Neural Computing and Applications (2022) 34:7209–7225
https://doi.org/10.1007/s00521-021-06841-7

ORIGINAL ARTICLE

Generative adversarial network with hybrid attention and compromised normalization for multi-scene image conversion

Jinsheng Xiao[1] · Shuhao Zhang[1] · Yuntao Yao[1] · Zhongyuan Wang[1] · Yongqin Zhang[2] · Yuan-Fang Wang[3]

Received: 26 June 2021 / Accepted: 12 December 2021 / Published online: 29 January 2022

Abstract
In order to generate high-quality realistic images, this paper proposes an image conversion algorithm based on hybrid attention generation adversarial network. This network is composed of the generator and discriminator, both of which are jointly trained through the loss function. The generator is constructed by using a three-stage structure of down-sampling, residual and up-sampling blocks where the residual block uses a hybrid attention mechanism. The compromised instance and layer normalization is also proposed by weighting the output of the fully connected layer. The multi-scale PatchGAN is introduced as the discriminator. The proposed network can produce more realistic images using a new loss function, which comprises four iterms: generation adversarial loss, L_1 regularization loss, VGG loss and feature matching loss. The experimental results demonstrated that the proposed method can produce more realistic and detailed images than the state-of-the-art methods.

Keywords Image conversion · Deep learning · Generative adversarial networks · Image generation

1 Introduction

Many computer vision issues can be regarded as image conversion issues, which map the image in one domain to the corresponding image in another domain. In fact, it is the mapping between pixels. For example, super-resolution can be considered a problem of mapping a low-resolution image to a corresponding high-resolution image, and image coloring can be viewed as mapping a gray-scale image to a corresponding color image. Thus, image conversion is an important research field of computer vision. Specifically, image conversion is also called as image-to-image translation, including multiple types, such as style conversion, color conversion, content conversion and scene conversion. These image transformations, texture adjustments and stylized editing are used in art, scientific research and engineering.

Image conversion issue can be studied in supervised and unsupervised learning environments. In unsupervised learning, there are only two independent images instead of a pair of image training sets. Unsupervised image-to-image conversion requires more accurate and complex network to produce realistic images, which is difficult to implement, like CycleGAN [1] and MUNIT [2]. In supervised learning, the corresponding images can be trained in different domains [3]. The mapping relationship between the generated image and the input image pixels is more accurate through supervised learning method. However, it is difficult to obtain matching image pairs in natural environments. Unsupervised learning can solve this problem, but the lack of paired images cause the uncontrollable results. In the specific task targeted by the algorithm in this paper, we can obtain image pairs, so we use supervised learning

✉ Jinsheng Xiao
xiaojs@whu.edu.cn

✉ Yongqin Zhang
zhangyongqin@nwu.edu.cn

Yuan-Fang Wang
yfwang@ucsb.edu

[1] School of Electronic Information, Wuhan University, Wuhan 430072, People's Republic of China

[2] School of Information Science and Technology, Northwest University, Xi'an 710127, People's Republic of China

[3] Department of Computer Science, University of California, Santa Barbara, CA 93106, USA

framework to produce more clear and natural images. Supervised learning is performed using a convolutional neural network (CNN) [4], which also minimizes the loss function when generating images as a standard for network adjustment. However, CNN requires minimizing the euclidean distance between the predicted image and the ground truth, which may produce blurred results [5, 6] as for the Euclidean distance averages the output of all pixels. Therefore, it is a tricky problem to develop a specific loss function for a specific conversion task in CNN network. GAN (Generative Adversarial Networks) comprises two specific network structures, namely generative model and discriminative model,which are mutually adversarial. The discriminator is used to judge whether an image is from the generator or the real world. The generator is used to synthesize images that attempt to deceive the discriminator as much as possible [7]. GAN [8] attempts to classify that the output image is real or fake, while training the generated model. Traditionally, the loss function of GAN has to be used for different types of tasks. Under such a background, using the optimized GAN network for supervised learning and image conversion has gradually become a research hotspot. This paper proposes a novel image scene conversion algorithm. The multi-scene image conversion algorithm flow chart designed in this paper is shown in the Fig. 1. The contributions of this paper are as follows.

1. We propose a novel hybrid attention module that uses different pooling methods to integrate spatial and channel information. It is a lightweight module group that can be easily embedded in other networks. Inspired by the current advanced normalization methods, especially adaptive instance normalization (AdaIN [9]), we designed a new normalization method called CILN (Compromised Instance and Layer Normalization) that is a compromise between instance and layer normalization.
2. With the hybird attention module and CILN embedded, we propose a novel image scene conversion algorithm that can make fast multi-scene conversion.
3. We conduct four scene conversion experiments, which are image hazing, satellite image to map conversion, optic image to SAR conversion and day to night conversion. Our algorithm has certain advantages in subjective observations and objective indicators compared with the five state-of-the-art algorithms.

