Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.027727

ARTICLE

Single Image Desnow Based on Vision Transformer and Conditional Generative Adversarial Network for Internet of Vehicles

Bingcai Wei, Di Wang, Zhuang Wang and Liye Zhang[*]

College of Computer Science and Technology, Shandong University of Technology, Zibo, 255049, China

*Corresponding Author: Liye Zhang. Email: zhangliye@sdut.edu.cn

Received: 11 November 2022 Accepted: 29 January 2023 Published: 28 June 2023

ABSTRACT

With the increasing popularity of artificial intelligence applications, machine learning is also playing an increasingly important role in the Internet of Things (IoT) and the Internet of Vehicles (IoV). As an essential part of the IoV, smart transportation relies heavily on information obtained from images. However, inclement weather, such as snowy weather, negatively impacts the process and can hinder the regular operation of imaging equipment and the acquisition of conventional image information. Not only that, but the snow also makes intelligent transportation systems make the wrong judgment of road conditions and the entire system of the Internet of Vehicles adverse. This paper describes the single image snow removal task and the use of a vision transformer to generate adversarial networks. The residual structure is used in the algorithm, and the Transformer structure is used in the network structure of the generator in the generative adversarial networks, which improves the accuracy of the snow removal task. Moreover, the vision transformer has good scalability and versatility for larger models and has a more vital fitting ability than the previously popular convolutional neural networks. The Snow100K dataset is used for training, testing and comparison, and the peak signal-to-noise ratio and structural similarity are used as evaluation indicators. The experimental results show that the improved snow removal algorithm performs well and can obtain high-quality snow removal images.

KEYWORDS

Artificial intelligence; Internet of Things; vision transformer; deep learning; image desnow

1 Introduction

Artificial intelligence's recent rapid development has driven the Internet of Things [1,2]. At the same time, with the development of the automobile industry and the increasing number of vehicles, vehicle safety and management issues have become increasingly important. The Internet of Vehicles can improve the efficiency of road traffic and make it possible to exchange information between vehicles [3,4]. As a considerable part of the Internet of Things, the Internet of Vehicles can use multiple communication technologies for data interconnection. Specifically, based on the Internet of Vehicles technology, through the camera, radar and other sensors [5], the information of vehicles, people and roads can be scanned, and the real-time detection and sharing of vehicle information, road traffic and personnel information can be realized. Therefore, vehicle networking technology can learn

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

multiple connections between vehicles and roads, vehicles and people, and vehicles and vehicles. The Internet of Vehicles can facilitate travel while ensuring the safety of people, cars and roads. In many information collection methods, the sensor of machine vision technology can take the lead in cost, interactivity and other aspects, so it has been a broad concern by researchers. Therefore, individual vehicles in the Internet of Vehicles need to have the ability to extract information from visual data, which is indispensable, such as pedestrian detection and recognition. Vision-based traffic information extraction generally requires high-quality acquired data.

Most information obtained based on vision is the number of vehicles and pedestrians, vehicle license plate number and traffic signs. These are collectively referred to as traffic images, which contain a large amount of primary information. This information is of great significance for intelligent transportation and autonomous driving. Most of the vehicle scenarios are outdoors. However, the impact of bad weather is particularly significant. For example, on snowy days, the size of snowflakes changes significantly, which usually causes different degrees of occlusion to the image, which may cause the reduction of image quality, thus affecting the regular operation of vision-based sensors. For tasks requiring high precision, such as intelligent transportation or automatic driving, the wrong signal of the visual sensor may cause very severe consequences.

In daily life, the main ways people obtain information are visual and auditory, so it is crucial to get clear images. Image degradation caused by lousy weather is widespread, and this kind of weather significantly impacts human vision. For example, visibility is abysmal when driving in heavy snow, rain or fog, which easily causes traffic accidents. In some monitoring systems, the blurring of the image can also cause the degradation of the computer recognition performance and thus create unpredictable dangers. In recent years, with the continuous development of computer vision, restoring blurred images, especially those caused by bad weather, is in full swing. Snow is typically lousy weather in daily weather, coupled with the complexity and difficulty of calculation recognition caused by the different morphology and opacity. So, snow removal from a single image is a research work of practical significance.

Image degradation caused by bad weather is usually due to atmospheric particle noise such as fog, rain and snow. In the traditional image snow or rain removal algorithm, Luo et al. [6] proposed an algorithm based on image decomposition to separate rain through sparse coding and dictionary learning. Kim et al. [7] proposed an adaptive rain pattern algorithm. Chen et al. [8] proposed a rain-grain removal method based on Error Optimized Sparse Representation Model (EOSR). Finally, Li et al. [9] proposed a Gaussian mixture model for image rain removal. All of the above methods are model-driven rain removal algorithms, which use the known prior knowledge such as the direction, size and density of rain grain to establish the rain removal model and then design and optimize the algorithm to obtain relatively clear rain-free images. With the development of deep learning, the data-based rain removal algorithm has become more and more mainstream. For example, JORDER [10] proposed a rain model based on location and regional distribution. Fu et al. [11] tried to remove rain lines through a deep convolutional neural network (DerainNet). After that, many CNN-based methods have been proposed, such as [12,13], which use labels to learn nonlinear mappings.

