ORIGINAL RESEARCH PAPER

MSAR-DefogNet: Lightweight cloud removal network for high resolution remote sensing images based on multi scale convolution

Ying Zhou1,2 | Weipeng Jing1,2 | Jian Wang3 | Guangsheng Chen1,2 | Rafal Scherer4 | Robertas Damaševičius5

1 Northeast Forestry University, Xiangfang District, Harbin, People’s Republic of China
2 Key Laboratory of Forestry Data Science and Cloud Computing of State Forestry Administration, Xiangfang District, Harbin, People’s Republic of China
3 Aerospace Information Research Institute Chinese Academy of Sciences, Haidian District, Beijing, People’s Republic of China
4 Department of Intelligent Computer Systems, Czestochowa University of Technology, Czestochowa, Poland
5 Silesian University of Technology, Gliwice, Silesia, Poland

Correspondence
Weipeng Jing, Northeast Forestry University, Xiangfang District, Harbin, People’s Republic of China.
Email: jwp@nefu.edu.cn

Funding information
Heilongjiang Province Applied Technology Research and Development Program Major Project, Grant/Award Numbers: GA18B301, GA20A301; Fundamental Research Funds for the Central Universities, Grant/Award Number: 2572017Z09; China State Forestry Administration Forestry Industry Public Welfare Project, Grant/Award Number: 201504307; National Natural Science Foundation of China, Grant/Award Number: 31770768

Abstract
High resolution remote sensing image cloud removal can bring a lot of convenience for human activities. However, the existing cloud removal algorithms have a variety of disadvantages. First of all, they have the disadvantages of long computing time and large consumption of computing resources. Secondly, the effect of recovery needs to be improved. In order to improve the above two points, a near real-time effective algorithm is proposed, namely MSAR-Defognet (multiple scale attention residual network using for cloud remove), which consumes less computing power and space and has superior cloud removal effect. On the one hand, several different large-scale filters are chosen to extract the weak information effectively, while can save the computing power and shorten the image processing time. On the other hand, the fine-grained convolution residual block with channel attention mechanism is used to enhance the network’s ability to extract cloud features. In addition, a data set which is closer to the real cloud shape and has higher richness to train the cloud removal network, so that the parameters obtained by training have stronger robustness and can adaptively remove clouds with different thickness. Experiments show that, compared with other advanced network models, the network not only has the advantage of fast processing speed, but also has better image restoration effect in high-resolution remote sensing image restoration. It can meet the requirements of many hard real-time tasks, so that remote sensing images can play a greater value for human activities.

1 | INTRODUCTION

After nearly 60 years of rapid development, remote sensing image has gradually penetrated into many fields of human production and life, and plays an important role in resource survey [1], dynamic monitoring of urban development [2], agriculture and forestry vegetation classification [3] and disaster detection [4], military command [5] and so on. However, there is a lot of cloud noise in remote sensing image, which greatly reduces the utilisation of remote sensing image and increases the cost of remote sensing technology. Therefore, removing cloud noise in remote sensing image with cloud and obtaining real land surface information is a topic that many scholars have been studying.

Recently, most common image defogging algorithms are proposed for a single outdoor foggy or indoor smoke image and have gotten a large quantity of achievements. Taken together, the methods for removing cloud and fog include defogging algorithms based on image processing, image restoration methods based on atmospheric physical models, and methods based on deep learning.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. IET Image Processing published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology
The dehazing algorithms based on image processing mostly rely on colour constancy and brightness constancy. The principle of this kind of algorithm is relatively simple and mature. Through the processing of this kind of algorithm, the contrast and saturation of the image are increased. So that people can extract the valuable information in the image more clearly, while this kind of method is not in the true sense of cloud and fog. The more representative methods of this kind are histogram equalisation [6], logarithmic transformation, power law transformation, sharpening, wavelet transformation [7], homomorphic filtering [8] etc. This kind of method mostly considers the imaging model of the image too little, and cannot get very satisfactory results when the image information is more complex.

