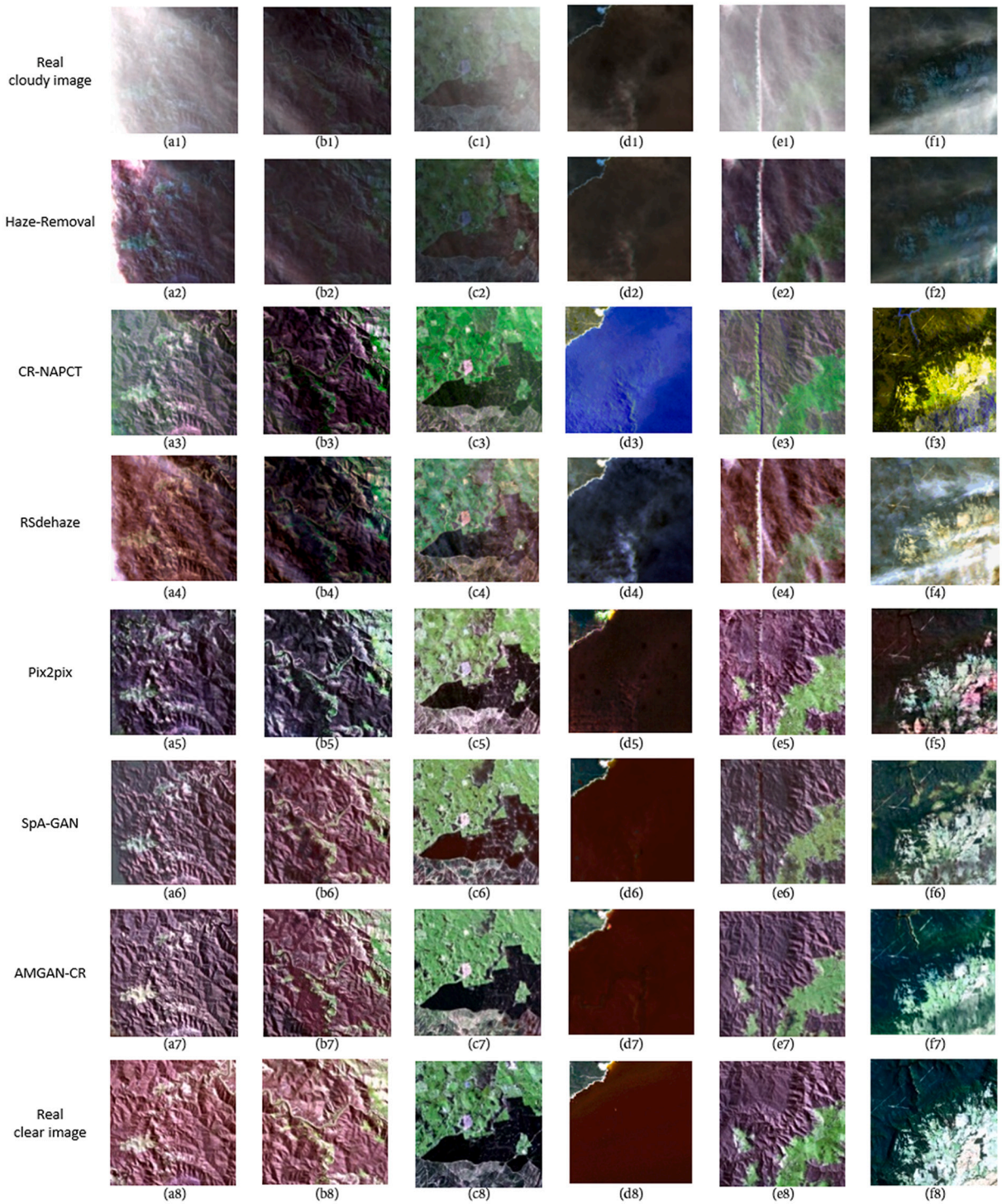

Fig. 8. Experimental results on real data displayed in the composition of bands 2, 3, and 4. The first row shows the six cloudy images. The second to seventh rows show the cloud removal images generated by the Haze-Removal, CR-NAPCT, RSdehaze, Pix2pix, SpA-GAN and AMGAN-CR methods, respectively. The last row shows the corresponding reference images with no cloud cover.

blocks (SARB) and one spatial attentive module (SAM), as was originally proposed in Wang et al. (2019). For the network used for extracting the attention map, in Pan (2020), a two-round four-directional identity matrix initialization architecture was used to obtain the features in the four directions and then the weights in the four directions were multiplied to concatenate the results. Finally, the attention map is output through additional convolutions and sigmoid activations and then used the residual network to subtract the attention map from the cloud image to obtain a cloud-free image. The optimizer of SpA-GAN is Adam, with a learning rate of 0.0004, number of epochs equals to 200, and minibatch equal to 1. In contrast, it is worth noting that in our method the attentive recurrent network with several blocks in which the features are extracted through five residual networks to obtain the attention map. This bottleneck structure can improve the speed of calculation. First, convolutions and ReLUs are used to perform a regular feature extraction on the cloudy image. However, the performance of a single residual

Table 7
Average PSNR and SSIM values calculated for the scenes in Fig. 8. The best performance in each evaluation metric is marked in bold. RGB denotes the measurement on all three bands.

