Multidimensional Systems and Signal Processing (2022) 33:1119–1138
https://doi.org/10.1007/s11045-022-00835-x

TransRA: transformer and residual attention fusion for single remote sensing image dehazing

Pengwei Dong[1] · Bo Wang[1]

Received: 7 November 2021 / Revised: 14 April 2022 / Accepted: 16 April 2022 /
Published online: 20 May 2022

Abstract
Haze seriously reduces the quality of optical remote sensing images, resulting in poor performance in many applications, such as remote sensing image change detection and classification. In recent years, deep learning models have achieved convincing performance in image dehazing, which has attracted more and more attention in haze removal of remote sensing images. However, the existing deep learning-based methods are less able to recover the fine details of remote sensing images that suffered from haze, especially the cases of nonhomogeneous haze. In this paper, we propose a two-branch neural network fused with Transformer and residual attention to dehaze a single remote sensing image. Specifically, our upper branch is a U-shaped encoder–decoder architecture, using an efficient multi-head self-attention Transformer for capturing long-range dependencies. The lower branch is an attention stack of residual channels to enhance fitting capability of models and complement fine-detailed features for upper branch. Finally, the features of the two branches are stacked and mapped to the haze-free remote sensing image by fusion block. Extensive experiments demonstrate that our TransRA achieves superior performance against other dehazing competitors both qualitatively and quantitatively.

Keywords Transformer dehazing · Residual attention · Fusion block · Remote sensing images

1 Introduction

The camera of optical remote sensing images is easily disturbed by clouds, atmospheric humidity, and bad weather, resulting in color distortion and information loss Liu et al. (2011). The area covered by haze in remote sensing images directly affects the analysis and utilization of the image, thus limiting the application of remote sensing images. Therefore, haze removal technology has attracted more and more attention in improving the visibility of opti-

Pengwei Dong and Bo Wang have contributed equally to this work.

✉ Bo Wang
tjuwb@nxu.edu.cn

[1] School of Physics and Electronic-Electrical Engineering, Ningxia University, Yinchuan 750021, China

cal satellites. However, how to achieve haze removal in a single degraded remote sensing image is still a challenging and ill-posed task, because the area covered by haze contains both the ground surfaces and haze components Li and Chen (2020).

The early image dehazing processing methods relied on different assumptions and prior knowledge to solve this problem. Dehazing algorithms can be roughly divided into two categories: image enhancement and image restoration Long et al. (2014). Image enhancement is to modify pixels, including contrast enhancement, histogram equalization, homomorphic filtering, and Retinex theory Kim et al. (2013); Makarau et al. (2014). These enhancement methods can only improve the quality of the image to a certain extent, but cannot substantially remove haze. They treated remote sensing images as a whole and not considered the high-resolution features.

While image restoration employs the imaging principle of hazy images and estimates the latent parameters of atmospheric scattering model (ASM) Oakley and Satherley (1998) to restore clear images. Mathematically, ASM can be formulated as follows:

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

where I and J respectively represent haze and non-haze images, A is the global atmospheric light, $t(x) = e^{-\beta d(x)}$ represents the transmission map, and β and $d(x)$ represent the atmospheric scattering parameters and scene depth, respectively.

Although many ASM-based methods Berman and Avidan (2016); Fattal (2015); He et al. (2011); Meng et al. (2013) have achieved good performance, A and t must be accurately estimated to restore clear images with ASM. However, the inaccurate estimation of A and t usually leads to unpleasant results when the images captured in a complex environment. In addition, since t s determined by the depth of the scene, ASM requires strong assumptions. In other words, the thickness of the haze is closely related to the depth of the scene. Due to this attribute, it is difficult for ASM-based methods to process degraded remote sensing images. In recent years, deep learning-based method LeCun et al. (2015) has achieved great success in single image dehazing field Cai et al. (2016); Li et al. (2017); Qin et al. (2020); Ren et al. (2018); Zhang and Patel (2018); Zhou et al. (2020). By using deep learning, the shortcomings of ASM can be avoided to a certain extent. With the availability of powerful convolutional neural networks (CNN), one can readily train them on large-scale datasets to learn a correct mapping from input hazy images to clear outputs. However, it is costly (and in some cases impossible) to obtain a large number of hazy images and corresponding clear ground truths in the real world.

Recently, Transformer Vaswani et al. (2017) has attracted widespread attention in the field of computer vision. Transformers were initially used for natural language processing (NLP), such as translation Wang et al. (2019) and question answering Tan and Bansal (2019). It can achieve such great success in NLP tasks because it has been pre-trained based on the transformer model on a large text corpus and fine-tuned on the dataset of the specific task. Transformer variants, such as BERT Devlin et al. (2019) and GPT-3 Brown et al. (2020), further enriched the training data and improved the pre-training performance. There are some interesting attempts to extend the transformer to the computer vision. For example, Wang et al. (2017) and Fu et al. (2019) use a self-attention model to obtain global image information; Carion et al. (2020) use a transformer architecture for end-to-end object detection; Dosovitskiy et al. (2020) introduce the visual transformer (ViT) treats the input image as 16 × 16 words, and has achieved good results in image recognition; Chen et al. (2020) proposed a pre-training model that based on transformer architecture for image processing, which uses multi-heads and multi-tails design handling three different tasks at the same time, including super-resolution, denoising, and deraining, and has achieved good results on the low-level visual tasks.

In this paper, Transformers are employed for remote sensing image dehazing for the first time. Specifically, we designed a neural network integrating Transformer and residual channel attention for haze removal from remote sensing images. Our neural network is divided into two branches. The first branch is called Transformer U-Net, and the efficient transformer block in Zhang and Yang (2021) is used as the main backbone to form a u-shaped encoder–decoder structure for modelling the long-range dependencies. While in the second branch, called RCAG-Net, is an end-to-end network structure. Due to the powerful feature mapping capabilities of the residual channel attention network (RCAN) Zhang et al. (2018), we design it as a residual channel attention group (RCAG-Net) to enhance fitting capability of models and complement fine-detailed features for upper branch. To better fuse the output features of the two branches, we design a fusion block to learn and aggregate these features of the two branches to form the final clear remote sensing image.

In general, the main contributions of our work are: (1) We employ Transformer model in the task of remote sensing image dehazing for the first time. (2) In order to learn a specific data distribution, we use a residual attention group, which can extract more features from the current data distribution and provide feature compensation for the upper branch. (3) We design a learnable fusion block to better fuse the outputs of the two branches and produce the final clear remote sensing image. (4) We present a large number of experimental results to demonstrate the effectiveness of our TransRA network in remote sensing images dehazing.