In the study of single-image haze removal, Tian [14] removed haze locally by maximizing blurred images; Fattal et al. [15] inferred the blurred image's projection medium by estimating the scene's albedo to eliminate the blur. Huang et al. [16] proposed the Laplacian visibility restoration technology to refine the projection graph and solve the colour transmission problem. However, this traditional hand-made method is influenced by the prior clues, and the restored images could be better. Some recent deep learning-based algorithms focus on learning the mapping between blurred pictures and their corresponding clear images, which are more effective than traditional methods. The dehazing algorithm for a single image is devoted to extracting global features, which usually restores the blur of the whole area.

Finally, a separate model algorithm is usually proposed to remove snow particles from a single image. Unlike rain and fog particles, which are highly similar in their characteristics, transparency, and trajectories, snow particles have a more complex variation in shape and size. Traditional snow removal methods include [17,18], etc. Similar to the rain and fog removal model, the known prior information is used to decompose frequency features and other image-based decomposition methods. Liu et al. [19] first proposed a single-image snow removal network based on deep learning, which can be divided into two stages: residual restoration and translucent restoration. Then Zhang et al. [20] proposed a Deep Dense Multi-scale network (DDMSNet) using semantics and depth priors for snow removal.

Before being introduced into the field of computer vision, Transformer [21] was an important neural network model in the field of natural language processing. The success of Visual Transformer (ViT) [22] also borrows heavily from its application in the field of natural language processing. Dosovitskiy et al. [22] first introduced the Transformer structure into the visual realm and outperformed convolutional neural networks. Subsequently, the application of Transformer in vision covered many research fields [23–27]. In terms of low-level vision tasks, Transweather [28] and Restormer [29] have achieved quite good image restoration by combining convolutional neural networks with visual transformers.

Influenced by deep learning and generative adversarial network, the generator network structure of the traditional GAN is improved in this paper to incorporate the transformer attention mechanism. A variety of Loss functions, such as SSIM Loss, MSE Loss and adversarial Loss, are applied to improve the quality of snow removal images. The effectiveness of the proposed enhanced network structure is verified by the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) on the Snow100K dataset.

2 Related Works

2.1 Degraded Image Formation

The early formula for rainy and snowy image can be expressed in Eq. (1):

$$O = B + S \tag{1}$$

O represents an image with noise, B represents a clean background layer, and S represents a noise (rain streaks or snowflakes) layer. But this model can lead to excessive smoothing where there are multiple densities and no noise in an image. Later, researchers put forward a new model, which can be expressed in Eq. (2):

$$O = B + SR \tag{2}$$

where R represents some values of 0 or 1, representing different regions, respectively, 0 represents non-noise regions, and 1 represents noisy regions. Furthermore, due to the superposition of rain streaks in extreme weather conditions, more complex models are needed to represent them. The models representing multiple rain streaks are shown in Eq. (3):

$$O = \alpha \left(B + \sum_{t=1}^{s} S_t \right) + (1 - \alpha) A \tag{3}$$

where t is the index of the rain-streak layers, A is the global atmospheric light, α is the atmospheric transmission.

The model for snowy image particles adopted in this paper is shown in Eq. (3). Suppose that image I with snowflake noise is composed of a snow-free image y and an independent snow mask z. Then, the image with snowflake noise can be described in Eq. (4):

$$I = (1 - \mathrm{z}) \circ y + A \circ r \tag{4}$$

among them, $\circ$ represents the multiplication of elements between matrices, I $\in R^{C\times M\times N}$ represents the image occluded by snow particles, y $\in R^{C\times M\times N}$represents the clean background layer, r $\in R^{C\times M\times N}$represents the snow particle layer, and A $\in (0, 1)^{C\times M\times N}$represents the transparency matrix.

2.2 Transformer Model

The transformer model was proposed in [21], including two parts: encoder and decoder, as shown in Fig. 1. The encoder part is composed of a multi-head self-attention mechanism and a fully connected feedforward network. Residual connection and Layer normalization operations are used in each module. Instead of using the recurrent neural network in the encoder, there is a self-attention mechanism. The calculation method can be described in Eq. (5):

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

where Q represents the query matrix, K represents the key value matrix, V represents the value matrix, and dividing by $\sqrt{d_k}$ is more convenient. The multi-head self-attention mechanism comprises multi-group attention mechanisms, which extract and splices the relationships among Q, K, V. The formula can be described as follows:

$$\begin{gathered} MultiHead\,(Q, K, V) = Concat\,(head_1, head_n)\; W^{\circ} \\ \text{wherehead}_i = AttentionQW_i^Q, KW_i^K, VW_i^V \end{gathered} \tag{6}$$

where $W_i^Q \in R^{d_{model}*d_q}$, $W_i^K \in R^{d_{model}*d_k}$, $W_i^V \in R^{d_{model}*d_v}$, $W^O \in R^{hd_v*d_{model}}$, d_{model} represents the number of hidden units of the model. $head_i$ illustrates the transformation matrix of the heads.