Metric	Method	Band 2	Band 3	Band 4	RGB
PSNR	Haze-Removal	12.90	12.29	11.58	12.18
	CR-NAPCT	12.32	13.45	14.39	12.49
	RSdehaze	12.18	13.76	12.18	12.58
	Pix2pix	16.16	16.05	12.91	14.58
	SpA-GAN	19.05	18.47	16.24	17.65
	AMGAN-CR	**19.80**	**18.99**	**17.70**	**18.63**
SSIM	Haze-Removal	0.46	0.44	0.35	0.42
	CR-NAPCT	0.45	0.50	0.56	0.51
	RSdehaze	0.41	0.52	0.50	0.47
	Pix2pix	0.68	0.66	0.56	0.63
	SpA-GAN	0.78	0.75	0.69	0.74
	AMGAN-CR	**0.82**	**0.81**	**0.77**	**0.80**

Table 8
The performance of each band in the real cloud dataset.

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7
PSNR	16.03	19.80	18.99	17.70	17.32	17.24	16.68
SSIM	0.72	0.82	0.82	0.77	0.79	0.78	0.66

block for feature extraction was found to be insufficient, so we use five residual blocks. Then, LSTM is used to process the non-linear structure of the image and retain valuable information. Finally, a two-dimensional attention map is obtained through convolutions. Furthermore, we also reduced the number of network layers to further improve the performance as seen from the experimental results.

Three parameters θ_1, θ_2 and θ_3 are the weighting factors, aiming to make three individual losses contribute equally to the total loss minimization $\mathscr{L}_{\mathrm{AMGAN-CR}}$ in AMGAN-CR. According to Isola et al. (2017), the quality of images achieves the best when $\theta_1 = 0.01$, $\theta_2 = 0$, and $\theta_3 = 10$ in our case. Therefore, we set $\theta_1 = 0.01$ and $\theta_3 = 10$ as the recommended and typical selection in most of the well-known studies. The attention loss $\mathscr{L}_{\mathrm{Att}}$ is used to calculate the difference between the attention map and the cloud mask. The influence of $\mathscr{L}_{\mathrm{Att}}$ and $\mathscr{L}1$ are expected equal. Based on the definition of these two losses, they are both used to minimize the distances between outputs and ground truth inputs. Therefore, θ_2 is set to 10. For the purpose of examining the effects of different values of these parameters to the results, we keep $\theta_2 = 10$ and $\theta_3 = 10$ unchanged and θ_1 is set to 0.01, 0.03, 0.05, 0.07, and 0.1. Table 5 shows the average PSNR, SSIM and RMSE values with different values of θ_1 on the simulated testing dataset. It can be seen that changing the values of θ_1 from 0.01 to 0.1 reduces slightly the performance of AMGAN-CR and all metrics have the best result when $\theta_1 = 0.01$. Thus, we set $\theta_1 = 0.01$, $\theta_2 = 10$, and $\theta_3 = 10$ in our experiment.

The training epoch is set as 200. It can be seen from Fig. 6, when the number of epochs reaches 200, the values of RMSE and PSNR tend to stabilize. Therefore, we set training epoch to 200 as the effective number to achieve the desired performance. The number of epochs is an important hyperparameter that indicates the number of complete cycles for the whole training dataset to learn the process of the algorithm. The internal model parameters of the dataset are updated per epoch. Our training and testing experiments are executed on a Windows 10 operating system with an Intel(R) Core(R) i5-4590 CPU @ 3.30 GHz and an NVIDIA Tesla P100 PCIe with 16 GB RAM using the PyTorch framework. Since all the baseline methods except for CR-NAPCT are developed for natural images, we select only bands 2, 3, and 4 of the Landsat images to implement all the comparative analyses. It should be noticed that the presented AMGAN-CR method is developed with seven bands of cloudy images. The dataset given in Section 3 is used to train all algorithms. We also conduct an ablation study to investigate the influence of K used in the attentive recurrent network, which is described in Section 2.2.1, the influence of N in Eq. (5), which is the number of bands in the input cloudy image, and the influence of the so-called cloud mask $\mathbf{M}$ in Eq. (4) on the performance of AMGAN-CR.