2 Related works

2.1 Single image dehazing

Single image dehazing methods can be roughly divided into two categories: priori-based methods and learning-based methods. Based on priori methods, the transmission map and global atmospheric light in ASM Oakley and Satherley (1998) are estimated. Many prior-based methods Berman and Avidan (2016); Fattal (2015) have shown good performance in single image dehazing. However, in practice, since the priors are easily violated, the prior-based methods are not always robust when encountering complicated scenarios. In addition, most of the existing dehazing algorithms are designed for natural images. Considering that the physical models of remote sensing images are different from those of natural images, IDeRS, Xu et al. (2019) proposed a concept of virtual depth for physical models of remote sensing images and introduced an iterative dehazing algorithm for remote sensing images, which has achieved good results on some existing public datasets. With the success of deep CNN, methods based on deep learning have received extensive attention in recent years. DehazeNet Cai et al. (2016) is the first dehazing model based on deep learning methods. It uses CNN to estimate the transmission map and uses the ASM Oakley and Satherley (1998) to produce the final dehazing clear image. Unlike DehazeNet, AOD-Net Li et al. (2017) reformulated the ASM Oakley and Satherley (1998) and presented a lightweight CNN to produce clear images from hazy images. In addition, there are currently many methods that can restore haze-free images without using ASM Oakley and Satherley (1998). Ren et al. (2016) proposed a multi-scale CNN by designing a set of coarse and fine scale networks to predict the image transmission map independently, then using multi-scale fusion to complete image dehazing. Qin et al. (2020) proposed FFA-Net with a new feature attention module, including pixel attention and channel attention. This work has achieved high performance on the RESIDE Li et al. (2019) dataset. GFN Ren et al. (2018) is a gated fusion network that restores haze images by transforming the input image using white balance and gamma correction. Yu et al. (2021) proposed a two branch dehazing framework for non-uniform dehazing by integrated learning, which uses the idea of transfer learning with channel attention fusion and achieves good results on all four different datasets. Shao et al. (2020) introduced a domain adaptation framework of dehazing tasks, which builds a bridge between the real world and the synthetic haze image. It is sub-optimal to directly use these above-mentioned methods to dehaze remote sensing images because the remote sensing image is inconsistent with the natural image in many aspects. To solve this problem, some studies have explored different learning strategies to extract haze-related features or heuristic clues in remote sensing images, such as CNN in Qin et al. (2018), Ke et al. (2019), Guo et al. (2021), Grohnfeldt et al. (2018) and Generative Adversarial Networks (GAN) in Huang et al. (2020), Mehta et al. (2021). For example, Qin et al. (2018) introduced a deep CNN with residual structure, and Ke et al. (2019) designed a complete CNN to remove haze from RS images. Huang et al. (2020) introduced conditional GAN preprocessing on SAR images to eliminate image haze. FCTF-Net Li and Chen (2020) is a first-coarse-then-fine two-stage dehazing neural network combining a channel attention mechanism with a basic convolutional block, which is trained in an end-to-end manner and has achieved better results on synthetic datasets and some real images. SkyGAN Mehta et al. (2021) is proposed to use HSI guidance and multi-color input for haze removal from satellite images. However, previous networks are less effective in preserving good spatial details and encoding contextual information. Thus, an effective dehazing model should be designed that not only removes the haze from degraded images but also retains the required spatial details (e.g., real textures and edges).

2.2 Transformer

Transformer Vaswani et al. (2017) and its variants have proven its success as a powerful unsupervised or self-supervised pre-training framework in various natural language processing tasks. For example, GPTs Radford et al. (2018, 2019); Brown et al. (2020) are pre-trained in an autoregressive manner to predict the next word in a huge text dataset. Due to the success of the Transformer model in NLP, many attempts have been made to explore the application of the Transformer model in computer vision tasks. These attempts can be roughly divided into two categories. First, self-attention is introduced in the traditional convolutional neural network. Yuan et al. (2018) introduced spatial attention in image segmentation. Fu et al. (2019) proposed DANET that combines spatial attention and channel attention to use contextual information. Wang et al. (2018), Chen et al. (2019), Jiang et al. (2019) and Zhang et al. (2019) also used self-attention to enhance features to improve the performance of the model in some high-level visual tasks. The other is to replace convolutional neural networks with self-attention blocks. For example, Kolesnikov et al. (2020), Dosovitskiy et al. (2020), use a transformer block for image classification. Carion et al. (2020) and Zhu et al. (2020) implemented Transformer-based models in detection. Chen et al. (2020) introduced a pre-trained GPT model for generation and classification tasks. Wu et al. (2020) and Zhao et al. (2020) proposed a transformer model-based pre-training method for image recognition tasks. Jiang et al. (2021) introduced TransGAN that used Transformer to replace convolution to form a strong GAN. However, there are few related studies focusing on low-level visual tasks. In this paper, we explore a novel remote sensing image dehazing network including

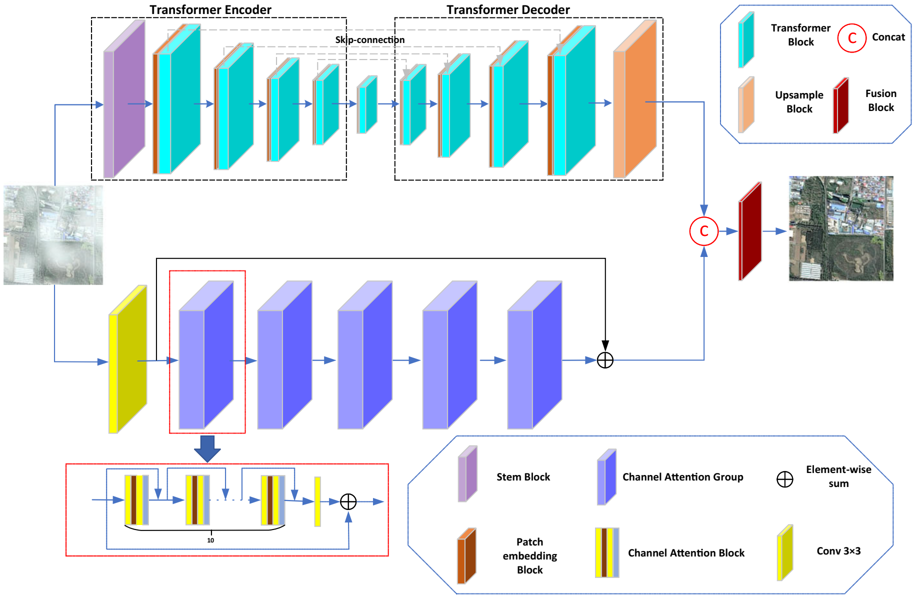

Fig. 1 We propose the structure of TransRA model. Our model includes transformer U-Net branch and RACG-Net branch. The model finally combines the output of the two branches to obtain the dehazing result through fusion block

Transformer and residual attention group. Specifically, two branches are trained separately and then passed through the fusion module to obtain a haze-free remote sensing image.

3 Proposed method

In this section, we will introduce two branches of remote sensing image dehazing neural network in detail, including the network of Transformer U-Net, RCAG-Net, and loss function.

3.1 Network structure

Our method includes both upper and lower networks, namely Transformer U-Net and RCAG-Net (as shown in Fig. 1). Each network is used for a specific purpose: Transformer U-Net extracts a robust global representation from the input remote sensing image by encoding and decoding. Here, Transformer is used to form the upper network because its own structure contains a self-attention mechanism, which can obtain more robust global context information than CNN. RCAG-Net is designed to work on a specific training image domain of the current remote sensing data to obtain more fine-grained remote sensing image features. The output feature maps of the two networks are concatenated together and then sent to the fusion block to output a clear haze-free image.