The image restoration method based on the atmospheric physical model uses the atmospheric transmission model to recover clear images by solving the inverse process of the physical process of image degradation, belongs to the category of physical restoration. This type of algorithm starts with the cause of fog, builds a model of image degradation with known or partially known prior knowledge, and then restores the ideal image through the inverse process of the physical model. Among the more outstanding researches, He [9] proposed a dark channel defogging method in 2009, but this method often brings colour distortion and halo phenomenon. In order to solve these problems, Liu et al. [10] combined image depth information and guided filtering to improve the defogging effect of dark channel prior algorithm. In the field of remote sensing, Pan et al. [11] noted that the average intensity of the dark channel in remote sensing images is low, but not close to zero. Therefore, they added a constant term to the image degradation model for haze removal to better complete the cloud and fog removal task for remote sensing images. This kind of method usually needs to solve nonconvex optimisation problems, and the amount of calculation is large. In addition, in practical application, when the assumed prior knowledge is easy to be inconsistent with the actual situation, it may lead to inaccurate transmission approximation. So it is unable to perform the task of cloud removal well [9].

With the rapid development of convolutional neural networks, end-to-end deep learning neural network models have emerged to solve the problems of the above two methods. It is understood that the Dehazenet proposed by Cai et al. [12] in 2016 is the first neural network architecture for smoke removal. However, the haze in the image has obvious spatial variability. What is more, the complex nonlinear relationship between haze propagation and image spectral spatial information leading to the shallow neural network does not perform well when remove cloud from remote sensing images with huge amount of information [13]. Zhang and Patel [14] proposed to estimate the transmission map through a densely connected pyramid network, while estimating atmospheric light through the U-net [15]. However, when the estimated transmission map and atmospheric light are not accurate, these methods may also cause large deviation. To solve this problem, some end-to-end methods do not estimate the transmission map or atmospheric light [16–18], but learn to directly restore the clean image. Ren et al. [17] proposed a gated fusion network by fusing three images from the original fuzzy input (for example, white balance, contrast enhancement, and gamma correction). These algorithms are not suitable to remove the fog with uneven thickness. Thus, the FFA-net [19] proposed by Qin et al. in 2020 combines a novel feature attention module and a feature extraction module, which can flexibly allocate attention and process different types of information. Liu et al. [20] proposed a new multi-scale convolution network GridDehazeNet with pre-training module based on attention mechanism to rich data diversity and improve the effect of single image denoising. There are small quantity of neural network models specifically for defogging remote sensing images. Among them, Jiang et al. [21] and Qin [22] respectively proposed a convolution neural network based on residual structure for high-resolution remote sensing images and multi-spectral remote sensing images. Inspired by the super-resolution algorithm, Dong et al. [23] proposed a dense feature fusion module in u-net structure to enhance the function of restoring haze image. Shao et al. [24] proposed a domain adaptation parameter by sharing weight parameters in GAN to solve the problem that the existing methods perform well on synthetic data but not on real images. It can show superior effect on both synthetic data and real data.

Although many machine learning algorithms have achieved excellent results in improving PSNR (peak signal to noise ratio), SSIM (structural similarity) and other full reference image quality evaluation indicators. There are still some problems to be solved in this field. On the one hand, the images processed by deep learning will also have the problems of halo and low accuracy. On the other hand, they effectively improve the ability of the model to remove fog in different focuses. Without exception, these network systems are very heavy, and it takes too long to process a single image, which is difficult to meet the requirements of some special tasks.

In order to make improvements in the above two aspects, we propose a new end-to-end multi-scale residual neural network model MSAR-DefogNet for single remote sensing image cloud removal. After testing, the MSAR-DefogNet we proposed has shown superior cloud removal capabilities for high-resolution remote sensing images on both public data sets and self-made data sets. At the same time, it has lower memory requirements, lower hardware requirements and faster speed.

Specifically, our contributions are as follows:

(1) According to research, there are not many data sets related to remote sensing image de-clouding. The weight trained with the previous open source data set cannot be used to remove the cloud from remote sensing images with uneven cloud distribution. In order to deal with this problem, we fully considered the diversity of cloudy remote sensing images, and proposed a new data set. Experiments have proved that the weight trained with the new data set has stronger robustness and adaptability, and has been greatly improved in SSIM and PSNR indicators.

(2) Different from the large-scale deep neural network in the past, our network greatly reduces the computational cost and memory overhead. Naturally it also greatly improves
(3) We found that the pixel attention mechanism is not suitable for connection in multi-scale context convolution. We integrate the channel attention mechanism in the residual block so that our network can better represent learning.

The remaining work is summarised as follows: Section 2 introduces the details of the proposed method. Section 3 presents the experimental results and compares it with the existing advanced dehazing methods. In addition, we have also analysed and compared the computing power, memory and computing time required for various deep learning models to run. Finally, we summarise the work in Section 4.