4.1. Experimental results on the simulated datasets

Fig. 7 presents the outcomes of several approaches on six cloudy image in RGB color utilizing bands 2, 3, and 4. The first row shows the simulated cloudy images synthesized by the method mentioned in

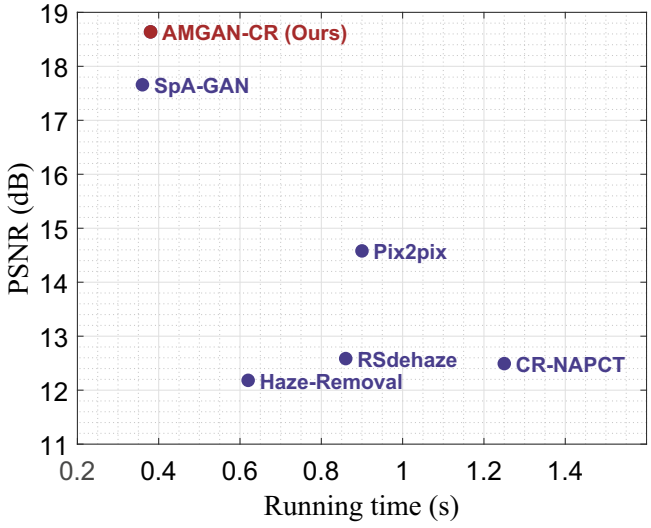

Fig. 10. Average running time and PSNR of six methods.

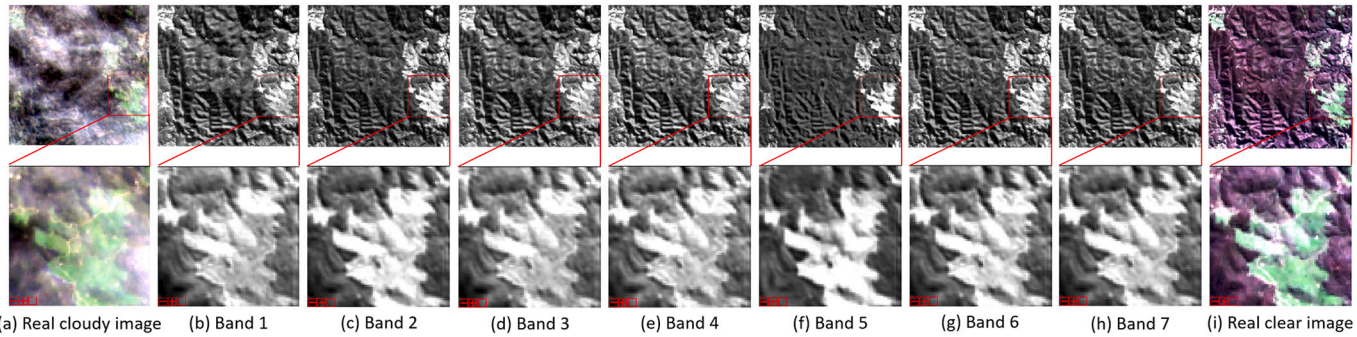

Fig. 9. Columns (b)-(h) display each band of our cloud removal results corresponding to Table 8. The second row is the magnified view of the red box in the first row, and the first and last columns are the cloudy and cloud-free reference images, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