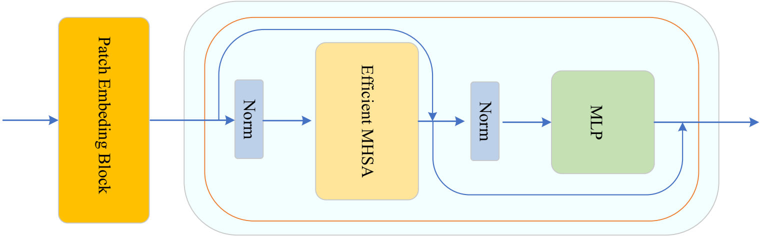

Fig. 2 Illustration of the transformer block. Among them, Norm represents the layer normalization operation, and MLP adopts the multi-layer perceptron operation proposed in Zhang and Yang (2021)

3.1.1 Transformer U-Net

The Transformer U-Net is an encoding–decoding structure, see the upper branch in Fig. 1. In order to better distinguish haze information from background texture information, and extract global robust features. Inspired by Zhang and Yang (2021), we use the ResT-large four-stage down-sampling structure as the encoder layers for our network. The decoder is also a four-stage process, each stage consists of a deconvolution upsampling patch embedding module and the transformer block shown in Fig. 2. There is an intermediate transition layer between encoding and decoding, which is composed of a down-sampling patch embedding module and a transformer block to transmit the encoded feature information to the decoding stage. To be specific, in the encoding stages, our Transformer U-Net consists of the following basic modules, namely stem block, patch embedding layer, and transformer block.

Stem block: The stem block is designed to effectively extract low-level information, reduce the height and width of the input feature, and expand the dimensions of the channels. It first introduces three convolutional layers with strides [2,1,2], followed by Batch Normalization and ReLU activation after the first two convolutional layers. After the stem block, the spatial resolution is reduced by 4 times, and the dimension of the channel is expanded from 3 to 96.

Patch embedding block: In the 2, 3, and 4 stages of the encoder structure, the patch embedding block is used to down-sample the spatial dimension by 4× and increase the channel dimension by 2×. This function can be implemented using standard 3×3 convolution with a stride of 2 and the BatchNorm layer. After this stage, there is an indispensable module called position encoding (as shown in Fig. 3a). Position encoding is the key to use sequence order. Here, our position encoding uses a simple and effective pixel-wise attention (PA) module (as shown in Fig. 4b). To be precise, the PA module uses a 3×3 depth-wise convolution (with padding 1) operation to obtain the pixel-wise weight and then uses the sigmoid function to scale. The position code of the PA module can be expressed as:

$$PA(x) = sigmoid(DWConv(x)) \tag{2}$$

where Dwconv denotes 3 × 3 depth-wise convolution with stride 1. Since the input tokens of each stage are obtained by convolutional operations, we can embed the position encoding into the patch embedding module, so the patch embedding of stages 2, 3, 4 combined with the position encoding can be written as:

$$PE(x') = x' * PA(x) = x' * sigmoid(DWconv(x')) \tag{3}$$

$$x' = BN(C_s * x) \tag{4}$$

where x and x' denote the feature map, C_s epresents a convolutional layer with kernel size $s+1$ and stride s. To achieve hierarchical feature extraction, we set the hyperparameter to 2.

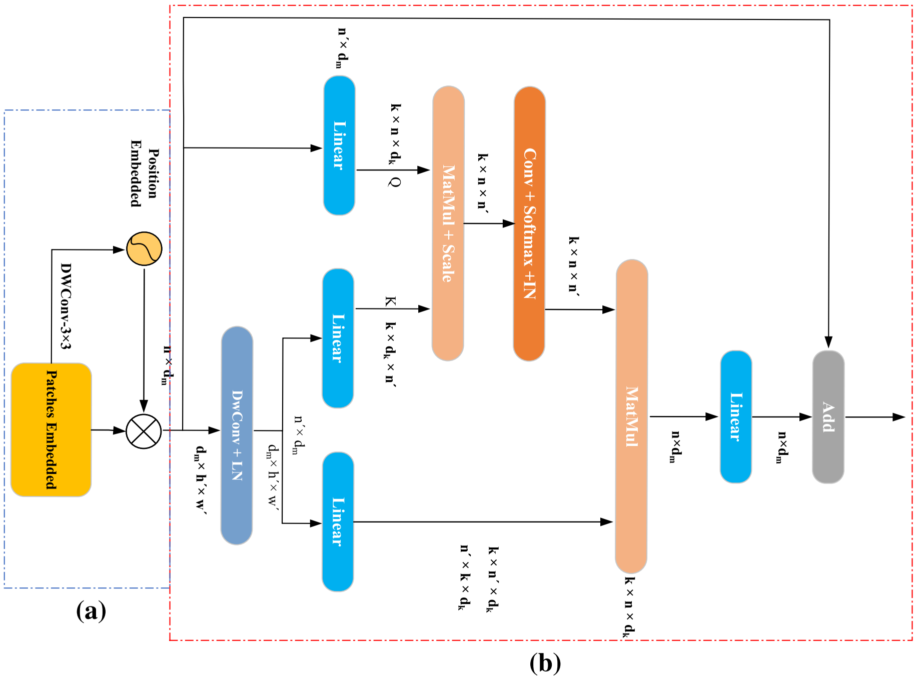

Fig. 3 Illustration of position embedded in patch emdeding block (**a**) and effection multi-head self-attention (EMHSA) in transformer block (**b**).These two parts of the structure constitute the transformer encode–decoder in our proposed transformer U-Net branch

It is worth mentioning that there is also a PA layer in our Stem Block to add the input feature map in the first stage position information.

Transformer block: Transformer block is composed of an efficient multi-head self-attention (EMHSA) (see in Fig. 3b), multi-layer perceptron [MLP Zhang and Yang (2021)], and layer normalization Ba et al. (2016) operations (see in Fig. 2). Each output of the Efficient transformer block can be written as:

$$\hat{x} = x + EMHSA(LN(x)) \tag{5}$$

$$y = \hat{x} + MLP(LN(\hat{x}) \tag{6}$$

where x denotes the feature map obtained after the patch embedding block and $\hat{x}$ denotes the intermediate feature map, as shown in Eq. 5. The feature map x first passes through the layer norm, then EMHSA, and then sums with itself to obtain the feature map $\hat{x}$. It is worth noting that EMHSA is the core of the whole efficient transformer block, which can be obtained in the following way

$$EMHSA(Q, K, V) = IN\left(\text{Softmax}\left(\text{Conv}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)\right)V \tag{7}$$

Here, Conv denotes a standard 1 × 1 convolution, IN denotes instance normalization, and d_k denotes head dimension. For more details about EMSA and MLP, see Zhang and Yang (2021).