2 Related work

2.1 Generative adversarial networks

Generative Adversarial Network (GAN) [7, 10] is applied to a series of tasks such as image generation [11], image restoration [12] and image translation [13, 14], in which GAN has obtained impressive results. In training, the generator aims to generate realistic images to deceive the discriminator, and the discriminator tries to distinguish the generated image from the real image. Mixture variational

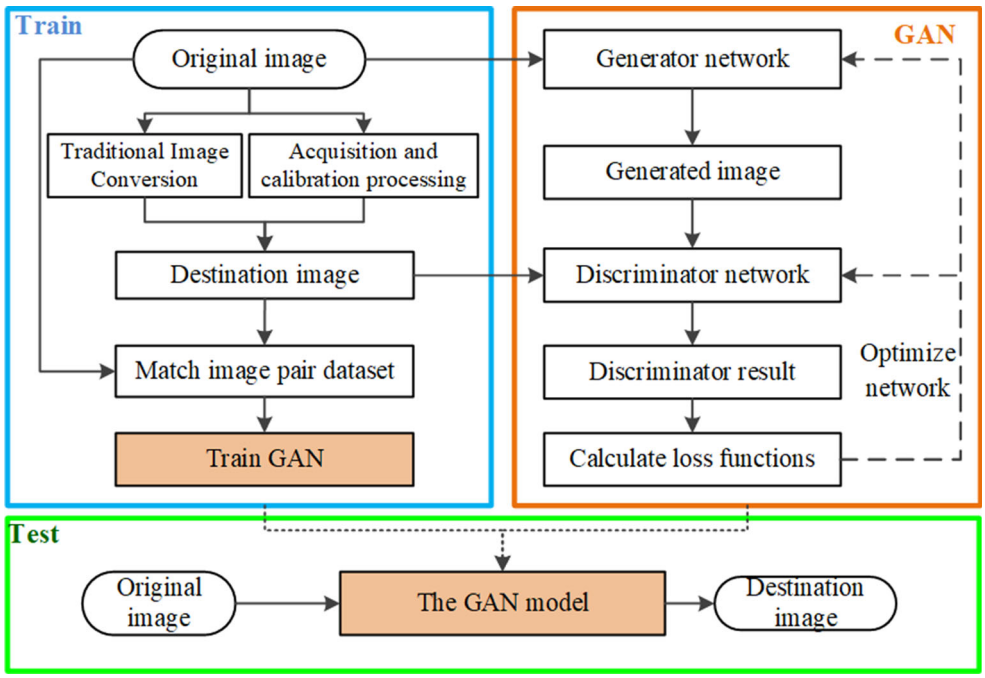

Fig. 1 The flowchart of proposed algorithm

autoencoders (VAE) [15] as a form of depth generation model is a generated network structure that is based on variation Bayes. VAE play an important role in unsupervised learning and representation learning. But the isotropic generative model in VAE cannot sufficiently utilize the latent representative space. The VAE improve the generative performance by enlarging the latent representative space. VAEGAN [16] presents an autoencoder that leverages learned representations to better measure similarities in data space. By combining a VAE with a generative adversarial network, VAEGAN can use learned feature representations in the GAN discriminator as basis for the VAE reconstruction objective. Training robust GAN is a non-trivial task due to the problem of mode collapse. Xu et al. [17] propose a novel generative adversarial network-based model, InjectionGAN, to learn a many-to-many mapping. Li et al. [18] propose a new approach to training GAN to overcome mode collapse by employing a set of generators, an encoder and a discriminator. Various multi-stage generation models [19] and better training targets [20] have been proposed to generate more realistic images. Inspired by their successes, we propose a new novel method for image conversion with a new attention module and a new combined loss function.