2 | PROPOSED METHOD

2.1 | Overall introduction

Because the useful information in the image affected by cloud noise is weak information, we must use deeper neural network to extract this kind of surface information. The residual neural network [25] structure proposed by He et al. in 2016 solves the problem of gradient dispersion that occurs in deep neural networks when the number of network layers increases by connecting different layers. In short, the principle of residual connection is to add a shortcut connection between the input and output of a layer, and judge whether to add the layer in the network according to the result. The specific principle of residual network is shown in the following equation:

$$y = F(x, W_f) + W_f'x$$  \hspace{1cm} (1)

Here \(x\) and \(y\) are the input and output vectors of the layers considered. The function \(F(x, W_f)\) represents the residual mapping to be learned. \(W_f'\) is a linear projection, which ensures that the dimensions of \(X\) and \(F\) are equal, and \(W_f'\) is the weight parameter of the layer.

Our MSAR-DefogNet also uses the idea of residual neural network. The model consists of pre-treatment module, feature extraction module, multi-scale context convolution module and thinning module. We add the residual connection structure between each delicate feature block and each multi-scale convolution block to connect the information obtained by different convolution cores and achieve more efficient information fusion. The specific information of the structure is shown in Figure 1. The remaining connections are dotted orange lines in Figure 1.

2.2 | Delicate feature module

The specific structure is shown in Figure 2. Before that, we use a 3 x 3 convolution kernel to convert the 3 channel information into 64 channel information, which is used to map the RGB channel information in high-resolution remote sensing image to a higher dimensional space, and complete the preliminary feature extraction. At the same time, it provides the basis for the use of subsequent channel attention module and multi-scale feature extraction module.

In the Delicate Feature module, we first use the convolution with kernel size of 3 to extract more fine-grained features. The Relu activation function is added between the two convolutions. This nonlinear factors are added to improve the expression ability of the neural network to the model, so as to solve the problems that the linear model cannot solve.

The attention mechanism imitates the selective visual attention mechanism of human beings, that is, human beings scan the whole situation quickly to obtain the areas that need to be focused on, which is commonly known as the focus of attention. Then more attention resources are put into this area to obtain more detailed information of the target, and suppress other useless information. In denoising tasks, some works will apply
channel attention mechanism and pixel attention mechanism. For one thing, the pixel attention module selectively aggregates the features of each location and allocates different attention resources for different locations by weighted sum of the features of all locations. For another, in the research of dark channel prior algorithm, we will find that the intensity of some pixels in different channels is different, which indicates that each channel needs to be allocated specific weight. Therefore, channel attention selectively emphasizes the interdependent channel mapping by integrating the relevant features among all channel mappings. Experimental results show that no matter what kind of multi-scale convolution algorithm combined with the pixel attention algorithm, the network effect will become worse, which may be due to the pixel attention mechanism for each pixel redistribution affects the transmission of information between different scales. In the end, we only use the channel attention mechanism to balance the attention allocation between different channels.

The channel attention mechanism model is shown in Figure 2. Because different channels are affected by the cloud to different degrees, the global average pool layer is first used to convert global spatial information into channel descriptors. See Equation (2) for details:

\[ g_c = A_p (F_i) = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} x_c(i,j) \]  

(2)

where \( A_p \) represents the average pool operation, \( F_i \) represents the original input information, \( H \) and \( W \) are the resolution of the image, and \( x_c(i,j) \) represents the value of the \( c \)th dimension at the position \((i,j)\).

First go through a convolution layer with a convolution kernel size of 1, change the number of channels to 1/8 of the original, merge different channel information. And then connect a convolution layer with a convolution kernel size of 1 expanding the number of channels to the size of the original input. After two convolutions, the Relu activation function layer and the Sigmod activation function layer are added to obtain the weights of different channels respectively. And then multiply it with the original input to get the channel attention distribution weight coefficient matrix of the current image. This process can be described as Equation (3):

\[ F_i^* = \sigma(C(\sigma(C(g))))) \times F_i \]  

(3)

Where \( F_i^* \) represents the result obtained through the channel attention module, \( C \) represents the convolution layer, \( \sigma \) is the activation function, \( F_i \) stands for the original input.