In the decoding stage, our dehazing network includes up-patch embedding block, efficient transformer block, and up-sampling block. Specifically, our up-Patch embedding block uses

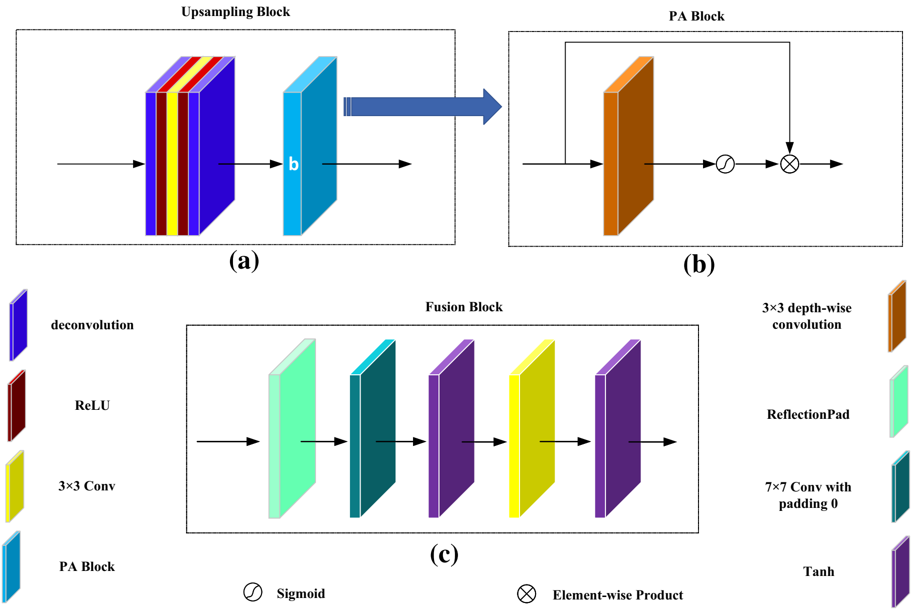

Fig. 4 The structure diagram of upsampling block (**a**), PA block (**b**) and fusion block (**c**). Among them, upsample block is used in our proposed transformer U-Net branch to restore the feature size to be consistent with the network input image; PA Block is used in Patch embedding block and upsample block to add position information. Fusion block is to better fuse the output features of the two branches

deconvolution to enlarge the feature map of the i-th decoding stage upwards by the same parameter factor s as the patch embedding block. Then pass through BatchNorm and add the position module PA to form a complete up-patch embedding block. The efficient transformer block in the encoding–decoding stage is the same and is used to extract the long-distance dependency in the feature map. The up-sampling block is shown in Fig. 4a, which contains two layers of 4×4 deconvolution with stride of 2, and two ReLU activation layers. The 3×3 convolution layer is the third layer of the block, and the PA block is the final output layer. The up-sampling block converts the feature map obtained after decoding to the same resolution as the input image, so as to better fuse with the feature map output from the lower branch.

3.1.2 RCAG-Net

Our RCAG-Net is based on the remaining channel attention block Zhang et al. (2018) (see the lower part of Fig. 1). This block contains the convolutional layers and the channel attention modules. Due to the remaining design and long-hop connections, the network is unlikely to have the problem of vanishing gradients Pascanu et al. (2013). In addition, channel attention highlights salient features to enhance the ability of the model to fit the current data. Moreover, to preserve the fine features, this branch avoids the use of down-sampling and up-sampling operations. The detailed features can be seen as a supplement to the upper branch network. Since our RCAG-Net is trained from scratch and is an end-to-end network built for full resolution, it can complement fine-detailed features for upper branch.

3.1.3 Fusion block

Our fusion block serves as an integration method to produce the final output (shown in Fig. 4c). Specifically, the fusion block fuses the features of the two branches and then learns to map the features to a clear image. The fusion block is constructed from two different convolutional layers and hyperbolic tangent activation functions. Although the two branches have provided enough and distinct features for recovering a clear image, the number of channels after fusion is 20 times more than that of the input image. Therefore, we use a two-layer convolution operation to ensure that the number of channels is decreasingly learned layer by layer to finally output a clear image. In contrast, we find that the single convolution operation may jeopardize the overall generalization ability of our method and lead to performance degradation. To further illustrate the effectiveness of our fusion blocks, we empirically investigate the effect of employing different sizes of fusion blocks from Table 3 in Sect. 4.3. It can be observed that the performance of our method decreases accordingly by using different convolution operations.

3.2 Loss functions

Our loss function consists of four different parts. Each loss function has a specific purpose.

3.2.1 Smooth L1 loss

We used the smooth L1 loss to ensure that the predicted image is close to a clean image. It is a robust L1 loss Girshick (2015), which has been proven to be better than L2 loss in many image restoration tasks. The smooth L1 loss can be written as:

$$L_{smooth} = \frac{1}{N}\sum_{i}^{N} smooth_{L1}(y_i - f_\sigma(x_i)) \tag{8}$$

$$smooth_{L1}(z) = \begin{cases} 0.5z^2 & if|z| < 1 \\ |2| - 0.5 & otherwise \end{cases} \tag{9}$$

where y_i and x_i represent the ground truth and hazy image at the pixel point i, respectively, $f_\sigma(\cdot)$ represents the neural network with parameter σ, and N denotes the total number of pixels.

3.2.2 Perceptual loss

We used perceptual loss Zhu et al. (2017) to provide additional supervision in the high-level feature space. It has been shown that training with perceptual loss allows the model to better reconstruct details. The loss network is a pre-trained VGG-16 on ImageNet Simonyan and Zisserman (2014). The loss function is as follows:

$$L_{per} = \frac{1}{N}\sum_{i}^{N} \frac{1}{C_j H_j W_j} \left\|\phi_j\left(f_\sigma(x)\right) - \phi_j(y)\right\|_2^2 \tag{10}$$

where x and y represent the input with haze and the ground truth image, respectively, $f_\sigma(x)$ represents the image after dehazing, and $\phi_j(\cdot)$ denotes the feature map of size $C_j \times H_j \times W_j$. The feature reconstruction loss is the L2 loss. N is the number of features used in the perceptual loss.

3.2.3 Adversarial loss

Contrast loss is effective in helping to recover photo-realistic images Ledig et al. (2017). Especially for small-scale datasets, pixelated loss functions usually do not provide enough monitoring signals to train the network to recover photo-realistic details. Therefore, we finally implemented the adversarial loss with discriminators in Zhu et al. (2017). The adversarial loss can be expressed as follows:

$$L_{adv} = \sum_{i}^{N} - \log D\left(f_{\sigma}\left(x\right)\right) \tag{11}$$

where $D\left(\cdot\right)$ is a discriminator. The probability that the dehazed image $f_{\sigma}(x)$ is a ground truth image is shown as $D\left(f_{\sigma}\left(x\right)\right)$.