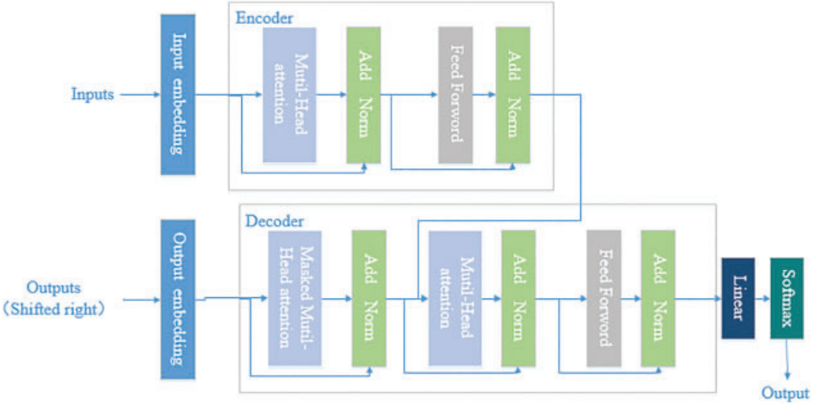

Figure 1: Structure of transformer

The Add & Norm layer consists of add and norm, which the formula can be expressed as:

$$\begin{aligned}&LayerNorm\left(X + MultiHeadAttention\left(X\right)\right)\\&LayerNorm\left(X + FeedForward\left(X\right)\right)\end{aligned} \tag{7}$$

X indicates the input of Multi-Head Attention or Feed Forward Network, and $FeedForward\left(X\right)$ and $MultiHeadAttention\left(X\right)$ illustrate the output. Add represents $X + MultiHeadAttention\left(X\right)$, which is a residual connection. Norm represents Layer Normalization.

The feed forward network module of the framework comprises a two-layer full connection and RELU activation function. The formula can be described as follows:

$$FFN\left(x\right) = \max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{8}$$

where b_1 and b_2 are different bias, W_1 and W_2 are different weights.

2.3 Generative Adversarial Network

The GAN [30] is an unsupervised learning model consisting of generator and discriminator, as shown in Fig. 2. The generator is constantly optimized to generate more realistic simulated data to deceive the discriminator. Meanwhile, the discriminator is updated and iterated to make its judgment more accurate. Eq. (8) can describe the antagonistic relationship between these two:

$$min_G max_D V\left(D, G\right) = E_{x\sim P_{data(x)}}\left[logD\left(x\right)\right] + E_{z\sim P_{data(z)}}\left[\left(1 - logD\left(G\left(z\right)\right)\right)\right] \tag{9}$$

among them, x represents the input real data, z illustrates the input noisy data (gaussian noise), $E_{x\sim P_{data}(x)}$ means the expectation of the input real data, $D\left(\cdot\right)$ represents the output of the discriminator, and $G\left(\cdot\right)$ represents the output of the generator.

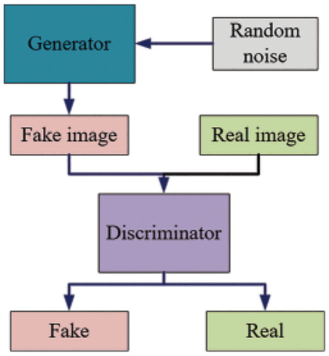

Figure 2: Structure of traditional GAN

Generative adversarial networks have been widely used in many deep learning tasks since their inception. With the in-depth study, the generative adversarial network also has many variants, including DCGAN [31], WGAN [32], CGAN [33], improved WGAN [34] and other network models. DCGAN combines a GAN network with convolutional neural networks for the first time to improve the stability of the network and has achieved very good results in the field of image processing. WGAN theoretically explains the EM distance and solves the disadvantage of instability in the training process of the GAN. The improved WGAN uses the gradient penalty instead of weight clipping and can generate higher-quality samples while accelerating convergence. CGAN can generate real samples that meet the constraints, which further improves the performance of generative adversarial networks. Compared with traditional generative adversarial networks, conditional generative adversarial networks input additional conditional variables into discriminators and generators, so they can avoid pattern collapse to a certain extent and make the generated images more realistic. For the underlying vision task, the conditional generative adversarial network behaves as feeding a degraded image into the generator. For example, for a single image rain pattern removal task, the input image received by the generator is a rain pattern attachment image; For single-image desnow task, the input image is the snowflake attachment image, not random noise. In this paper, we use conditional generative adversarial network to recover sharp images in an end-to-end manner, and in order to take into account the overall performance of the model, the WGAN network model is adopted simultaneously.

2.4 U-Net

The U-Net [35] structure was first proposed to be applied in medical image segmentation. This network structure can effectively extract context information and location information. The network structure diagram is shown in Fig. 3. The left side of the structure (encoder) is the feature extraction part, that is, the convolution and pooling operations. The second part (decoder) is for up-sampling processing and feature stitching. The model can extract different spatial scale features when processing image features. At the same time, the low-resolution image features can be fused and stacked with the up-sampled high-resolution features by skip connection, and the rich multi-scale features can finally be obtained.