### 2.3 Muti-scale block

The most notable feature of remote sensing images affected by clouds is cloud pollution with different thicknesses in different regions. However, the pixel information contaminated by the cloud is very close to the surrounding pixel information. Extracting features of different scales and fusing these features can well complete the processing of the current pixel. The acquisition of features at different scales depends on the receptive field of the convolutional neural network. There are two ways to obtain large-scale receptive fields: one is to increase the depth of the neural network, and the other is to expand the convolution kernel. Increasing the depth of the network will greatly increase the computational consumption and the effect is not good. The size of the receptive field is also a key concept in the extraction of target features by convolutional neural networks. If the receptive field is too small, only local features can be observed. If the receptive field is too large, a lot of invalid information can be obtained. We selected reasonable convolution kernels of different sizes to extract features of different sizes, and fuse multi-scale features.

As shown in Figure 1, the size of convolution kernel is 11, 9 and 7 in coarse scale convolution block, 7, 5 and 3 in fine scale convolution block, convolution layer with 3 kernel size is adopted in equal scale convolution block. Instead of other complex upsampling operations, we choose to fill the boundary with 0 to ensure that the image size after multi-scale convolution remains unchanged. The number of fillings is determined according to Equation (4).

\[ F_o = \frac{F_{in} + 2 \times p - k}{s} + 1 \]  

(4)
Where $F_0$ represents the output image size, $F_{in}$ represents the input image size, $p$ represents the number of zeros to be filled, $k$ represents the size of convolution kernel, and $s$ represents the step size.

In addition, in each scale convolution block, we do 64-16-64 processing for the dimension. This operation not only reduces the amount of calculation, but also facilitates multi-dimensional information fusion. The residual connection structure represented by the orange dotted line in Figure 1. So that it can effectively learn more extensive, powerful and abstract information by fusing features of three different scales.

The module can also have a strong enhancement effect when used alone. However, due to the high-resolution characteristics of remote sensing images, the image processed by the multi-scale convolution operation is prone to patch effects. In order to solve this problem, we splice the residual blocks before multi-scale convolution.

### 2.4 Loss function

The most common loss function is $L_1$ loss function and perceived loss function. $L_1$ loss is the mean square error (MSE) between two images, which is specifically described as follows:

$$L_1 = ||\text{DefogNet}(I) - R||^2$$  \hspace{1cm} (5)

Among them, the $I$ represents the blurred image, DefogNet($I$) represents the clear image generated after processing, and $R$ represents the clear image in the original data set.

$L_1$ loss function is a pixel by pixel comparison loss function, that is, the distance between the output image and the real image in each pixel colour space, so as to minimise the distance. In fact, there are always drawbacks in this method. For example, two images are basically the same, the difference of each pixel between the two images is 1. The basic visual effect and image structure are consistent, but the calculated loss function is relatively large. Therefore, we use the perceptual loss function (feature reconstruction loss function) to obtain the difference by pre-training the high-level image features extracted from CNN. The generated image and the target image pass through a non-trainable VGG-19 network, using the output features of the first, second, and third layers of the VGG-19 network as a measurement to determine the loss function. According to the overall loss of content and style to better dynamically plan the training process. The equation is as follows:

$$L_p = \sum_{i=1}^{3} \frac{1}{C_iH_iW_i} ||\text{VGG}_i(R) - \text{VGG}_i(\text{DefogNet}(I))||^2$$ \hspace{1cm} (6)

Where $I$ is the $i$th layer of the network, and $C_i$, $H_i$ and $W_i$ are the channel number, height and width of the image respectively. $I$ is the blurred image, and $R$ is the clear image for comparison in the dataset.

We combine $L_1$ and $L_p$ to guide the recovery process of the network. Summarising the previous research, we found that the size of $\lambda$ is not fixed. As long as it is within a certain range, the deep neural network can adapt to the value of $\lambda$ to achieve the desired effect. Therefore, we imitate the state-of-the-art model FFA-Net [17]. $\lambda$ is set to 0.5, and the total loss $L_s$ is expressed as follows:

$$L_s = L_1 + \lambda L_p$$ \hspace{1cm} (7)