3.2.4 Wavelet SSIM loss

Wavelet SSIM loss was proposed in Yang et al. (2020), which is a loss function based on SSIM, and it has been demonstrated that the Wavelet SSIM loss can improve PSNR and SSIM. The equation is as follows:

$$L_{W\text{-}SSIM}(f_{\sigma}(x), y) = \sum_{0}^{i} r_i L_{SSIM}\left([f_{\sigma}(x)]_i^{\omega}, y_i^{\omega}\right), \omega \in \{LL, KL, LH, HH\} \tag{12}$$

where r_i controls the weight of each patch, x and y represent the input with haze and the ground truth image, respectively, $f_{\sigma}(x)$ represents the image after dehazing. The L_{SSIM} refers to the SSIM loss function, which can be written as:

$$L_{SSIM} = -SSIM\left(x, y\right) \tag{13}$$

$$SSIM\left(x, y\right) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{14}$$

where x and y represent the dehazed image and ground truth image, respectively. Equation 14 calculates the SSIM of the two images, where μ and σ represent the average, standard deviation, and covariance of the image. For the details of this loss function, see Yang et al. (2020).

The total loss function is defined as:

$$L = \lambda_1 L_{smooth} + \lambda_2 L_{per} + \lambda_3 L_{adv} + \lambda_4 L_{W\text{-}SSIM} \tag{15}$$

where λ_1, λ_2, λ_3 and λ_4 are the hyperparameters to balance between different losses.

4 Experiments

In this section, we start with the description of the datasets, training details, and evaluation indicators. Then we conduct ablation studies to clarify the impact of the different modules in our method. Finally, we compared the method with other competing dehazing algorithms qualitatively and quantitatively.

4.1 Experiment settings

4.1.1 Datasets

To verify the effectiveness of our network, we conducted dehazing experiments on two recently synthesized remote sensing datasets, namely SateHaze1k Huang et al. (2020) and RICE Lin et al. (2019). In particular, the SateHaze1k dataset was created by Huang et al. (2020), which provided 1200 pairs of hazy remote sensing images, and divided the haze into three levels: thin, medium, and thick. Each level of the dataset contains 400 pairs of hazy images, where the training dataset accounts for 80% of the total data of each level, the validaton dataset and the testing dataset contain 35 pairs and 45 pairs of remote sensing images respectively. The RICE dataset was collected on Google Earth and consists of 500 pairs of hazy remote sensing images.

4.1.2 Implementation details

We use 90°, 180°, 270° of random rotation, horizontal flip, and vertical flip to increase the training dataset. The input image is randomly cropped to a size of 256 × 256. We use the Adam optimizer with batch 2 and the default $\beta 1 = 0.9$ and $\beta 2 = 0.999$. The whole process was trained for 1000 epochs, with an initial learning rate of 0.0001, and then reduced by half to the 300th, 600th, and 800th epochs. The hyperparameters λ_1 , λ_2 , λ_3 , and λ_4 of the loss function are 2.0, 0.001, 0.005, and 1.0, respectively. We implemented our model with PyTorch on a single NVIDIA TITAN XP GPU. The training takes about 50 h.

4.1.3 Evaluation metric and compatitors

To evaluate the performance of our method, we adopt the peak signal to noise ratio (PSNR) and the structural similarity index (SSIM) as the evaluation metrics, which are usually used as criteria to evaluate image quality in the image dehazing task. We compare with various dehazing methods including DCP He et al. (2011), IDERS Xu et al. (2019), DehazeNet Cai et al. (2016), MSCNN Ren et al. (2016), AOD-Net Li et al. (2017), GFN Ren et al. (2018), FFA Qin et al. (2020), FCTF-Net Li and Chen (2020) and Two-Branch Yu et al. (2021).

4.2 Comparisons with various dehazing methods

4.2.1 Results on SateHaze1k dataset

We demonstrate the performance of our TransRA and various dehazing methods on the SateHaze1k dataset. As can be seen from the Table 1, our TransRA achieves the best results at all three different levels of haze, with 25.20 dB, 26.50 dB, 22.73 dB in terms of PSNR, and 0.9302, 0.9474, 0.8451 in terms of SSIM, respectively. In particular, our TransRA achieved significant improvements averaging 0.54 dB PSNR and 0.0064 SSIM over three different levels of haze data compared to the suboptimal method. We also compare our TransRA with various dehazing methods on the quality of the restored images, which is shown in Figs. 5, 6, and 7. We can observe that DCP, IDERS, DehazeNet, MSCNN, AOD-Net, and GFN cannot successfully remove haze from remote sensing images and exist color distortion, especially in the cases of thick and medium haze. Compared to DCP, IDERS, DehazeNet, MSCNN, AOD-Net, and GFN, the methods of FFA, FCTF-Net and Two-Branch achieve the restored remote

Table 1 Quantitative comparisons over SateHaze1k for different methods

Method	Thin haze		Moderate haze		Thick haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP He et al. (2011)	15.61	0.7566	16.50	0.7902	15.94	0.7240
IEDRS Xu et al. (2019)	15.07	0.7699	14.73	0.7767	11.76	0.6958
DehazeNet Cai et al. (2016)	19.48	0.8698	18.86	0.8773	12.77	0.6633
MSCNN Ren et al. (2016)	19.61	0.8808	18.49	0.8972	13.10	0.7080
AOD-Net Li et al. (2017)	17.67	0.8565	14.77	0.7104	14.50	0.7143
GFN Ren et al. (2018)	17.15	0.8585	16.92	0.8833	13.82	0.7441
FFA Qin et al. (2020)	24.41	0.9261	25.85	0.9368	22.26	0.8638
FCTF-Net Li and Chen (2020)	20.41	0.8848	23.59	0.9313	18.37	0.8022
Two-branch Yu et al. (2021)	24.20	0.9226	25.89	0.9434	22.50	0.8637
Ours	**25.20**	**0.9302**	**26.50**	**0.9474**	**22.73**	**0.8751**

The best results are in bold and the second best are with underline

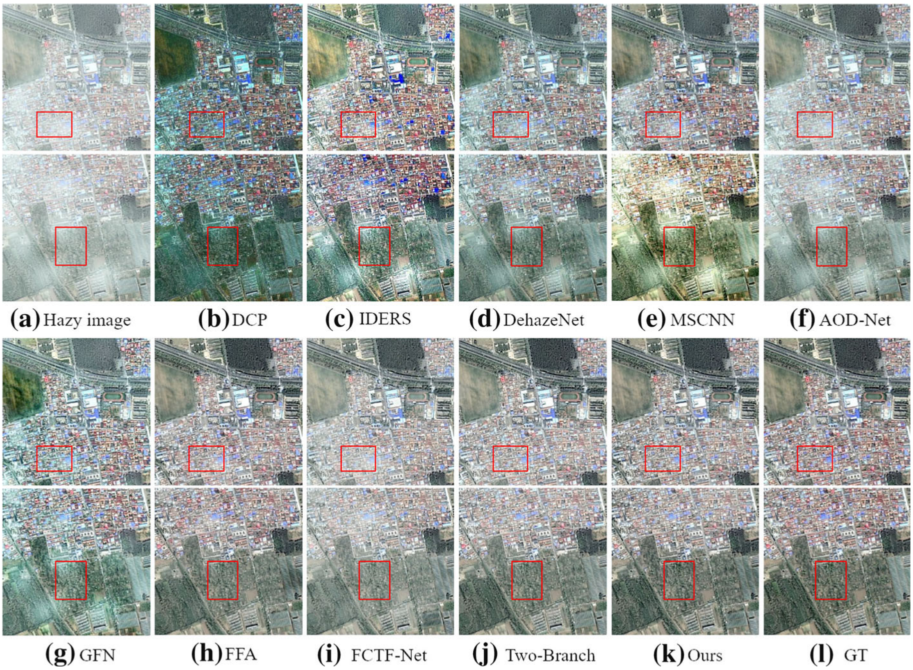
(a) Hazy image (b) DCP (c) IDERS (d) DehazeNet (e) MSCNN (f) AOD-Net
(g) GFN (h) FFA (i) FCTF-Net (j) Two-Branch (k) Ours (l) GT