Due to its ability to extract multi-scale features and simple implementation, the U-Net network model has been widely used in various low-level tasks, such as image restoration, and has achieved excellent performance.

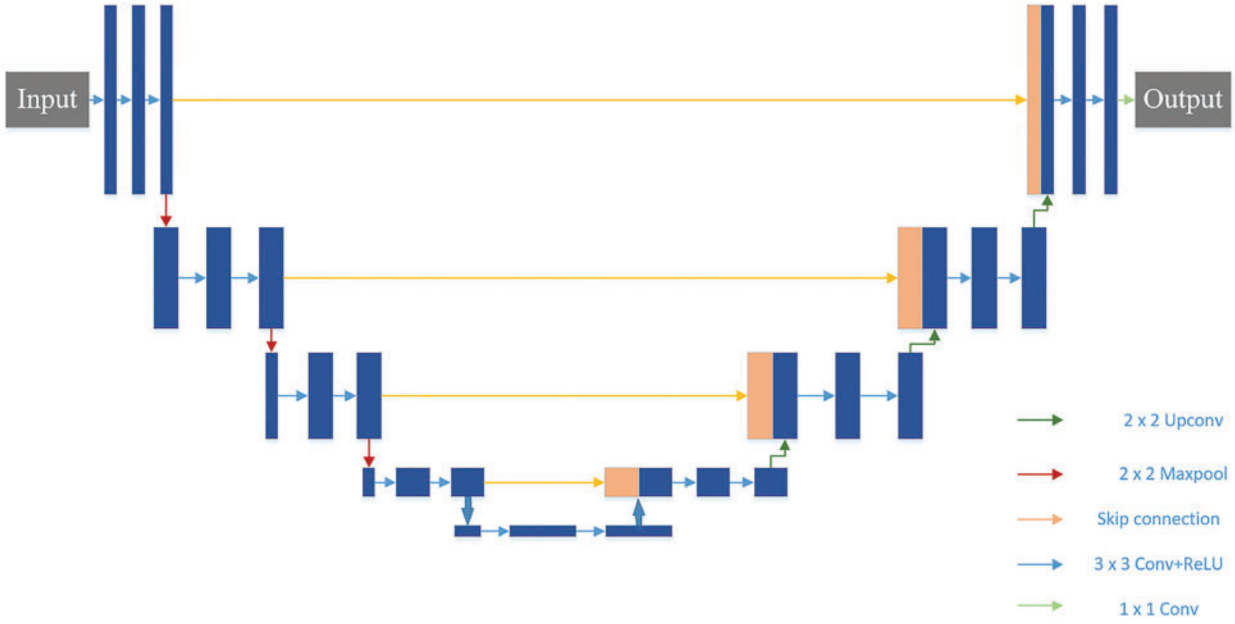

Figure 3: Structure of U-Net. As a symmetric neural network, U-Net can first extract the abstract features of the input image, compress them, and then gradually obtain the feature information that needs to be output at a specific size

3 Proposed Method

3.1 Network Structure of the Generator and Discriminator

Inspired by the attention module and U-Net, the network structure of the generator proposed in this paper is shown in Fig. 4. Enter a 256 × 256 color image with three channels, after the convolution operation with a convolution kernel size of 7 × 7 and a step size of 1, two downsampling operations are performed, and the number of channels becomes 256. Then the patch embedding operation is performed, and the output image is generated after 5 Transformer Block structures, two layers of upsampling and one layer of convolution. Finally, the predicted residual image and the input image are added globally by element and output as an RGB image.