### 3 EXPERIMENTS

#### 3.1 Experimental environment

Our model is deployed on the GPU of Tesla V100, and the basic framework is implemented using Pytorch 1.1.0 and python 3.7. The total number of training algorithms is $1 \times 10^4$, and verification is performed every 200 steps, using the Adam optimiser, where $\beta 1$ and $\beta 2$ respectively take the default values of 0.9 and 0.999. The initial learning rate $\eta$ is set to $1 \times 10^{-4}$, and the cosine annealing strategy [25] is used to complete the learning rate change during the training process. $T$ is the total number of batches. When running to batch $t$, the learning rate $\eta_t$ is and is computed as:

$$\eta_t = \frac{1}{2} \left( 1 + \cos \left( \frac{\pi t}{T} \right) \right) \eta$$ \hspace{1cm} (8)

#### 3.2 Dataset

Many image restoration methods are data-driven, that is to say, we must provide a pair of remote sensing images with and without clouds to guide the network to learn a set of weight coefficients that can effectively identify and remove clouds. Therefore, the image pairs used for training play an important role in this process. However, the number of publicly available data sets for removing clouds from high-resolution remote sensing images is not large. Lin et al. [27] proposed a data set for removing clouds from remote sensing images. The data set is divided into two parts: RICE1 and RICE2. All the images are 512 $\times$ 512 in size. RICE1 uses the cloud settings on Google Earth to obtain 500 pairs of images, and RICE2 collects 753 pairs of images from Landsat 8 OLI/TIRS by setting a certain time interval. But this data set didn’t consider the cloud shape and data richness. As shown in Figure 3(a), the clouds and fog distribution are relatively uniform and lack cloud forms with uneven thickness. The ground information under the cloud cover in Figure 3(b) is too simple to effectively evaluate the advantages and disadvantages of cloud removal model algorithm. The cloud in Figure 3(c) is too heavy and erases all available information on the ground. It is not applicable to the cloud removal model. To avoid these shortcomings, we also imitate the method of RICE1 to select a certain time point in the historical image of Google Earth in a certain scene by setting the cloud layer and not setting the cloud layer to obtain a pair of
cloud-free and cloud-free RGB images that are exactly the same. We have selected 400 pairs of data in total, and the resolution size of each image is $1317 \times 727$. We named this data set PRSC (paired remote sensing images with clouds). In the process of making the data set, we not only keep the beneficial shape of the original dataset, but also fully consider the various forms of fog. As shown in Figure 3(d–f), there are local clouds, thick clouds and white areas in the ground information that are easy to be confused. These rich and real scenes can make our data set more conducive to application to real scenes and have stronger adaptability.

In order to verify the validity of our proposed data set, we use the data set RICE to compare with our data set PRSC. We use the MSAR-DefogNet and FFA-Net and GridDehazeNet to

| Cloud Image | Ground Truth | FFA-Net | GridDehazeNet | ours |
|-------------|--------------|---------|---------------|------|
| RICE        | ![Image](image1) | ![Image](image2) | ![Image](image3) | ![Image](image4) |
| PRSC        | ![Image](image5) | ![Image](image6) | ![Image](image7) | ![Image](image8) |

**FIGURE 3** Samples about two datasets

**FIGURE 4** Experimental results on different datasets
Evaluating indicator

Our network has two goals: one goal is to remove clouds in high-resolution remote sensing images with cloud noise to improve the clarity of the picture. Another goal is to reduce the complexity of the deep learning framework to perform image enhancement on high-resolution images in real time.

For the first goal, we use PSNR (peak signal-to-noise ratio) and SSIM (Structural Similarity) indicators to evaluate the pros and cons of cloud removal algorithms. For the second goal, we use FLOPs (floating point operations) to evaluate the computational power required for forward propagation of deep learning models, which reflects the level of hardware (GPU) performance requirements. We use the number of parameters (parameters) to describe the memory required by the model, and use the time required to process a picture to intuitively evaluate the real-time nature of our model.

PSNR is a full-reference image quality evaluation index, which can effectively reflect the cloud removal effect of the image. The larger the value, the better the cloud removal effect.

\[
\text{PSNR} = 10 \log_{10} \left( \frac{2^8 - 1)^2}{\text{MSE}} \right)
\]

\[
\text{MSE} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X(i, j) - Y(i, j))^2
\]

Where MSE is the mean square error of the current image \(X\) and the reference image \(Y\), \(H\) and \(W\) are the size of the image, and \(v\) is the number of bits per pixel, which is generally taken as 8, that is, the gray scale of the pixel is 256.