Fig. 5 Qualitative comparisons of all the comparing methods and ours TransRA(k) on the StaeHaze1k (thin haze) dataset

senesing images with higher quality. However, they still exist some defects in processing high-frequency details. In contrast, our method can better retain the high-frequency details and recover the color information of the whole remote sensing image.

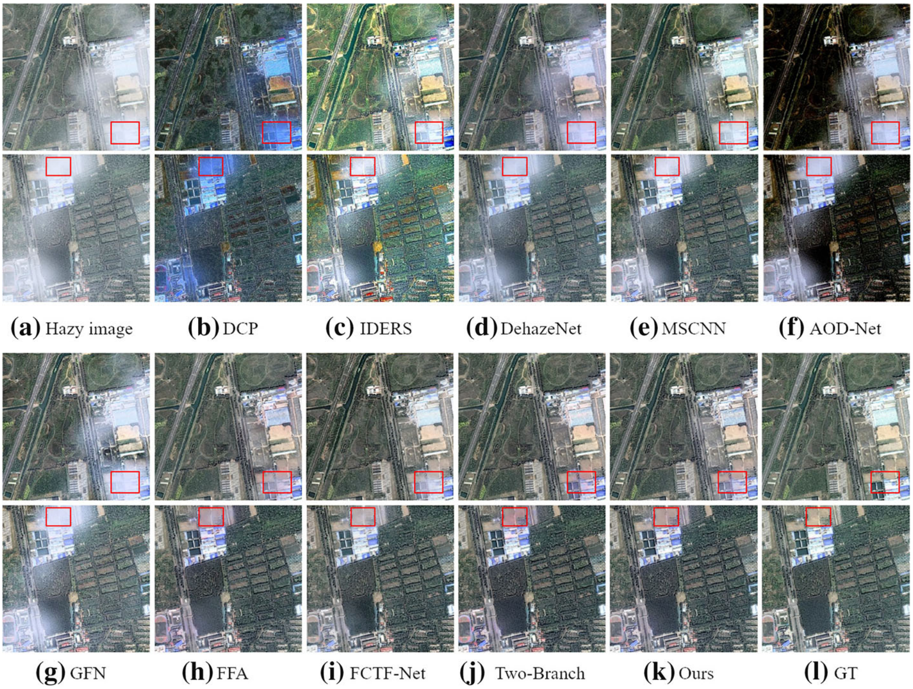

Fig. 6 Qualitative comparisons of all the comparing methods and ours TransRA(k) on the StaeHaze1k (moderate haze) dataset

4.2.2 Results on RICE dataset

We also compare the performance of our TransRA and various dehazing methods on the RICE dataset. As shown in Table 2, we can observe that our TransRA has a PSNR of 31.13 dB and a SSIM of 0.9551 on the RICE dataset, outperforming other dehazing methods except FFA. Note that FFA only achieves 0.65 dB over our method in term of PSNR, however, our results are closer to the ground truth on subjectively, which can be seen in Fig. 8. In addition, our results are the best both qualitatively and quantitatively compared with other methods. Although the results of FFA are higher than our method in terms of PSNR, our method performs better in terms of SSIM and qualitative performance.

4.3 Ablation studies

To analyze and evaluate the effectiveness of each component of our approach, we combine five factors for the ablation study: fusion block, PA block, Transformer U-net, RCAG Net, contribution of each loss function and ablation study of its hyperparameters. The ablation experiments are shown below: (1) Differences in fusion blocks: Ablation studies were performed using a different fusion block than the one we proposed. (2) Transformer U-Net and fusion block TF: Only the combination of Transformer U-Net and fusion block is used. (3) RCAG with fusion block RF: Only RCAG with fusion block is used. (4) Contribution of Each Loss: The constraints of each loss on the algorithm in this paper. (5) Without PA Block:

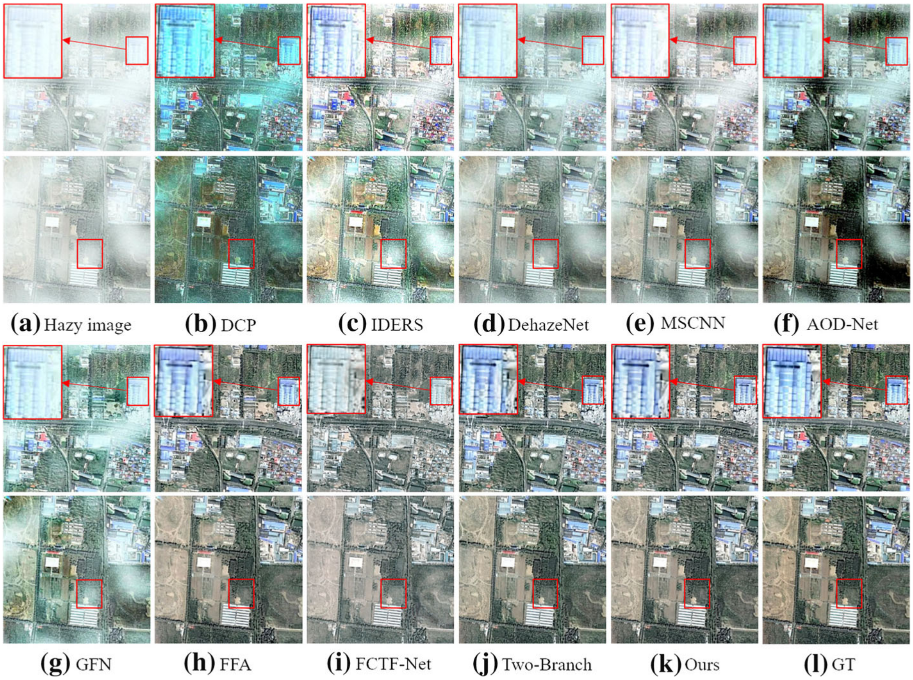

Fig. 7 Qualitative comparisons of all the comparing methods and ours TransRA(k) on the StaeHaze1k (thick haze) dataset

Table 2 Quantitative comparisons over RICE for different methods

Method	RICE	
	PSNR	SSIM
DCP He et al. (2011)	16.91	0.7590
IEDRS Xu et al. (2019)	14.21	0.6364
DehazeNet Cai et al. (2016)	23.74	0.8645
MSCNN Ren et al. (2016)	22.72	0.8792
AOD-Net Li et al. (2017)	19.84	0.7396
GFN Ren et al. (2018)	16.53	0.7792
FFA Qin et al. (2020)	**31.78**	0.9543
FCTF-Net Li and Chen (2020)	28.49	0.9170
Two-branch Yu et al. (2021)	30.49	0.9525
Ours	31.13	**0.9551**

The best results are in bold and the second best are with underline

We without to add position information by used PA Block in Transformer U-Net. (6) Ours: Using a combination of Transformer U-Net and RCAG with our fusion block.