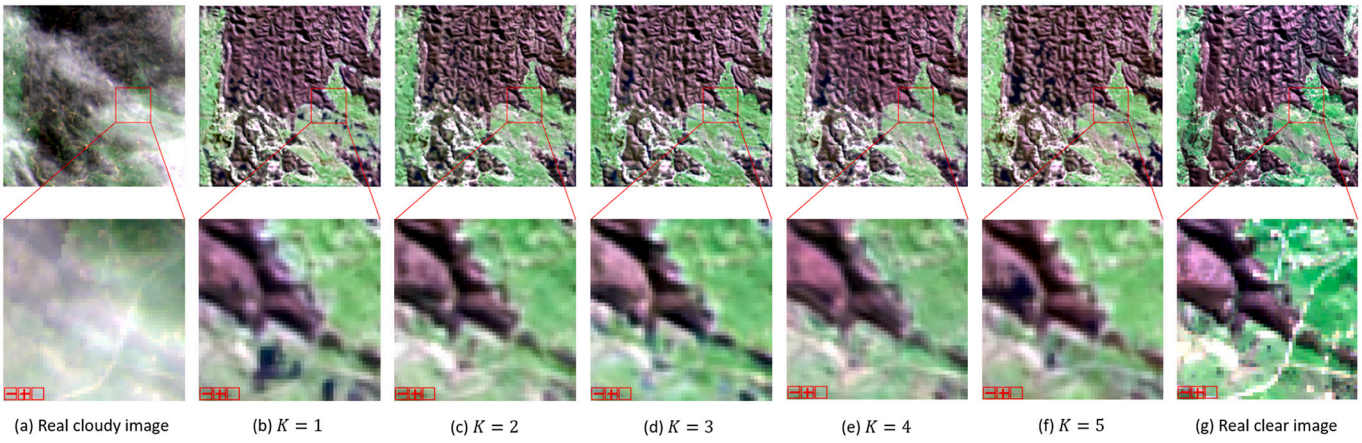

Fig. 11. An example image produced with $K = 1$ to $K = 5$ in the attentive recurrent network. The second row is an enlarged version of the red box in the first row, and the first and last columns are the cloudy and cloud-free reference images, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Average PSNR and SSIM values calculated with different K on the real datasets.

Metric	K	Band 2	Band 3	Band 4	RGB
PSNR	K=1	18.394	18.622	18.037	18.292
	K=2	18.438	18.839	18.290	18.448
	K=3	18.685	18.869	18.311	18.555
	K=4	**18.713**	18.942	**18.434**	**18.633**
	K=5	18.678	**18.983**	18.282	18.565
SSIM	K=1	0.710	0.723	0.693	0.709
	K=2	0.709	0.723	0.693	0.709
	K=3	0.711	0.722	0.695	0.709
	K=4	**0.713**	**0.725**	**0.698**	**0.712**
	K=5	0.710	0.723	0.689	0.707

Section 3, and the last row shows the ground-truth images corresponding to the simulated cloudy images. The six cloudy images are covered by representative ground surfaces and clouds of different shapes and thicknesses. The first three cloudy scenes all include water areas, which are always ignored by cloud removal algorithms. It can be seen from Fig. 7 that the three traditional methods, which are Haze-Removal, CR-NAPCT and RSdehaze, did not work well in removing simulated clouds. The Haze-Removal method has a better result in Fig. 7(f2) than the other two traditional methods. Due to the particular shape of cloud cover over water areas in Fig. 7(a1), the three methods are not able to remove clouds effectively. This common problem occurred only with traditional methods, which may be because of the different distributions of simulated clouds and real clouds. The results of the deep learning methods, which are Pix2pix, SpA-GAN and the proposed AMGAN-CR, are all visually better for the cloudy images than those of the above three methods. Similarly, Fig. 7(a5)-(a7) present the boundaries after cloud removal, but in Fig. 7(b7) and (c7), the clouds are completely removed, without the blurred areas that exist in Fig. 7(b5), (c5) and (b6), (c6). The results of Fig. 7(d1), (e1), and (f1) processed by our method outperform those of Pix2pix and SpA-GAN when compared with the ground-truth image visually.

To make a quantitative comparison, the average values of PSNR, SSIM and RMSE calculated on bands 2, 3, 4 and RGB are listed in Table 6. The best results in each assessment metric are marked in bold. It is clearly seen that the proposed AMGAN-CR achieved the best results among all six methods, which are consistent in the visual display. Specifically, the values of PSNR, SSIM and RMSE on band 2 are better than those of other bands for almost all the methods owing to the greater effects of clouds on shortwave bands than longwave bands.