Because the simple calculation based on difference profile does not conform to the evaluation results of human visual system (HVS). We try to use objective image quality assessment to evaluate the image. SSIM (structural similarity) measures image similarity from brightness, contrast and structure. The larger the SSIM value is, the more obvious the cloud removal effect is.

\[
\text{SSIM} = \frac{(2\mu_x \mu_y + \epsilon_1)(2\sigma_{xy} + \epsilon_2)}{(\mu_x^2 + \mu_y^2 + \epsilon_1)(\sigma_x^2 + \sigma_y^2 + \epsilon_2)}
\]

Where the value of \(\mu_x, \mu_y\) represents the average brightness of the image. \(\epsilon_1\) and \(\epsilon_2\) are constant terms. When the mean value is close to 0, adding these two constant terms can avoid the denominator being 0. Take \(\mu_x\) as an example

\[
\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

The value of \(\sigma_x\) and \(\sigma_y\) indicates the contrast of the image, that is, the sharpness of the change in the brightness of the image, which is the standard deviation of the pixel value. Take \(\sigma_x\) as an example, the calculation method is shown in Equation (13):

\[
\sigma_x = \left( \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2}
\]

\(\sigma_{xy}\) is the covariance of \(x\) and \(y\), which is shown in Equation (14):

\[
\sigma_{xy} = \left( \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \right)
\]

Control experiment

In order to verify the efficiency of multi-scale context convolution in cloud removal tasks. We only use multi-scale convolution module for cloud removal and only use the same size of scale-free change module for cloud removal. The experimental results are shown in Figure 5, which can prove that multi-scale convolution has better defogging ability under the condition of consuming the same computing power. However, it is not difficult to find the shortcomings of this method. When it comes to large colour areas such as ocean, multi-scale convolution has obvious patch effect due to its coarse-grained characteristics. In addition, when the cloud is thick, the structure cannot extract the weak information from the image. Therefore, we add the delicate feature module before this module to formed our model.

The specific experimental process is shown in Algorithm 1. In this section we compare the cloud removal ability of our MSAR-DefogNet and the most advanced image denoising algorithms from qualitative and quantitative aspects. Based on the new data set, we have done experiments on DCP [8], FFANet [18], GridDehazeNet [19] and MSAR-DefogNet. Since the DCP algorithm doesn’t need to be trained, the DCP algorithm can be used directly to process the image. The FFANet, GridDehazeNet and MSAR-DefogNet algorithms are deep neural
TABLE 1  Recovery effect of each algorithm

| Parameters | DCP | FFA-Net | DehazeNet | Ours |
|------------|-----|---------|-----------|------|
| PSNR       | 19.06 | 21.48   | 20.28     | 21.29|
| SSIM       | 0.77  | 0.83    | 0.82      | 0.85 |

TABLE 2  Calculation power consumption of each algorithm

| FLOPs | FFA-Net | GridDehazeNet | ours |
|-------|---------|---------------|------|
| Parameters | 219.19 G | 16.50 G | 31.26 G |
| Time   | 5.89s   | 1.21 s | 1.21 s |

ALGORITHM 1  Training strategy

Require: Training set \((X, Y)\), validation set \((X_v, Y_v)\), Kaiming initialize weight \(w\), validation cycle \(eval\_step\), total step \(epochs\)

1: for \(t = 1\) to \(epochs\) do
2: Get training sample \((I, I^*)\) from \((X, Y)\)
3: Compute \(I_o = MSAR - DefogNet_w(I)\)
4: Optimise learning rate \(\eta_t\) by \{Equation (8)\}
5: Update Weight \(w\) by \(\min(L_s)\)
6: if \(t \% eval\_step = 0\) then
7: get validating sample \((I_v, I_v^*)\) from \((X_v, Y_v)\)
8: Compute \(I_{vo} = MSAR - DefogNet_w(I_v)\)
9: Compute PSNR\((I_{vo}, I_v^*)\) by \{Equation (9)\}
10: Compute SSIM\((I_{vo}, I_v^*)\) by \{Equation (11)\}
11: if PSNR > MAX-PSNR and SSIM > MAX - SSIM then
12: Save weight
13: end if
14: end if
15: end for