2.2 Image conversion

Many researchers have leveraged adversarial learning for image-to-image translation [21], whose goal is to translate an input image from one domain to another domain given input-output image pairs as training data. For example, Gatys et al. [22] proposed an art-style transformation algorithm whose main step is to solve the problem of extracting deep features from Gram matrix from content images and style images. Li et al. [23] propose a Simplified Unsupervised Image Translation (SUIT) model for domain adaptation on semantic segmentation, which can significantly improve the performance of the model on the target domain. Many style-transfer algorithms [24, 25] surpass Gatys' method in performance and speed. High-fidelity image stylization is related to image-to-image translation problems [26], and its goal is to learn to translate images from one domain to another. Luan et al. [27] improve the realism of the stylized output calculated by the style conversion algorithm by adding a new loss function to the optimization target, thereby better preserving the local structure in the content photo. Pix2pix [3] uses conditional GAN for different image transformations. However, Pix2pix learns a one-to-one mapping between a domain to another domain. Therefore, when the gap between the input data and the training data is large, the result is likely to be meaningless. Based on Pix2pix, Wang et al. [19] propose a new coarse-to-fine generator and multi-scale discriminator architectures suitable for conditional image generation at a much higher resolution. In the absence of training pairs, various image-to-image translation methods [28] have also been proposed. CycleGAN [1] uses unpaired images for training and proposes a loss of cycle consistency, which is widely used. Compared to Pix2pix, the CycleGAN does not require pairs of images, which can reduce the cost of making samples. MUNIT [2] decomposes the image through image coding to obtain an invariable content space and a variable style space, thereby achieving a many-to-many mapping relationship. Chen et al. [29] pointed out that it is difficult to generate high-resolution images for the conditional GAN training due to training instability and optimization problems. In order to avoid this difficulty, a perceptual loss is proposed [30]. The resulting image is high resolution, but often lacks detail and realistic texture.

2.3 Attention module

Attention mechanism has recently improved the success of various computer vision tasks recently and continues to be an omnipresent component in state-of-the-art models. In broad terms, attention can be viewed as guidance to bias the allocation of available processing resources toward the most informative components of an input. In the field of image generation, learning high-dimensional and complex image distribution through the attention mechanism has proved to be effective [31, 32].The self-attention module [33] uses the weighted sum of all feature points to reconstruct each feature point, significantly improving the correlation between distant relevant regions in the feature map. The Research Institute of Defiance proposes a new method [34] for modeling long-distance relationships through normalization on conditional image generation tasks, it expands based on instance normalization and describes long-range dependence through attention normalization(AN).

2.4 Normalization

Normalization is widely used in deep learning, especially in the field of computer vision. It can speed up training and improve the accuracy of the network model. There are currently four basic normalization methods, namely Batch Normalization, Layer normalization, Instance normalization and Group normalization. It is worth noting that some novel normalization methods have appeared in image generation tasks, including conditional batch normalization (CBN [34]), adaptive instance normalization (AdaIN [9]), spatially-adaptive (de)normalization (SPADE [35]), etc. Generally speaking, these conditional normalization methods perform better in specific tasks.

3 Image conversion algorithm based on generative adversarial networks

The proposed image conversion algorithm based on GAN in this paper could be divided into two stages of training and testing. In the training phase, the generator and the discriminator are cross-trained and learn from each other to get an excellent generator. In the test phase, we input images to the generator to get the converted images. Our model has experienced no model collapse problem. The main reason is the diversity of the training samples and the closely matched complexity of the generating network and the discriminant network. This section introduces our network structure in detail from three aspects: generator, discriminator and loss function. The multi-scene image conversion algorithm flow chart is shown in the Fig. 1.

3.1 Generator structure

In terms of the generator, we adopt a three-stage structure of down-sampling, residual and up-sampling blocks. Among them, down-sampling block initially extracts image features. The residual module performs deep feature extraction. Up-sampling block reconstructs the feature map and finally outputs the generated image. It is worth noting that this algorithm introduces a new attention mechanism in the middle of the residual network, which is called hybrid attention. Besides, we have also improved the normalization method of some residual modules. This new normalization method is called CILN (Compromised Instance and Layer Normalization). It is a balance of two normalization methods of instance and layer. The network structure is shown in Fig. 2.

Assuming that there are source domain and target domain spaces {X, Y}, an instance x of the source domain space needs to be converted to the target domain space, and the generator realizes this conversion. According to the network structure, the generator can be divided into three parts: down-sampling, residual module and up-sampling. According to functions, it can be divided into two parts: encoder C_E and decoder C_D. Specifically, the encoder comprises a down-sampling module and a 9-layer residual block. The decoder consists of a 9-layer CILN residual block and an up-sampling module. The down-sampling and up-sampling modules, respectively, include three layers of convolution and deconvolution, and each convolution (deconvolution) layer contains three parts: convolution, normalization and activation function. It is worth noting that the first layer of the down-sampling module and the last layer of the up-sampling module both use a 7×7 convolution kernel. The purpose is to increase the receptive field and capture enough details and context information. The network introduces a hybrid attention module between the encoder and the decode. The hybrid attention module is in the middle of the residual blocks and TILN residual blocks as shown in Fig. 2. The hybrid attention is used to enhance the control of the generator. It pays more attention to the edges and objects in the image, keeping the necessary texture details in the conversion process. The network structure is shown in Fig. 3.