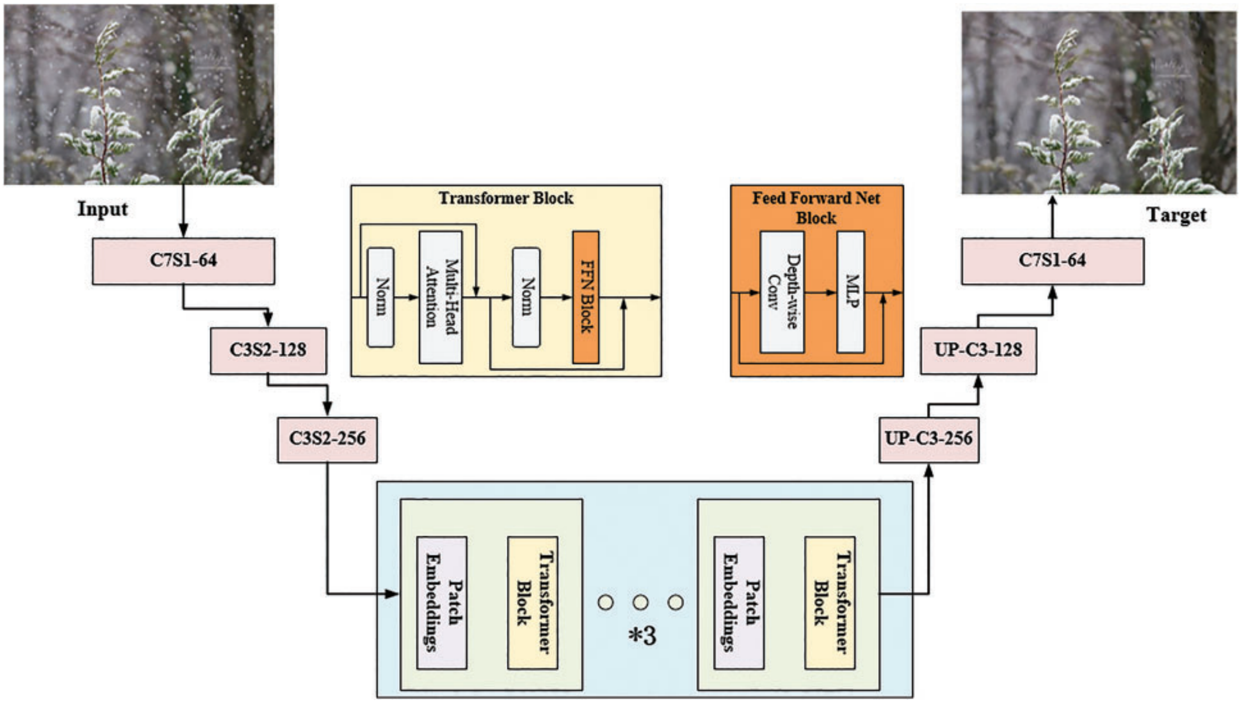

Figure 4: Structure of generator. The features in the input picture are first extracted by a multi-layer convolutional neural network, then further processed by the Transformer module stacked in the middle, and finally upsampled by a multi-layer convolutional neural network to obtain a clear image after removing snowflakes

This paper uses the classical PatchGAN [33] as the discriminator. In the original GAN, the discriminator is designed to output only one evaluation value, which is the overall evaluation of the image generated by the generator. PatchGAN is a discriminator that punishes only within the range of patches. It performs a convolution operation on images and classifies each n × n patch in the image to judge whether they are true or false. PatchGAN designed five convolution layers to superimpose patches and expand the receptive field to 70 times. See Fig. 5 for details, the image is effectively modelled as a Markov random field, which can retain the high resolution and details of the image.

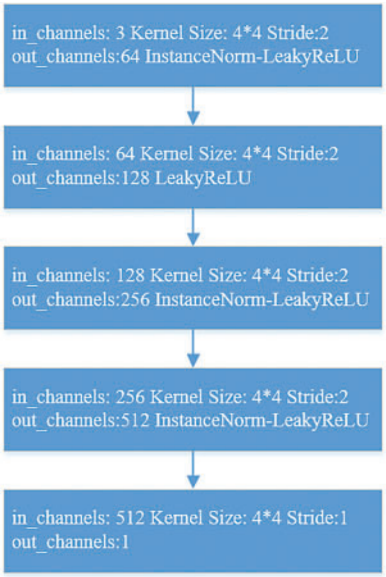
in_channels: 3 Kernel Size: 4*4 Stride:2
out_channels:64 InstanceNorm-LeakyReLU
in_channels: 64 Kernel Size: 4*4 Stride:2
out_channels:128 LeakyReLU
in_channels: 128 Kernel Size: 4*4 Stride:2
out_channels:256 InstanceNorm-LeakyReLU
in_channels: 256 Kernel Size: 4*4 Stride:2
out_channels:512 InstanceNorm-LeakyReLU
in_channels: 512 Kernel Size: 4*4 Stride:1
out_channels:1

Figure 5: Structure of PatchGAN

3.2 Loss Function

The choice of loss function has a great influence on the training result. This paper uses a combination of different loss functions to improve the performance. In this paper, MSE loss function is used, which is expressed in Eq. (10):

$$L_{mse} = \frac{1}{N}\sum_{i=1}^{N} |y - f(x)|_2^2 \tag{10}$$

where N represents the number of sample data, y represents the actual value of the sample, and $f(x)$ represents the predicted value of the sample.

The SSIM loss in [36] is used to enhance further and refine the structure awareness of the network model, which can be expressed in Eq. (11):

$$L_{SSIM} = 1 - SSIM(f(x), y) \tag{11}$$

The perceptual loss [37] measures the gap between the feature maps of restored images and clear images after VGG-19 [38] conv3.3. It can be expressed in Eq. (12):

$$L_{Perp} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}\left(\Phi_{i,j}\left(I^S\right)_{x,y} - \Phi_{i,j}\left(G_{\theta_G}\left(I^B\right)\right)_{x,y}\right)^2 \tag{12}$$

where $\Phi_{i,j}$ is the feature map obtained by the convolution layer within the VGG19 [38] network and $W_{i,j}H_{i,j}$ are the dimensions of the feature maps.

The loss function of the generator and discriminator of the generative adversarial network can be defined in Eq. (13):

$$L_{ADV} = -D\left(f\left(x\right)\right)$$
$$L_{Generator} = L_{ADV}$$
$$L_{Discriminator} = E_{f(x)\sim P_G}\left[D\left(f\left(x\right)\right)\right] - E_{y\sim P_r}\left[D\left(y\right)\right] \tag{13}$$

where D represents the discriminator model, and P_G represents the distribution of generators, and P_r represents the distribution of actual data.