Ensure: MSAR = DefogNet\(_w\)

network models that need to be trained. We divide the PRSC dataset into train set, validation set and test set according to the ratio of 3:1:1. In the training process, the training set is used to adjust the parameters of the network, and the validation set is used to determine which parameter is ultimately retained. After training, the test set is used to evaluate the effects of different models. The experimental results are shown in Figure 6. The cloud removal ability test results on different models are shown in Table 1. It can be seen from the Figure 6 that because the imaging distance of the remote sensing image is different from that of the natural image, the colour of the remote sensing image restored by the DCP algorithm is darker, causing serious colour shift. This method is not suitable for the cloud removal task of high-resolution remote sensing image. Next, from Figure 6, we can see that the recovery degree of GridDehazeNet in detail is relatively insufficient. Especially when the cloud concentration is too heavy, the restoration of the model is clear and the image ability is greatly reduced. In contrast, our algorithm and FFA-Net algorithm can completely retain the details of the image, and the image restored by FFA-Net often appears darker. Ours is more in line with the imaging characteristics of remote sensing image. In addition, as shown in the third group of pictures in Figure 6, when there are white interfering objects similar to clouds on the ground, the image restored by our model will not be as blurred as GridDehazeNet, nor will it be removed after the white objects on the ground are mistakenly identified as cloud noise as FFA-Net. Our algorithm and FFA-net algorithm have achieved good results in PSNR and SSIM indicators. In terms of PSNR indicators, it is 0.19 lower than FFA-Net overall. In terms of SSIM indicators, our MSAR-DefogNet has achieved the best results, which is better than FFA-Net 0.02 on average.

For the comparison of calculation of consumption cost, see Table 2:

The proposed MSAR-DefogNet is slightly higher than GridDehazeNet in terms of flops parameters, i.e. computational
power. But it has the lowest consumption in terms of parameters and time required to recover a single image. Compared with FFA-Net, which can recover more details from high-resolution remote sensing images with cloud noise. The computational complexity of our proposed MSAR-DefogNet is about 1/7 of that of FFA-Net, and the total number of parameters required is about 1/8 of that of FFA-Net. This is because our multi-scale convolution network makes full use of the context features in the image, which cannot only effectively extract the cloud features and restore the image details, but also process the clouds with different thickness to different degrees. We only need to connect a small number of residual blocks to refine the image restoration effect without significantly increasing the size of the network. Without losing the effect of image restoration, our network needs less computing resources. It is not only very friendly to training, but also can greatly shorten the processing time.

The imaging principles of remote sensing optical instruments and ordinary cameras are basically the same. For RGB format images, the main difference between the two is the imaging distance. Remote sensing image passes through many atmospheric media in the process of imaging. Atmospheric molecules can scatter, refract and reflect. This makes them different in visual effect. The proposed algorithm takes into account the differences between different images, and achieves a balance between the effectiveness of cloud removal and the saving of computing power according to the characteristics of remote sensing images. It is very conducive to the migration to the devices with limited memory and computing resources. The processing speed is very fast and can meet the needs of many real-time tasks.

Although our newly proposed MSAR-DefogNet network has achieved good results, it is not difficult to see that our model still has a lot of room for improvement. As shown in the first row in Figure 6, when this type of feature is involved, the image recovered by our model still has a large colour shift, and the detail clarity of the feature needs to be further improved. Moreover, there is still a lot of room for improvement in terms of computational power consumption and time consumption of the model.

In the later stage, the research goal can be focused on improving the clarity of the image by optimising the network structure, and the network can be further lightweighted through solutions such as knowledge distillation and the use of lightweight networks.

4 | CONCLUSION

In this paper, we introduce a new MSAR-DefogNet for remote sensing image cloud removal. The network consists of multi-scale convolution and residual block with channel attention mechanism. It makes full use of the multi-scale convolution operation from coarse to fine to capture the fog details of the fog image and the law of haze space change from the whole to the local. It uses the minimum memory cost to complete the initial defogging operation of the image. We also concatenate the residual blocks before the multi-scale module to extract detailed features to solve the problem of patch effect. In addition, we provide a new data set, including more forms of cloud, which can improve the adaptability and robustness of the network. Experimental results show that compared with other networks, this network has stronger feature extraction ability for cloud high-resolution remote sensing images, and only needs a small amount of processing time, memory overhead and computing cost.