4.2. *Experimental results on real datasets*

Fig. 8 shows the outcomes of several approaches on six cloudy

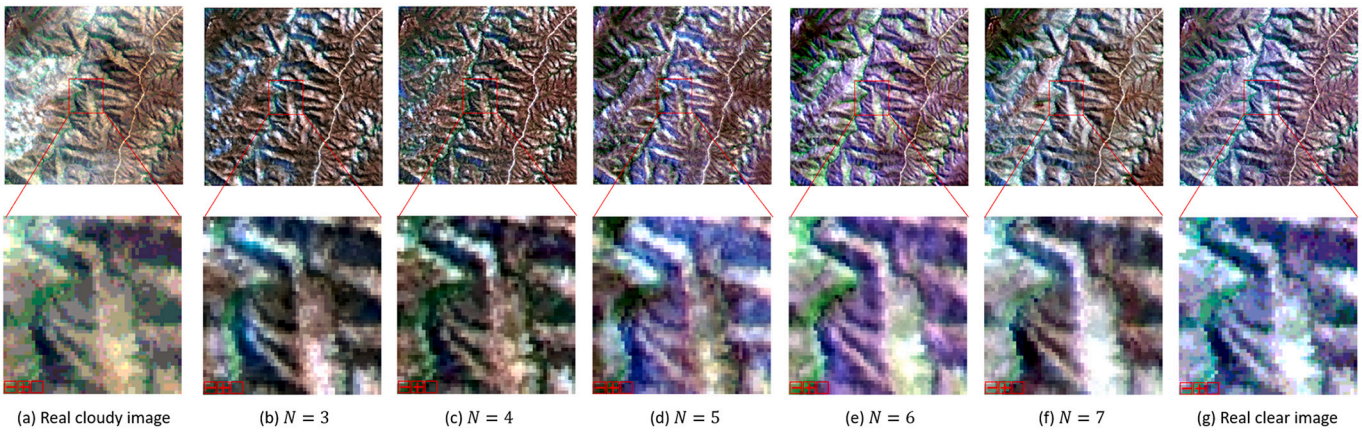

Fig. 12. An example image produced with $N = 3$ to $N = 7$ from the cloudy image. The second row is an enlarged version of the red box in the first row, and the first and last columns are the cloudy and cloud-free reference images, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 10
Average PSNR and SSIM values calculated with different numbers of bands N on the real datasets.

Metric	N	Band 2	Band 3	Band 4	RGB
PSNR	N=3	24.643	23.118	21.079	22.616
	N=4	24.261	22.379	20.780	22.137
	N=5	**27.965**	25.620	24.128	25.542
	N=6	27.731	25.250	23.807	25.232
	N=7	27.869	**25.867**	**24.297**	**25.642**
SSIM	N=3	0.875	0.860	0.823	0.853
	N=4	0.860	0.845	0.814	0.840
	N=5	0.924	0.903	**0.879**	0.902
	N=6	0.922	0.896	0.869	0.895
	N=7	**0.925**	**0.908**	**0.879**	**0.904**

Table 11
Average PSNR, SSIM and RMSE values calculated on the simulated datasets with different methods of calculating $\mathbf{M}$ in $\mathscr{L}_{Att}$. The algorithm used in detection is the Fmask zhu2015improvement cloud detection method.

Metric	Method of calculating $\mathbf{M}$	Band 2	Band 3	Band 4	RGB
PSNR	Detection	27.229	24.982	23.383	24.790
	Subtraction	**28.368**	**25.461**	**23.631**	**25.314**
SSIM	Detection	0.924	0.901	0.872	0.899
	Subtraction	**0.933**	**0.904**	**0.873**	**0.903**
RMSE	Detection	0.049	0.063	0.076	0.064
	Subtraction	**0.042**	**0.058**	**0.072**	**0.059**

images in natural color. The reference images without cloud cover are collected on near date to the cloudy images. Fig. 8(a1) and (e1) are covered by relatively thicker clouds than the others, and (e1) has an obvious strip of clouds. The Haze-Removal and RSdehaze methods do not correct the cloudy images effectively, with some clouds retained. CR-NAPCT can remove clouds from most of the scenes, while over-correction clearly occurs in the last three images, especially in the image with large water areas. SpA-GAN performs nearly as well as AMGAN-CR, the results of which are visually better than those of Pix2pix. However, the boundaries after correction in Fig. 8(e6) are more remarkable than those in (e7). Although the results of AMGAN-CR still present some blurred areas in the removal image, the overall performance is the best visually. The average values of PSNR and SSIM are provided in Table 7 for a quantitative comparison. The best results in each assessment metric are marked in bold, and it is shown that the proposed method produces the best result. We use seven bands of the Landsat 8 images in the experiments. Bands 2, 3, and 4 (RGB) are selected for visual inspection in qualitative evaluation and fair quantitative comparison with other methods. To present the performance comprehensively, we add the results of all bands in Table 8. In addition, we use Fig. 9 to show the intuitive visual effects of each band. As an example we can see that the texture is clearer and more complete after the correction in Band 2.

It is noteworthy that the proposed method requires about three hours for model training. Fig. 6 shows that it takes about 200 epochs until training convergence. While training requires a lot of computation time, our method is more effective than several other cloud removal methods in using the model to conduct cloud removal. Fig. 10 shows the average running time and PSNR of six methods over the tests on the 10 real cloudy images. Since Haze-Removal and RSdehaze are based on the Haze Imaging Model, and CR-NAPCT is based on the noise-adjusted principal components transform model, complex optimizations are still required to remove clouds of each new image, resulting in slower processing time. Among the three deep learning methods, Pix2pix is processed pixel by pixel so that the processing time is longer. Both our method and SpA-GAN are faster due to the direct application of the trained models. Although SpA-GAN takes a little shorten running time than AMGAN-CR, its PSNR is lower. As a result, our method has a good trade-off between computational costs and accuracy.