The hybrid attention module includes a global max pooling (GMP) layer, a global average pooling (GAP) layer, a 1×1 convolution, and the RELU activation function. It is a lightweight module group that can be easily embedded in other networks. Suppose $C_E(x)$ is the feature map output by C_E, and $C_E(x)$ is input to the global max pooling and average pooling layers, respectively. Global maximum pooling retains the maximum value in the feature map and plays the role of focusing on essential areas of the image, global average pooling takes a mean value for the feature map. This value reflects the importance of each feature map in the channel to a certain extent. The new

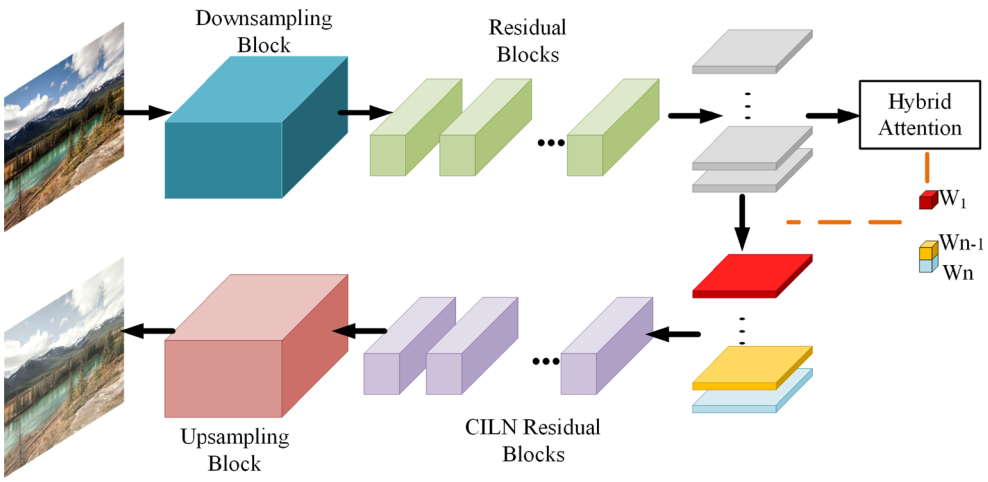

Fig. 2 Generator network structure

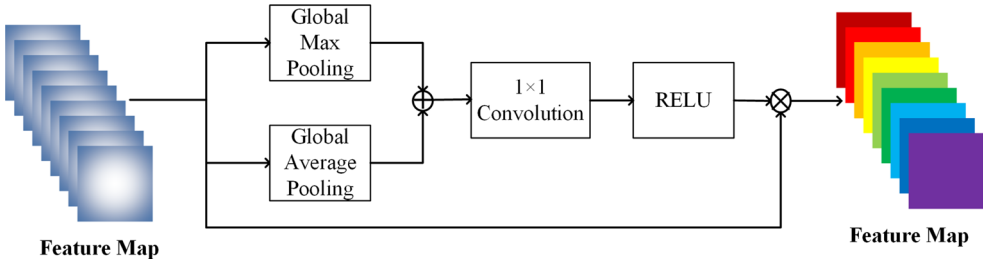

Fig. 3 Hybrid attention network structure

feature maps obtained by the two pooling layers are cascaded, and the potential combination relationship is extracted through 1×1 convolution. Finally, the weight coefficients of different channel feature maps are obtained through the Relu layer and then multiplied by the original feature map $C_{\mathrm{E}}(x)$ to obtain a set of attention feature maps that incorporate the region and channel information.

Our hybrid attention model is unique and different from SEnet [36] and CBAM [37]. Among them, SEnet only pays attention to the difference of the channel characteristics, and the zooming operation causes loss of image information to a certain extent. CBAM integrates spatial and channels attention, but it modularizes the two separately and then operates the two in series on the feature map. In this paper, we use different pooling methods to integrate spatial and channel information within the modules, and the structure is more succinct than CBAM. Later, we will explain this module's function by analyzing the feature maps before and after the attention module.