ACKNOWLEDGMENTS

The work described in this paper is supported by National Natural Science Foundation of China (31770768), Fundamental Research Funds for the Central Universities(2572017PZ04), Heilongjiang Province Applied Technology Research and Development Program Major Project(GA18B301,GAA20A301) and China State Forestry Administration Forestry Industry Public Welfare Project (201504307).
REFERENCES

1. Wan, Y.Q., Fan, Y.H., Jin, M.S.: Application of hyperspectral remote sensing for supplementary investigation of polymetallic deposits in huanunshan ore region, northwestern china. Scientific Reports 11(1), 440 (2021)
2. Judith, C., et al.: Using high-resolution remote sensing data for habitat suitability models of bromeliaceae in the city of mérida, venezuela. Landscapdee and Urban Planning 120, 107–118 (2013)
3. Wei, W., et al.: Study on remote sensing image vegetation classification method based on decision tree classifier. In: 2018 IEEE Symposium Series on Computational Intelligence, pp. 2292–2297.IEEE, Piscataway, NJ (2018)
4. Mohammed, A.A., Khamees, H.T.: Categorizing and measurement satellite image processing of fire in the forest greece using remote sensing. Indonesian Journal of Electrical Engineering and Computer Science 21(2), 843–853 (2021)
5. Benz, U.C., et al.: Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. ISPRS J. Photogramm. Remote Sens. 58(3-4), 239–258 (2004)
6. Stark, A.J.: Adaptive image contrast enhancement using generalizations of histogram equalization. IEEE Trans. Image Process. 9(5), 889–896 (2000)
7. Dippel, S., et al.: Multiscale contrast enhancement for radiographies: Laplacian pyramid versus fast wavelet transform. IEEE Trans. Med. Imaging 21(4), 343–53 (2002)
8. Ekstrom M. Digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing 28, (4), 484–486 (1980). http://doi.org/10.1109/tassp.1980.1163457
9. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis Machine Intelligence 33(12), 2341–2353 (2011)
10. Liu, X., et al.: Scene-adaptive single image dehazing via opening dark channel model. IET Image Processing 10(11), 877–884, (2017)
11. Pan, X., et al.: Haze removal for a single remote sensing image based on deformed haze imaging model. IEEE Signal Process. Lett. 22(10), 1806–1810 (2015)
12. Cai, B., et al.: Dehazenet: An end-to-end system for single image haze removal. IEEE Trans. Image Process. 25(11), 5187–5198 (2016)
13. Ghamisi, P., Chen, Y., Zhu, X.X.: A self-improving convolution neural network for the classification of hyperspectral data. IEEE Geoscience and Remote Sensing Letters 13(10), 1537–1541 (2017)
14. Zhang, H., Patel, VM.: Densely connected pyramid dehazing network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3194–3203.IEEE, Piscataway, NJ (2018)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.Springer, Berlin (2015)
16. Li, B., et al.: AOD-Net: All-in-one dehazing network. In: 2017 IEEE International Conference on Computer Vision, pp. 4780–4788.IEEE, Piscataway, NJ (2017)
17. Ren, W., et al.: Gated fusion network for single image dehazing. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3253–3261.IEEE, Piscataway, NJ (2018)
18. Li, R., et al.: Single image dehazing via conditional generative adversarial network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8202–8211.IEEE, Piscataway, NJ (2018)
19. Qin, X., et al.: Ffa-net: Feature fusion attention network for single image dehazing. arXiv:1911.07559, 2019
20. Liu, X., et al.: GridDehazeNet: Attention based multi-scale network for image dehazing. In: 2019 IEEE/CVF International Conference on Computer Vision, pp. 7313–7322.IEEE, Piscataway, NJ (2019)
21. Jiang, H., Lu, N.: Multi-scale residual convolutional neural network for haze removal of remote sensing images. Remote Sens. 10(6), 945 (2018)
22. Qin, M., et al.: Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(5), 1645–1655 (2018)
23. Dong, H., et al.: Multi-scale boosted dehazing network with dense feature fusion. In: IEEE Conference on Computer Vision Pattern Recognition, pp. 2154–2164.IEEE, Piscataway, NJ (2020)
24. Shao, Y., et al.: Domain adaptation for image dehazing. In: IEEE Conference on Computer Vision Pattern Recognition, pp. 2805–2814.IEEE, Piscataway, NJ (2020)
25. He, K., et al.: Deep residual learning for image recognition. in IEEE Conference on Computer Vision Pattern Recognition, pp. 770–778.IEEE, Piscataway, NJ (2016)
26. He, T., et al.: Bag of tricks for image classification with convolutional neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558–567.IEEE, Piscataway, NJ (2018)
27. Lin, D., et al.: A remote sensing image dataset for cloud removal. arXiv:1901.00600, (2019)