4 Experimental Results and Analysis

4.1 Experimental Details

The comparison and ablation experiments in this paper are carried out on the server configured as NVIDIA RTX3060. The server system is CentOS7, and the overall framework is Pytorch. Adam optimizer is used in ablation experiments and comparison experiments. Among them, the initial learning rate of the ablation experiment part is $1e^{-4}$, with 200 epochs of training. The learning rate warm-up is used to prevent the network from over fitting. The batch is set to 8, and the image size is 128×128. In the final experiment, the learning rate is $2e^{-4}$, the batch is set to 4, and the image size is 256×256. The learning rate is attenuated by cosine annealing. After several iterations of training, the final learning rate is linearly attenuated to 10^{-6}. In the training process, the images are flipped randomly, horizontally and vertically to increase the generalization performance and robustness of the model. The structure similarity and peak signal-to-noise ratio are used as evaluation indexes. This paper uses Snow100K data set to evaluate the effect of the snow removal experiment. The Snow100K dataset contains 1×10^5 composite snow images, as well as corresponding clear images and snow masks. The test set is divided into three parts, Snow100K-S, Snow100K-M, and Snow100K-L, which contain snow grains from less to more.

4.2 Ablation Experiments

For the structure part of the generator, we replaced it with residual blocks and compared the performance. At the same time, we compare the two models that estimate the snowflake mask at the same time, and find that this will cause performance loss, so the final model structure only estimates the prediction image of snow removal. The loss function of this part is MSE. The results of the ablation experiment are shown in Table 1.

Table 1: Ablation study of network structure

	ResBlock PSNR/SSIM	Mask residual PSNR/SSIM	Mask PSNR/SSIM	Transformer PSNR/SSIM
L	21.32/0.6668	18.80/0.3592	24.96/0.6874	25.10/0.7121
M	23.31/0.7450	19.24/0.4148	27.58/0.7938	28.07/0.8112
S	23.86/0.7524	19.30/0.4306	28.48/0.8098	29.12/0.8260

As shown in the above table, ResBlock and Transformer represent the BackBone used in the network, respectively; Mask represents simultaneous prediction of snowflake mask; Mask Residual indicates the predicted value of the network will be as residual snow image, that is, the final desnowing image is obtained by subtracting the entire neural network prediction image from the input snowflake attachment image. Experimental results show that this method significantly reduces the image quality, so it is not used in the end.

For the loss function selection part of the generator, we compared a variety of loss functions, including perceptual loss, adversarial loss, and structural similarity loss function, as shown in Table 2.

Table 2: Ablation study of loss function

	Perceptual loss PSNR/SSIM	Adversarial loss PSNR/SSIM	Perp.+ Adv.PSNR/SSIM	Adv.+ SSIM loss PSNR/SSIM
L	24.82/0.7061	24.99/0.7089	24.84/0.7063	24.91/0.7135
M	27.95/0.8023	28.03/0.8096	27.92/0.8029	28.08/0.8106
S	28.95/0.8144	29.07/0.8246	28.89/0.8146	29.21/0.8237

The final experimental results show that the best effect can be achieved using MSE loss to combine adversarial loss and structural similarity loss. Therefore, the final loss function is shown in the Eq. (14):

$$Loss_{all} = \alpha Loss_{mse} + \beta Loss_{ADV} + \gamma Loss_{SSIM} \tag{14}$$

4.3 Comparison Experiments

In the comparison experiment part, we compared our method with various related methods, including Zheng et al. [39], DerainNet [11], DehazeNet [40], DeepLab [41], JORDER [10] and other methods, as shown in Table 3. At the same time, Figs. 6 and 7 show the visual effect comparison between ours and other methods. The snowflake mask in the bottom row of Fig. 7 is output using the Mask network structure in the ablation experiment. In the overall method, our method does not predict the snowflake mask. From the table and figure, we can see that our method has a stronger snowflake removal ability than other methods and can restore the results closer to the real image.

Table 3: Quality evaluation results of different algorithms on Snow100K dataset

Method	L PSNR/SSIM	M PSNR/SSIM	S PSNR/SSIM
Synthesized	18.67/0.73	22.82/0.83	25.10/0.86
Zheng [39]	19.95/0.72	22.99/0.79	24.32/0.81
DerainNet [11]	19.18/0.74	23.36/0.84	25.74/0.86
DehazeNet [40]	22.61/0.79	24.16/0.86	24.96/0.88
DeepLab [41]	21.29/0.77	24.36/0.85	29.94/0.87
JORDER [10]	23.40/0.80	24.97/0.87	25.62/0.88
Ours	27.27/0.83	30.41/0.89	31.73/0.91