4.3. Ablation experiment

4.3.1. Influence of K in the attentive recurrent network

To show the impact of K (the total number of blocks in the attentive recurrent network), we performed five experiments, and Fig. 11 shows an example image produced with $K = 1$ to $K = 5$. The first and last columns are the cloudy and cloud-free reference images. The results with varied values of K are shown in the second to sixth columns. The second row is the magnified view of the red box in the first row. As seen from the figure, the image texture becomes clearer and sharper from left to right, and the results of the K numbers of 4 and 5 are visually the same. To reduce the time consumption, we adopt $K = 4$ in the attentive recurrent network of the proposed AMGAN-CR in the experiment. The average PSNR and SSIM values calculated with different K are listed in Table 9. When $K = 4$, the model has the highest PSNR value except for band 3, where its value is very close to the highest one, and the SSIM values are the maximum as well.

4.3.2. Influence of the number of bands N

We select five values of N for the cloudy image with K=4 in the attentive recurrent network to comprehend the impact of the number of bands: N=3 stands for using bands 2–4, N=4 stands for using bands 1–4, N=5 stands for using bands 1–5, N=6 stands for using bands 1–6 and N=7 stands for using bands 1–7. Fig. 12 shows an example image produced with N=3 to N=7 from the cloudy image. The second row is the magnified view of the red box in the first row, and the first and last columns are the cloudy and cloud-free reference images, respectively. It can be observed from the second row of the figure that when N=7, the image texture is closest to the cloudless reference image. The average PSNR and SSIM values calculated with different numbers of bands on the real datasets are shown in Table 10. We can easily see that when N=7, the model obtains the highest PSNR and SSIM values except for band 2. Using more bands are generally better, but not exactly. It should be noted that the PSNR values of $N = 4$ and $N = 6$ are slightly lower than the PSNRs of $N = 3$ and $N = 5$, respectively. Similarly, SSIM values reveal the same situation. It is important to learn using all the bands, $N = 7$, is necessary and they provides the best performance.

4.3.3. Influence of $\mathbf{M}$ on attention loss $\mathscr{L}_{Att}$

Two cases are analyzed to determine the influence of $\mathbf{M}$ on the attention loss $\mathscr{L}_{Att}$ for the performance of the AMGAN-CR method. Case 1 (Subtraction): In our experiment, $\mathbf{M}$ is required, by using the cloudy image, to subtract the ground-truth image in the simulation test or the cloud-free reference image and then to clip into a binary map containing values that are either 0 or 1. Therefore, $\mathbf{M}$ is a so-called cloud mask because it is not technically acquired by masking clouds. Case 2 (Detection): The second case of obtaining $\mathbf{M}$ is by exploiting a cloud detection method to calculate a real cloud mask. In our ablation experiment, we choose the Fmask (Zhu et al., 2015) algorithm for comparison. Specifically, a real cloud mask is generated by Fmask in the preprocessing step, and we assign $\mathbf{M}$ as the result, which is a binary map as well. To compare the two cases, the average PSNR, SSIM and RMSE values calculated on the simulated datasets are shown in Table 11. The table shows that the result of calculating $\mathbf{M}$ by subtraction is better than the result by detection. However, different cloud detection methods can lead to different cloud removal results. In our experiment, only the Fmask method is considered thanks to its open-source code developed in various programming languages.

5. Discussion

It should be noted that collection of training and testing datasets is time-consuming and difficult which is the primary issue for the application of deep learning models in remote sensing field. The cloudy and

clear images were paired by selecting the Landsat images with shortest time interval from the same location. Therefore, geometric registration is an important step in this case. Moreover, when dealing with data from different seasons or with longer intervals, the land cover changes between cloudy and reference images may bring negative influence on the training phase. It should keep in mind that"garbage in equals garbage out". The quality of the training image pairs is important, and a large quantity is expected to employ to reduce the effect of the poor samples. With the fast development of remote sensing missions, more image data will be collected and archived, which will ease the problem and improve the training image availability.

Furthermore, we converted the unsigned 16-bit integers to 0–255 by applying a linear 2% stretch on the original reflectance values, which is convenient for displaying the images before and after correction. Normalization is not a necessary step in the removal process. In the ablation experiment, although we compared the influence of mask **M** on the attention loss $\mathscr{L}_{Att}$ and observed that the use of a cloud mask achieved by the Fmask method cannot improve the accuracy of cloud removal, we could not conclude that the subtraction method must be better than the detection method because we did not implement other cloud detection methods.