Suppose the attention module is F, and the learned weight is W. This process can be expressed as:

$$W = F(C_{\mathrm{E}}^{k}(x)) \tag{1}$$

$$o(x) = W \cdot C_{\mathrm{E}}(x) = \left\{ w^{k} \cdot C_{\mathrm{E}}^{k}(x) | 1 \le k \le n_{\mathrm{c}} \right\} \tag{2}$$

where $o(x)$ denotes the feature map output by the attention module, $\cdot$ denotes multiplication, w^{k} denotes the weights learned by the attention module of the k-th channel, $C_{\mathrm{E}}^{k}(x)$ denotes the feature map output by the encoder of the k-th channel. n_{c} is the total number of channels of the encoder output feature map. Then, the decoder accepts $o(x)$ as input, which can be expressed as $C_{\mathrm{D}}(x)$. The decoder includes a CILN residual module and an up-sampling module, where CILN is a normalization method. We replace the IN in the original residual block with CILN to form a new CILN residual block. Inspired by the current advanced normalization methods, especially adaptive instance normalization (AdaIN [9]), we designed a new normalization method. It is a compromise between instance and layer normalization, so we call it CILN(Compromised Instance and Layer Normalization), which the following formula can express:

$$\mathrm{CILN}(\alpha, \delta, \omega, \varphi) = \omega(\delta\alpha_I + (1-\delta)\alpha_L) + \varphi \tag{3}$$

$$\alpha_I = \frac{\alpha - \mu_I}{\sqrt{\sigma_I^2 + \epsilon}}, \alpha_L = \frac{\alpha - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}} \tag{4}$$

where μ_{I}, μ_{L} and σ_{I}, σ_{L} are the mean and standard deviation of the instance and layer normalization methods, respectively. ω and φ are the zoom and translation parameters that are automatically updated by the fully connected layer in the normalization mode. δ is a weight parameter used to weigh the two normalization methods and is also dynamically calculated and updated by the fully connected layer.

Instance normalization (IN) calculates the characteristics of a single instance and a single channel, which makes it easier to maintain the content structure of the source domain image. This method ignores the correlation of features between different channels of an instance. Layer normalization (LN) on the other hand considers the correlation of different channel features. Still, this method calculates the global statistical information of the feature map and ignores the content structure of the source domain image. To combine the advantages of the two and overcome their shortcomings, we proposed CILN. CILN combines IN and LN's advantages by selectively changing content or style information and automatically updates the weight coefficients with reference to AdaIN [9]. This method helps to solve a wide range of image conversion trade-offs.

3.2 Discriminator structure

We used a multi-scale discriminator in this paper as shown in Fig. 4. To be specific, there is a discriminator in every different image scale. Each discriminator is similar in network structure, which can be seen in Fig. 5. For high-resolution images, multi-scale discriminators can improve the receptive field of the network.

In the field of neural networks, people usually use more network layers or larger convolution kernels to expand the receptive field of the network. However, both of these will increase the number of network parameters, resulting in overfitting or even the inability to train. In this paper,

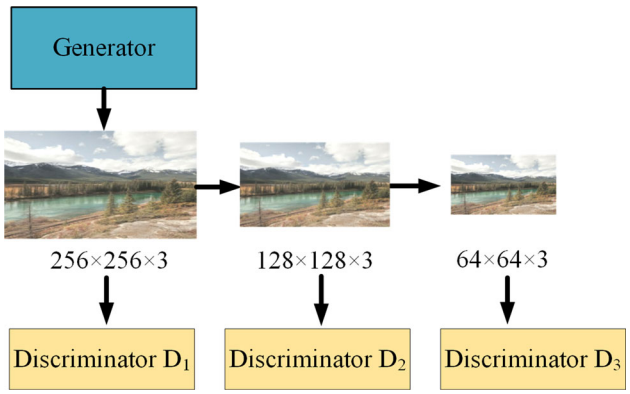

Fig. 4 Multi-scale discriminator

multiple discriminators of different scales are used, and only a few parameters are added to achieve the purpose of increasing the network's receptive field. Take the use of three discriminators as an example, represented by D_1, D_2 and D_3, respectively. D_1 judges the original size image, D_2 judges the image down-sampled once, and D_3 judges the image down-sampled twice. Each single discriminator has similar downsampling and output classification layer, except that the size of the input image is different.

For every single discriminator network, the convolution kernel size of the lower sampling layer is 4×4, the step length is 2, the number of sampling layers is 3. In this paper, the number of discriminators is set to 2.

3.3 Loss function

The proposed loss function comprises four parts, namely GAN loss, L_1 loss, VGG loss and FM loss. Firstly, because multi-scale discriminator is used, the optimization problem of generative adversarial networks is expressed as follows:

$$\min_{G} \max_{D_1,D_2,D_3} \sum_{k=1,2,3} L_{\mathrm{GAN}}(G, D_k) \tag{5}$$

GAN loss is expressed as follows:

$$L_{\mathrm{GAN}}(G, D_k) = E_{(x,y)}[\log D_k(y)] + E_x[\log(1 - D_k(G(x)))] \tag{6}$$

where x is the input image, and y is the target image.