Specifically, we used the thick haze dataset in SateHaze1k as the train and test datasets for our ablation study. The quantitative results of the ablation study are shown in Tables 3, 4 and 5. Table 3 shows the results of our fusion block and the other two types of single convolution fusion blocks. It can be seen that the double convolution fusion block designed by our can significantly improve PSNR and SSIM. Table 4 shows the experimental results

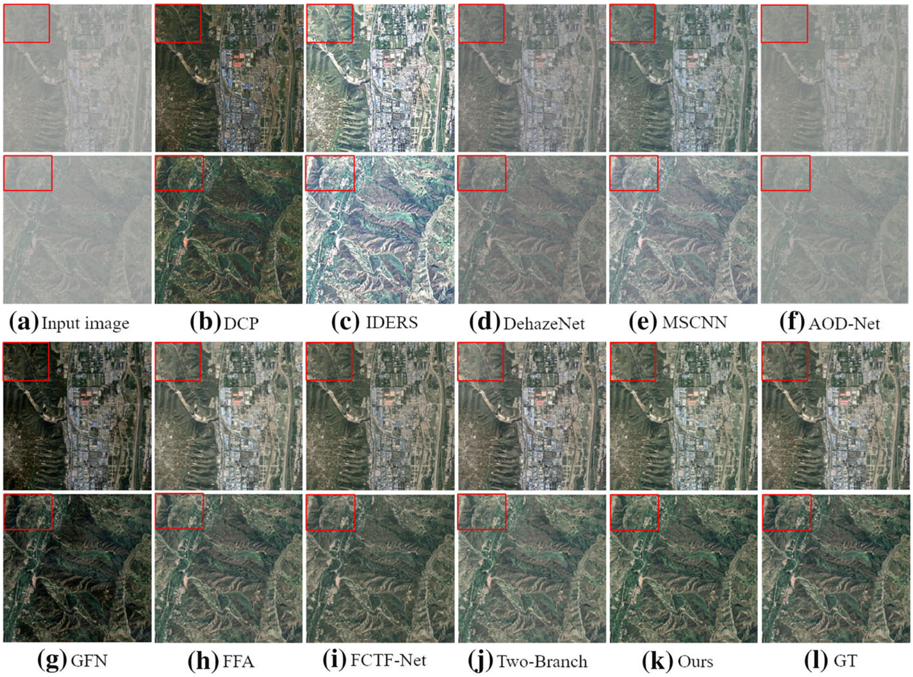

Fig. 8 Qualitative comparisons of all the comparing methods and ours TransRA(k) on the RICE dataset

Table 3 The performance of our proposed method using fusion blocks with different sizes

Fusion blocks	PSNR	SSIM
3 × 3 Conv + Tanh	22.62	0.8737
RefiectionPad + 7 × 7 Conv + Tanh	22.40	0.8726
Ours	**22.73**	**0.8751**

The best results are in bold
The evaluation is conducted on StakeHaze1k (thick haze) dataset

Table 4 Ablation studies of different combinations

Methods	PSNR	SSIM
TF[a]	21.95	0.8637
RF[b]	22.50	0.8736
w/o PA[c]	22.62	0.8740
Ours	**22.73**	**0.8751**

The best results are in bold
The evaluation is conducted on StakeHaze1k (thick haze) dataset
[a] "TF" denotes transformer U-Net and fusion block
[b] "RF" denotes RCAG Net and fusion block
[c] "w/o PA" denotes without PA block in our network

obtained by combining our two-branch network and the fusion block respectively. In addition, quantitative results without PA block in Transformer U-Net are also presented in Table 4. We can see that our two-branch network is much better than a single-branch network, which demonstrated the effectiveness of our two-branch design. The PSNR and SSIM of our model

Table 5 Quantitative ablation study of the contribution of each loss (smooth L1 loss L_{smooth}, perceptual loss L_{per}, adversarial loss L_{adv}, wavelet SSIM loss $L_{W\text{-}SSIM}$)

Ablation	Loss function type				Thick haze	
	L_{smooth}	L_{per}	L_{adv}	$L_{W\text{-}SSIM}$	PSNR	SSIM
1	w/o				22.46	0.8743
2		w/o			22.20	0.8726
3			w/o		22.64	0.8750
4				w/o	22.56	0.8629
Ours	✓	✓	✓	✓	**22.73**	**0.8751**

The best results are in bold

The evaluation is conducted on StakeHaze1k (thick haze) dataset

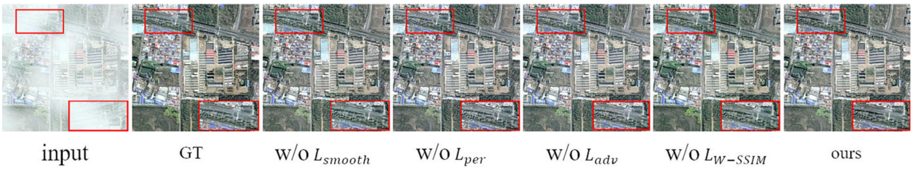

Fig. 9 Qualitative ablation study of the contribution of each loss (smooth L1 loss L_{smooth}, perceptual loss L_{per}, adversarial loss L_{adv}, wavelet SSIM loss $L_{W\text{-}SSIM}$) and ours TransRA on the StakeHaze 1k (thick haze) dataset

reached 22.73 dB and 0.8751, respectively. When there is without PA block to add position information in Transformer U-Net, the results obtained are significantly worse than the full structure. This shows that the role of the PA block cannot be ignored. These scores show that each component we considered plays a crucial role in performance of network.

We present the results of training TransRA with various loss combinations in Fig. 9 and Table 5. The results without smooth L1 loss L_{smooth} are partially noisy compared to the full results (e.g., the red boxed area in Fig. 9). This shows the importance of L_{smooth} in removing noisy luminous surfaces. Removing the perceptual loss L_{per} results in a reduction in overall contrast and a large difference from the full results in PSNR and SSIM metrics. The qualitative gap is not significant when the adversarial loss L_{adv} is discarded, but the full results outperform quantitatively, suggesting that the L_{adv} plays a role in the training process. Finally, removing the wavelet loss $L_{W\text{-}SSIM}$ causes the image color to be inconsistent with the GT image, with slight distortion, and differs from the full result by 0.17 and 0.0122 on PSNR and SSIM, respectively.