Although the experiments were carried out on Landsat 8 images only, the proposed method for cloud removal is not sensor dependent. The proposed method is based on generative adversarial networks (GANs) that learn the features of paired cloudy and cloud-free images and correct the target cloudy image using the trained model. The model is trained using the training samples from a particular sensor, then it works better for the data from that sensor. The training and testing images generally have the same spatial resolutions. The performance of the proposed method is affected by the configurations of the model and the training images. We can expect the trained model works for newly launched Landsat 9's OLI-2 images due to their similar imaging characteristics. Other types of satellite data can also be effectively handled as long as the training is conducted with the relevant training samples. It is possible to apply the AMGAN-CR method to other optical sensors (such as MODIS, Sentinel, and Gaofen) by retraining the parameters of the model. If training images have different spatial resolutions, the model should work for new images of different resolution. The proposed method may not work well but should be investigated further. The limitation of the proposed AMGAN-CR is that undercorrection may occur when the cloudy image is covered with thick clouds. Therefore, it is advised to separate the image areas covered with thick clouds and conduct the correction using data replacement techniques.

6. Conclusion

In this paper, we developed attention mechanism-based generative adversarial networks for cloud removal (AMGAN-CR) in Landsat 8 images. Based on the results of the study, the following is a summary of the conclusions: (i) The AMGAN-CR method is able to take cloudy images as input and provide cloudfree output images, which is crucial in the preprocessing of remote sensing images due to the severe effects of clouds all year round. (ii) The attention map produced by the attentive recurrent network of the generator is able to detect the distribution of cloud cover in the input cloudy image. (iii) Both simulated and real cloudy datasets were exploited to verify the usefulness of AMGAN-CR, and the results of the experiments show that the images reconstructed by AMGAN-CR were superior to those of five other traditional and current deep learning based cloud removal methods quantitatively and qualitatively. AMGAN-CR has many advantages. It can extract the spatial-spectral characteristics of clouds by taking advantage of deep learning. In contrast to the traditional model-based methods, it can learn the relationship between cloudy and clear images by fully exploiting the features at image level rather than examining each pixel separately. There is no need to acquire a cloud detection mask separately; instead, spatial attention maps are generated indicating which target areas the network should focus on. In the future, we will work on building a cloud annotation database and paired cloudy and cloud-free image databases that can be used for deep model training and validation. So the proposed algorithm can be tested on larger regions with heterogeneous land surfaces during different seasons. Development of lightweight models to enhance the effectiveness of parameter training is also a valuable future research direction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61901278, Grant 41971300, and Grant 62001303; in part by the Natural Science Foundation of Guangdong Province under Grant 2021A1515011413; in part by the Key Project of Department of Education of Guangdong Province under Grant 2020ZDZX3045; and in part by Shenzhen Scientific Research and Development Funding Program under Grant 20200803152531004, and Grant KQTD20200909113951005.

Finally, we would like to take this opportunity to gratefully thank the Editors and five anonymous reviewers for their outstanding comments and suggestions, which greatly helped us to improve the quality of our manuscript.

References

Chen, S., Chen, X., Chen, J., Jia, P., Cao, X., Liu, C., 2016. An iterative haze optimized transformation for automatic cloud/haze detection of Landsat imagery. IEEE Trans. Geosci. Remote Sens. 54, 2682–2694.

Chen, Y., He, W., Yokoya, N., Huang, T.Z., 2019a. Blind cloud and cloud shadow removal of multitemporal images based on total variation regularized low-rank sparsity decomposition. ISPRS J. Photogramm. Remote Sens. 157, 93–107.

Chen, Y., Tang, L., Yang, X., Fan, R., Bilal, M., Li, Q., 2019b. Thick clouds removal from multitemporal ZY-3 satellite images using deep learning. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 13, 143–153.

Cheng, Q., Shen, H., Zhang, L., Yuan, Q., Zeng, C., 2014. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. ISPRS J. Photogramm. Remote Sens. 92, 54–68.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-Based Models for Speech Recognition arXiv preprint arXiv:1506.07503.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N., 2017. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 48–56.

Gao, J., Yuan, Q., Li, J., Zhang, H., Su, X., 2020. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. Remote Sens. 12, 191.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems, p. 27.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hong, G., Zhang, Y., 2018. Haze removal for new generation optical sensors. Int. J. Remote Sens. 39, 1491–1509.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sens. 9, 95.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1125–1134.

Ji, S., Dai, P., Lu, M., Zhang, Y., 2020. Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 59 (1), 732–748.

Ju, J., Roy, D.P., 2008. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sens. Environ. 112, 1196–1211.

Kennedy, R.E., Cohen, W.B., Schroeder, T.A., 2007. Trajectory-based change detection for automated characterization of forest disturbance dynamics. Remote Sens. Environ. 110, 370–386.