Also, restrictions should be added to evaluate the generated image. For generators and discriminators, the more important thing is the result of the generator, which will introduce L_1 loss:

$$L_1(G) = E_{(x,y)}\left[\|y - G(x)\|_1\right] \tag{7}$$

In order to make the output image closer to the real image, feature matching loss is introduced. Specifically, features are extracted from multiple layers of the discriminator and learned to match the intermediate features between the real image and the composite image. Assuming that the ith layer of the kth discriminator network is $D_k^{(i)}$, then the FM loss $L_{\mathrm{FM}}(G, D_k)$ can be expressed as:

$$L_{\mathrm{FM}}(G, D_k) = E_{(x,y)} \sum_{i=1}^{T} \frac{1}{N_i} \left[\left\| D_k^{(i)}(x, y) - D_k^{(i)}(x, G(x)) \right\|_1 \right] \tag{8}$$

where T is the total number of layers of the discriminator, and N_i is the number of elements in each layer. Similarly, for the difference between the two image features, VGG loss is introduced, and the features of the image are extracted through the pre-trained VGG network, defined as:

$$L_{\mathrm{VGG}}(G) = \sum_{i=1}^{N} \frac{1}{M_i} \left[\left\| F^{(i)}(y) - F^{(i)}(G(x)) \right\|_1 \right] \tag{9}$$

where $F^{(i)}$ represents the ith layer of the VGG network, and M_i represents the number of elements in this layer. Therefore, the ultimate optimization objective of the total loss function of the algorithm in this paper can be expressed as:

$$\min_{G} \left(\left(\max_{D_1,D_2,D_3} \sum_{k=1,2,3} L_{\mathrm{GAN}}(G, D_k) \right) + \lambda_1 L_1(G) + \lambda_2 \sum_{k=1,2,3} L_{\mathrm{FM}}(G, D_k) + \lambda_3 L_{VGG}(G) \right) \tag{10}$$

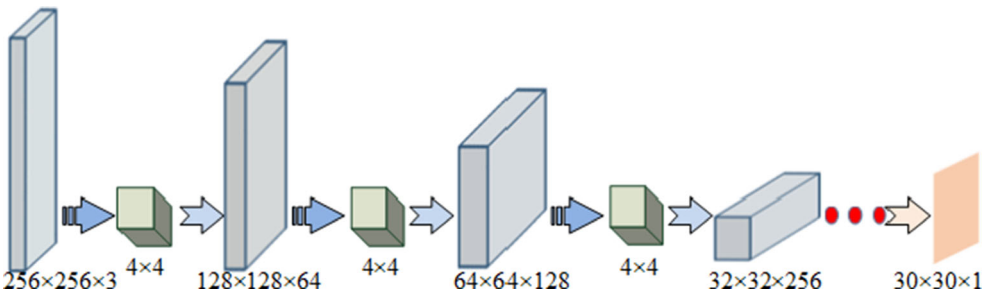

Fig. 5 Multi-scale discriminator

4 Experimental results and discussion

4.1 Experimental environment and datasets

The experiment includes four tasks, image hazing, satellite image to map conversion, optic image to SAR conversion and day to night conversion. Tests are implemented by PyTorch on a computer with CPU: Intel Core i7-5820K @ 3.30GHz × 12, GPU: NVIDIA GeForce TITAN X and 16G memory. The initial learning rate is 2e-4. The batch size is 1.

The image hazing dataset mainly comes from the RESIDE (Realistic Single Image Dehazing) dataset [38] which includes synthetic and real-world blurred images. In order to increase the diversity of dataset, we use the Adobe lightroom CC to haze clear images from the Middlebury Stereo Datasets [39] and the image from the internet.

The SAR image comes from the internet, a total of 1024 pairs of images.

The day and night conversion dataset comes from Pix2pix [3], a total of 1024 pairs of images.

The optic image to SAR conversion dataset comes from Google Map [3], a total of 1096 pairs of images.