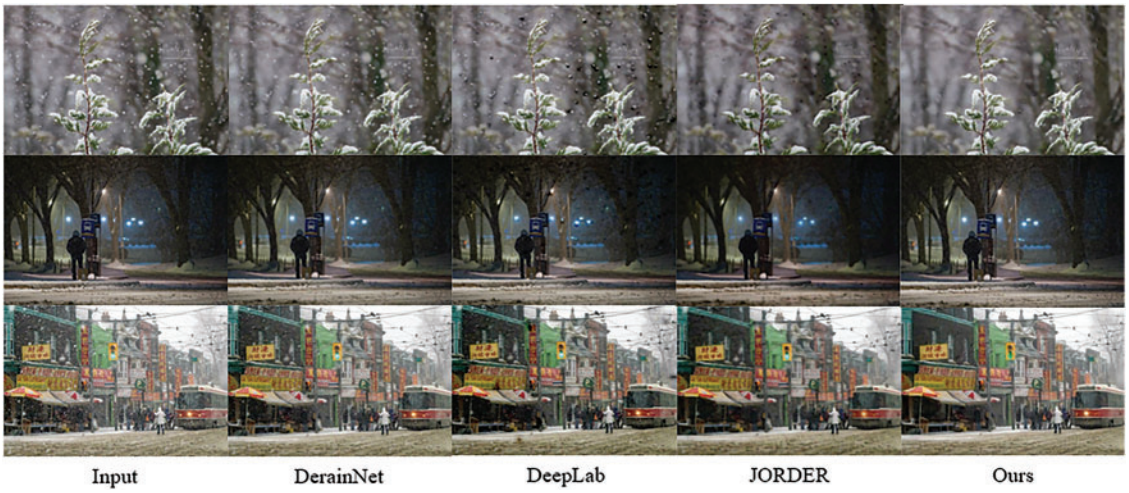

Figure 6: The visual effect of our algorithm and other algorithms on synthetic images

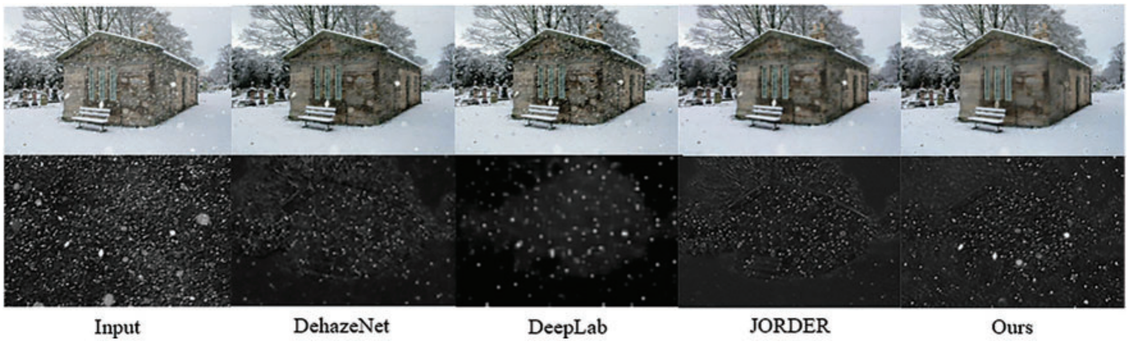

Figure 7: Effect comparison between different algorithms in natural images. The image of the underlying row is a snowflake mask predicted by the corresponding method

5 Conclusion

Aiming at the problem that snowflakes in the image reduce the image quality, this paper proposes a single image snow removal algorithm based on the transformer and conditional generative adversarial network. This paper first introduces the related works, such as vision transformer, snowflake image formation process, and generative adversarial network. Secondly, we present the transformer-based conditional generative adversarial network model proposed in this paper. Finally, through ablation studies and comparative experiments, the rationality and effectiveness of the proposed method are proved. Compared with other algorithms, the algorithm proposed in this paper can effectively remove snowflakes from images. More in-depth content will be carried out in the future.