Li, W., Li, Y., Chen, D., Chan, J.C.W., 2019a. Thin cloud removal with residual symmetrical concatenation network. ISPRS J. Photogramm. Remote Sens. 153, 137–150.

Li, Z., Shen, H., Cheng, Q., Li, W., Zhang, L., 2019b. Thick cloud removal in high-resolution satellite images using stepwise radiometric adjustment and residual correction. Remote Sens. 11, 1925.
Li, J., Wu, Z., Hu, Z., Zhang, J., Li, M., Mo, L., Molinier, M., 2020. Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion. ISPRS J. Photogramm. Remote Sens. 166, 373–389.
Lin, C.H., Tsai, P.H., Lai, K.H., Chen, J.Y., 2012. Cloud removal from multitemporal satellite images using information cloning. IEEE Trans. Geosci. Remote Sens. 51, 232–241.
Lin, C.H., Lai, K.H., Chen, Z.B., Chen, J.Y., 2013. Patch-based information reconstruction of cloud-contaminated multitemporal images. IEEE Trans. Geosci. Remote Sens. 52, 163–174.
Liu, Q., Gao, X., He, L., Lu, W., 2017. Haze removal for a single visible remote sensing image. Signal Process. 137, 33–43.
Lv, H., Wang, Y., Shen, Y., 2016. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. Remote Sens. Environ. 179, 183–195.
Maalouf, A., Carré, P., Augereau, B., Fernandez-Maloigne, C., 2009. A bandelet-based inpainting technique for clouds removal from remotely sensed images. IEEE Trans. Geosci. Remote Sens. 47, 2363–2371.
Mao, X., Shen, C., Yang, Y.B., 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. Adv. Neural Inf. Proces. Syst. 2802–2810.
Meraner, A., Ebel, P., Zhu, X.X., Schmitt, M., 2020. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J. Photogramm. Remote Sens. 166, 333–346.
Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets arXiv preprint arXiv:1411.1784.
Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., Lymburner, L., McIntyre, A., Tan, P., Curnow, S., et al., 2016. Water observations from space: mapping surface water from 25 years of landsat imagery across Australia. Remote Sens. Environ. 174, 341–352.
Nash, John F., 1950. Equilibrium points in n-person games. Proceedings of the National Academy of Sciences 36 (1), 48–49.
Pan, H., 2020. Cloud Removal for Remote Sensing Imagery via Spatial Attention Generative Adversarial Network arXiv preprint arXiv:2009.13015.
Pan, X., Xie, F., Jiang, Z., Yin, J., 2015. Haze removal for a single remote sensing image based on deformed haze imaging model. IEEE Signal Process. Lett. 22, 1806–1810.
Qin, M., Xie, F., Li, W., Shi, Z., Zhang, H., 2018. Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 11, 1645–1655.
Singh, P., Komodakis, N., 2018. 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. IEEE, pp. 1772–1775.
Tedlek, D., Kasetkasem, T., Khocmboonn, S., Kumazawa, I., Chanvimaluang, T., 2018. A cloud removal algorithm based on a level-set method: Case study multitemporal Landsat 8 OLI images. In: 2018 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE). IEEE, pp. 160–163.
Wang, J., Olsen, P.A., Conn, A.R., Lozano, A.C., 2016. Removing clouds and recovering ground observations in satellite image sequences via temporally contiguous robust matrix completion. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2754–2763.
Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W., 2019. Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12270–12279.
Xu, M., Pickering, M., Plaza, A.J., Jia, X., 2015. Thin cloud removal based on signal transmission principles and spectral mixture analysis. IEEE Trans. Geosci. Remote Sens. 54, 1659–1669.
Xu, M., Jia, X., Pickering, M., Plaza, A.J., 2016. Cloud removal based on sparse representation via multitemporal dictionary learning. IEEE Trans. Geosci. Remote Sens. 54, 2998–3006.
Xu, M., Jia, X., Pickering, M., Jia, S., 2019. Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. ISPRS J. Photogramm. Remote Sens. 149, 215–225.
Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., Dai, Q., 2019. STAT: spatial-temporal attention mechanism for video captioning. IEEE Trans. Multimedia 22, 229–241.
Zhang, Y., Guindon, B., Cihlar, J., 2002. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. Remote Sens. Environ. 82, 173–187.
Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L., 2017. Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. 26, 3142–3155.
Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56, 4274–4288.
Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S., 2017. Diversified visual attention networks for fine-grained object classification. IEEE Trans. Multimedia 19, 1245–1256.
Zhou, B., Wang, Y., 2019. A thin-cloud removal approach combining the cirrus band and RTM-based algorithm for Landsat-8 OLI data. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1434–1437.
Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277.