4.2 Performance analysis

4.2.1 Generator structure analysis

The algorithm in this paper is inspired by the Pix2pixHD [19], and the generator structure is relatively similar. The main difference is that this paper introduces a hybrid attention module and an compromised instance and layer normalization method. Both Pix2pixHD and the algorithm in this article adopt a three-stage architecture of down-sampling, residual module and upsampling. Among them, the algorithm in this paper first adds a hybrid attention module to the middle of the network (also the middle of the residual module group). And the residual module group has been partially replaced, that is, the normalization method of the residual module is replaced with the network-guided instance and layer normalization approach, thereby realizing a new residual module. In order to better analyze the generator architecture, this section uses a visual feature map to explain in detail the various processes in the network operation. A total of three virtual nodes are selected, before and after the attention module (that is, the middle section of the residual module, which is also the dividing line between the ordinary residual structure and the residual structure with CILN), and the node before downsampling (the node after the residual module). We conducted related experiments using smog conversion as an example, and the results are as follows:

In Fig. 6, (a) is the clear image in the input network, which is also the source domain image, (b) is the feature map before the attention module, that is, the part of the network after down-sampling and classic residual structure, (c) is the feature map after the attention module, (d) is the feature map after all the residual structures, and also the input feature map of the down-sampling module, (e) is the output image of the down-sampling layer and the generated image of the entire network. From the figure, we can find that although the feature map before the attention module contains a certain amount of image information, the overall appearance is rather messy, and the focus is not clear. After the attention module proposed by the algorithm in this paper, the contrast of the feature map is significantly improved, the key areas are clearly marked, and the texture structure is prominent. The feature map before the down-sampling module basically contains all the image information in a relatively complete manner. The picture's visual analysis fully illustrates the critical role of the attention module, which can extract important information of the image, such as edge information, texture structure and color information. In general, the addition of the attention module and CILN enables the network to synthesize images with higher quality, rich details and clear texture. This also shows that the algorithm's improvement direction in this paper is correct and has specific research significance.

4.2.2 The number of multi-scale discriminators

In this paper, the multi-scale discriminator is used, that is, multiple discriminators are carried out under input images of different sizes. Because the decision is made at different scales, the small-scale image as input can pay more attention to the overall structure and edge of the image. In comparison the large-scale image as input can pay more attention to the detail retention of the image. In theory, the more discriminator the better, but that's not the case. The more discriminators there are, the more complex the network and the more computations there are. Second, the number of discriminators is related to the size of the input image itself. If the size of the input image itself is appropriate, it is not necessary to have too many discriminators if the size is not large or super large. Therefore, it is appropriate to choose the number of discriminators, and the following experiments are carried out under the specific circumstances studied in this paper. In this paper, training was conducted on the fogging training set. The image input size was 256 × 256, and the results of iterating 60 epochs were tested with the number of discriminators being 1, 2 and 3, respectively.

Overall, the number of discriminators has little effect on the content of the generated image. But there will be

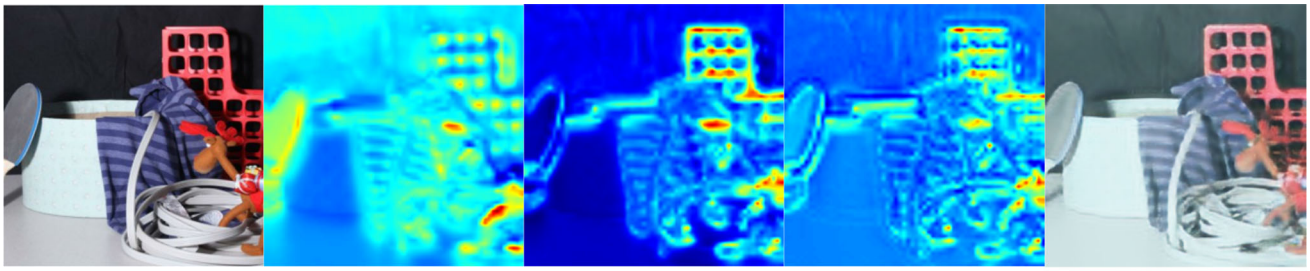

Fig. 6 Visual results of different stage of the generator

differences in detail. From the sky area in Fig. 7, since the sky area is brighter, it is still difficult to observe the difference after zooming in. Therefore, the sky area is processed, the RGB is converted to the YUV channel, and the Y channel is linearly pulled. Stretch to get the final result. The transformed image is shown in the enlarged area. When $\text{num}_D = 2$, the sky color is more uniform. Therefore, in the number of deciders, $\text{num}_D = 2$ for all experiments in this paper.

4.2.3 Settings for loss functions

The loss function of this paper consists of four parts: GAN loss, L_1 loss, FM loss and VGG loss.

This section compares the experimental results of using all loss functions (denoted as total loss), but not using VGG loss (denoted as no_VGGloss), but not using L_1 loss and FM loss (denoted as no_matchingloss). The experimental results in the three cases are shown in Fig. 8.