Funding Statement: This work is supported by School of Computer Science and Technology, Shandong University of Technology. This paper is supported by Shandong Provincial Natural Science Foundation, China (Grant Number ZR2019BF022) and National Natural Science Foundation of China (Grant Number 62001272).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Yu, B., Yin, H., Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. https://doi.org/10.48550/arXiv.1709.04875
2. Li, Y., Yu, R., Shahabi, C., Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. https://doi.org/10.48550/arXiv.1707.01926
3. Ahmed, E., Gharavi, H. (2018). Cooperative vehicular networking: A survey. *IEEE Transactions on Intelligent Transportation Systems, 19(3),* 996–1014. https://doi.org/10.1109/TITS.2018.2795381
4. Chen, M., Wang, T., Ota, K., Dong, M., Zhao, M. et al. (2020). Intelligent resource allocation management for vehicles network: An A3C learning approach. *Computer Communications, 151(3),* 485–494. https://doi.org/10.1016/j.comcom.2019.12.054
5. Bila, C., Sivrikaya, F., Khan, M. A., Albayrak, S. (2016). Vehicles of the future: A survey of research on safety issues. *IEEE Transactions on Intelligent Transportation Systems, 18(5),* 1046–1065. https://doi.org/10.1109/TITS.2016.2600300
6. Luo, Y., Xu, Y., Ji, H. (2015). Removing rain from a single image via discriminative sparse coding. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3397–3405. Santiago, Chile. https://doi.org/10.1109/ICCV.2015.388
7. Kim, J. H., Lee, C., Sim, J. Y., Kim, C. S. (2013). Single-image deraining using an adaptive nonlocal means filter. *2013 IEEE International Conference on Image Processing*, pp. 914–917. Melbourne, Australia, IEEE. https://doi.org/10.1109/ICIP.2013.6738189
8. Chen, B. H., Yeh, W. C., Kuo, S. Y. (2017). Error-optimized sparse representation for single image rain removal. *IEEE Transactions on Industrial Electronics, 64(8),* 6573–6581. https://doi.org/10.1109/TIE.2017.2682036
9. Li, Y., Tan, R., Guo, X., Lu, J., Brown, M. S. (2016). Rain streak removal using layer priors. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2736–2744. Las Vegas, USA. https://doi.org/10.1109/CVPR.2016.299
10. Yang, W., Tan, R. T., Feng, J., Liu, J., Guo, Z. et al. (2017). Deep joint rain detection and removal from a single image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1357–1366. Hawaii, USA. https://doi.org/10.1109/CVPR.2017.183
11. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J. (2017). Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing, 26(6),* 2944–2956. https://doi.org/10.1109/TIP.2017.2691802
12. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H. (2018). Recurrent squeeze-and-excitation context aggregation net for single image deraining. *European Conference on Computer Vision*, pp. 262–277. Munich, Germany. https://doi.org/10.1007/978-3-030-01234-2_16
13. Zhang, H., Patel, V. M. (2018). Density-aware single image de-raining using a multi-stream dense network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 695–704. Salt Lake City, USA. https://doi.org/10.1109/CVPR.2018.00079
14. Tan, R. T. (2008). Visibility in bad weather from a single image. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. Alaska, USA. https://doi.org/10.1109/CVPR.2008.4587643
15. Fattal, R. (2008). Single image dehazing. *ACM Transactions on Graphics, 27(3),* 1–9.
16. Huang, S. C., Ye, J. H., Chen, B. H. (2015). An advanced single-image visibility restoration algorithm for real-world hazy scenes. *IEEE Transactions on Industrial Electronics, 62(5),* 2962–2972. https://doi.org/10.1109/TIE.2014.2364798

17. Bossu, J., Hautiere, N., Tarel, J. P. (2011). Rain or snow detection in image sequences through use of a histogram of orientation of streaks. *International Journal of Computer Vision, 93(3),* 348–367. https://doi.org/10.1007/s11263-011-0421-7
18. Rajderkar, D., Mohod, P. S. (2013). Removing snow from an image via image decomposition. *2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, pp. 576–579. Tirunelveli, India, IEEE. https://doi.org/10.1109/ICE-CCN.2013.6528565
19. Liu, Y. F., Jaw, D. W., Huang, S. C., Hwang, J. N. (2018). DesnowNet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing, 27(6),* 3064–3073. https://doi.org/10.1109/TIP.2018.2806202
20. Zhang, K., Li, R., Yu, Y., Luo, W., Li, C. (2021). Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing, 30,* 7419–7431. https://doi.org/10.1109/TIP.2021.3104166
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.
23. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. et al. (2020). End-to-end object detection with transformers. *European Conference on Computer Vision*, pp. 213–229.
24. Zhu, X., Su, W., Lu, L., Li, B., Wang, X. et al. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv: 2010.04159.
25. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z. et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890.
26. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y. et al. (2021). Pre-trained image processing transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12299–12310. Online.
27. Zhou, L., Zhou, Y., Corso, J. J., Socher, R., Xiong, C. (2018). End-to-end dense video captioning with masked transformer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8739–8748. Salt Lake City, USA.
28. Valanarasu, J. M. J., Yasarla, R., Patel, V. M. (2022). Transweather: Transformer-based restoration of images degraded by adverse weather conditions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2353–2363. New Orleans, USA.
29. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S. et al. (2022). Restormer: Efficient transformer for high-resolution image restoration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739. New Orleans, USA.
30. Zhou, M., Lin, Y., Zhao, N., Jiang, Q., Yang, X. et al. (2020). Indoor WLAN intelligent target intrusion sensing using ray-aided generative adversarial network. *IEEE Transactions on Emerging Topics in Computational Intelligence, 4(1),* 61–73. https://doi.org/10.1109/TETCI.2019.2892748
31. Radford, A., Metz, L., Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv: 1511.06434.
32. Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein GAN. https://doi.org/10.48550/arXiv.1701.07875
33. Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134. Hawaii, USA. https://doi.org/10.1109/CVPR.2017.632
34. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C. (2017). Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028.

35. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Munich, Germany.
36. Zhao, H., Gallo, O., Frosio, I., Kautz, J. (2016). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging, 3(1),* 47–57. https://doi.org/10.1109/TCI.2016.2644865
37. Johnson, J., Alahi, A., Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*, pp. 694–711. Amsterdam, Netherlands.
38. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.
39. Zheng, X., Liao, Y., Guo, W., Fu, X., Ding, X. (2013). Single-image-based rain and snow removal using multi-guided filter. *International Conference on Neural Information Processing*, pp. 258–265. Berlin, Heidelberg.
40. 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. https://doi.org/10.1109/TIP.2016.2598681
41. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4),* 834–848. https://doi.org/10.1109/TPAMI.2017.2699184