5 Conclusion

In this paper, we propose a remote sensing image dehazing neural network based on the fusion of transformer and residual channel attention, and demonstrate its strong performance in remote sensing image dehazing tasks. To preserve the clear spatial detail and color fidelity of the resulting images, we use the fusion block to fuse the features of Transformer and residual channel attention. Our method has significant advantages on two current remote sensing dehazing datasets. It surpasses various dehazing methods in terms of qualitative and quantitative. Moreover, the ablation study demonstrates the effectiveness of our proposed individual modules.

Acknowledgements This work is supported by Higher Education Scientific Research Project of Ningxia (NGY2017009).

References

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Berman, D., & Avidan, S. (2016). Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1674–1682).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ...Amodei, A. (2020). Language models are few-shot learners. Preprint at arXiv:2005.14165.

Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing, 25*(11), 5187–5198.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., & Gao, W. (2020). Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp 12,299–12,310).

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In *International conference on machine learning, PMLR* (pp. 1691–1703).

Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., & Kalantidis, Y. (2019). Graph-based global reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 433–442).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp 4171–4186).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv:2010.11929.

Fattal, R. (2015). Dehazing using color-lines. *ACM Transactions on Graphics, 34*(1), 1–14.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 3146–3154).

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440–1448).

Grohnfeldt, C., Schmitt, M., & Zhu, X. (2018). A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images. In *IGARSS 2018–2018 IEEE international geoscience and remote sensing symposium* (pp. 1726–1729). IEEE.

Guo, J., Yang, J., Yue, H., Tan, H., Hou, C., & Li, K. (2021). Rsdehazenet: Dehazing network with channel refinement for multispectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 59*(3), 2535–2549. https://doi.org/10.1109/TGRS.2020.3004556

He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(12), 2341–2353.

Huang, B., Zhi, L., Yang, C., Sun, F., & Song, Y. (2020). Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1806–1813).

Jiang, P.-T., Hou, Q., Cao, Y., Cheng, M. M., Wei, Y., & Xiong, H. K. (2019). Integral object mining via online attention accumulation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2070–2079).

Jiang, Y., Chang, S., & Wang, Z. (2021). Transgan: Two transformers can make one strong GAN. arXiv preprint arXiv:2102.07074

Ke, L., Liao, P., Zhang, X., Chen, G., Zhu, K., Wang, Q., & Tan, X. (2019). Haze removal from a single remote sensing image based on a fully convolutional neural network. *Journal of Applied Remote Sensing, 13*(3), 036,505.

Kim, J.-H., Jang, W.-D., Sim, J.-Y., & Kim, C. S. (2013). Optimized contrast enhancement for real-time image and video dehazing. *Journal of Visual Communication and Image Representation, 24*(3), 410–425.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part V 16* (pp. 491–507). Springer.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2017). Aod-net: All-in-one dehazing network. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 4780–4788).

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., & Wang, Z. (2019). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing, 28*(1), 492–505.

Li, Y., & Chen, X. (2020). A coarse-to-fine two-stage attentive network for haze removal of remote sensing images. *IEEE Geoscience and Remote Sensing Letters*. https://doi.org/10.1109/LGRS.2020.3006533

Lin, D., Xu, G., Wang, X., Wang, Y., Sun, X., & Fu, K. (2019). A remote sensing image dataset for cloud removal. arXiv preprint arXiv:1901.00600.

Liu, C., Hu, J., Lin, Y., Wu, S., & Huang, W. (2011). Haze detection, perfection and removal for high spatial resolution satellite imagery. *International Journal of Remote Sensing, 32*(23), 8685–8697. https://doi.org/10.1080/01431161.2010.547884

Long, J., Shi, Z., Tang, W., & Zhang, C. (2014). Single remote sensing image dehazing. *IEEE Geoscience and Remote Sensing Letters, 11*(1), 59–63. https://doi.org/10.1109/LGRS.2013.2245857

Makarau, A., Richter, R., Muller, R., & Reinartz, P. (2014). Haze detection and removal in remotely sensed multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing, 52*(9), 5895–5905.

Mehta, A., Sinha, H., Mandal, M., & Narang, P. (2021). Domain-aware unsupervised hyperspectral reconstruction for aerial image dehazing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 413–422).

Meng, G., Wang, Y., Duan, J., Xiang, S., & Pan, C. (2013). Efficient image dehazing with boundary constraint and contextual regularization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Oakley, J. P., & Satherley, B. L. (1998). Improving image quality in poor visibility conditions using a physical model for contrast degradation. *IEEE Transactions on Image Processing, 7*(2), 167–179.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning, PMLR* (pp. 1310–1318).

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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(5), 1645–1655.

Qin, X., Wang, Z., Bai, Y., Xie, X., & Jia, H. (2020). FFA-NET: Feature fusion attention network for single image dehazing. pp 11,908–11,915

Radford, A., Child, R. W., Luan, D., Wu, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., & Yang, M. H. (2016). Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision* (pp. 154–169). Springer.

Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., & Yang, M. H. (2018). Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3253–3261).

Shao, Y., Li, L., Ren, W., Gao, C., & Sang, N. (2020). Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2808–2817).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. Preprint at arXiv:1908.07490v1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Wang, Q., Li, B., Bai, Y., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). Learning deep transformer models for machine translation. Preprint at arXiv:1906.01787.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677.

Xu, L., Zhao, D., Yan, Y., Kwong, S., Chen, J., & Duan, L. Y. (2019). Iders: Iterative dehazing method for single remote sensing image. *Information Sciences, 489*, 50–62.

Yang, H.-H., Yang, C.-H.H., & Tsai, Y.-C.J. (2020). Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing. In *ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2628–2632). IEEE.

Yu, Y., Liu, H., Fu, M., Chen, J., Wang, X., & Wang, K. (2021). A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 193–202).

Yuan, Y., Guo, J. H., Zhang, C., Zhang, C., Chen, X., & Wang, J. (2018). Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916.

Zhang, H., & Patel, V. M. (2018). Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3194–3203).

Zhang, Q., & Yang, Y. (2021). Rest: An efficient transformer for visual recognition. arXiv preprint arXiv:2105.13677.

Zhang, S., He, X., & Yan, S. (2019). Latentgnn: Learning efficient non-local relations for visual recognition. In *International conference on machine learning, PMLR* (pp. 7374–7383).

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 286–301).

Zhao, H., Jia, J., & Koltun, V. (2020). Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10,076–10,085).

Zhou, Z., Guo, M., Feng, Y., Feng, Y., & Zhao, M. (2020). Cggan: A context guided generative adversarial network for single image dehazing. *CoRR, 14*(15), 3982–3988.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Pengwei Dong received the B.S. degree in Ningxia University in 2020. He is currently working toward the M.S. degree in School of Physics and Electronic-Electrical Engineering in Ningxia University. His research interests include remote sensing image enhancement and computer vision.

Bo Wang received his Ph.D. degree from the School of Electrical and Information Engineering, Tianjin University in 2016. He is currently an associate professor in the School of Physics and Electronic-Electrical Engineering, Ningxia University. His research interests include image restoration and enhancement, image classification and medical image processing.