When the VGG loss is not used, image distortion occurs. For example, in Fig. 8, the red frame area, in the sky, the runway, and the like, an irregularly shaped white "foreign object" appears. This situation may occur due to data overflow, and without the limitation of VGG loss, image distortion will occur. When there is no loss of L_1 and FM, the image will not be distorted, but the color of the image will deviate. In the absence of L_1 and FM losses, the overall fogging is yellow. After using L_1 and FM loss, the color is standard. Therefore, using all loss functions has the best effect.

The three hyperparameters λ_1, λ_2 and λ_3 are the coefficients of the loss function that we supplement, and these three loss functions are used to measure the gap between the synthesized image and the real image, and the other is the GAN's counter loss. During training, we found that the value of the confrontation loss is in the single digit, while the value of other loss functions is often a few tenths. Therefore, we chose to set all three hyperparameters to 10. Similarly, we also refer to other algorithms, including Pix2pix, DRPAN and Pix2pixHD. The coefficient of their loss function except for the confrontation loss is generally 10. We take the haze conversion as an example to conduct an experimental analysis on the hyperparameters, and design $\lambda = 1, 10$ and 100 ($\lambda_1 = \lambda_2 = \lambda_3 = \lambda$), respectively. The experimental results are shown in Fig. 9.

We can see from the Fig. 9 that when λ is set to 1, 10, 100, the result of the network generation is similar. The image generated by $\lambda = 1$ is slightly blurred than $\lambda = 10$ and $\lambda = 100$. In view of the parameter settings of other algorithms and the order of different losses, $\lambda = 10$ is selected here.

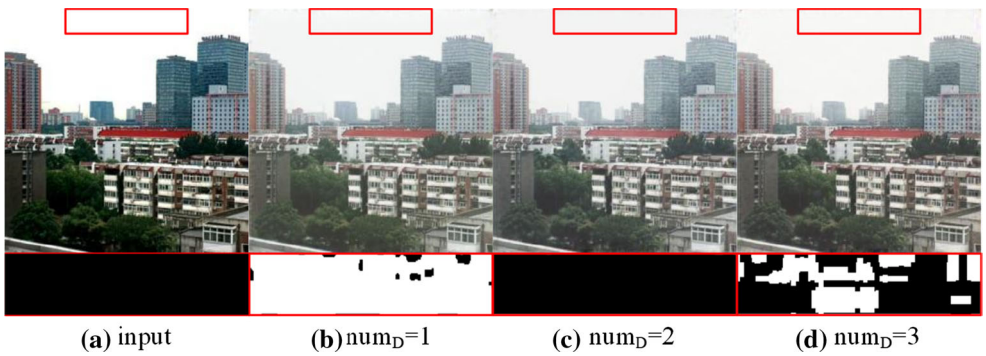

Fig. 7 Comparison of different number of discriminator

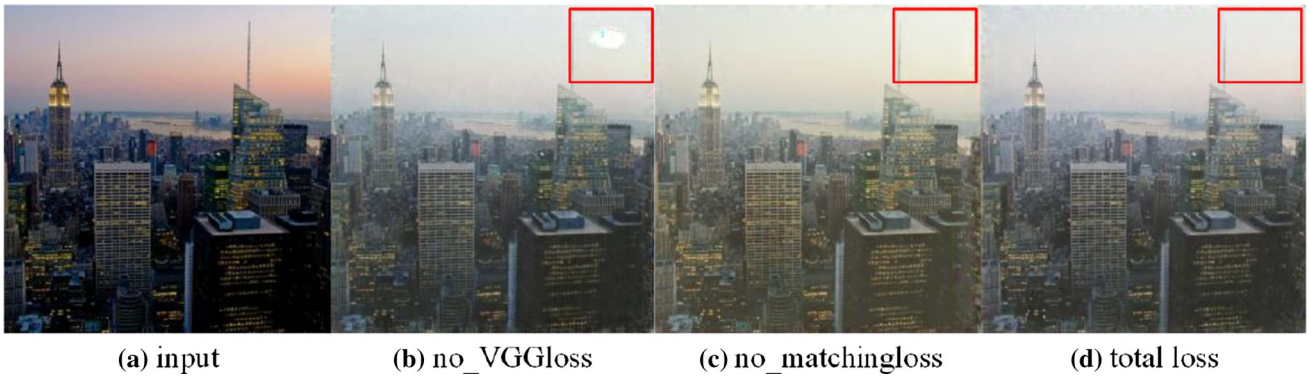

Fig. 8 Comparison of different loss function

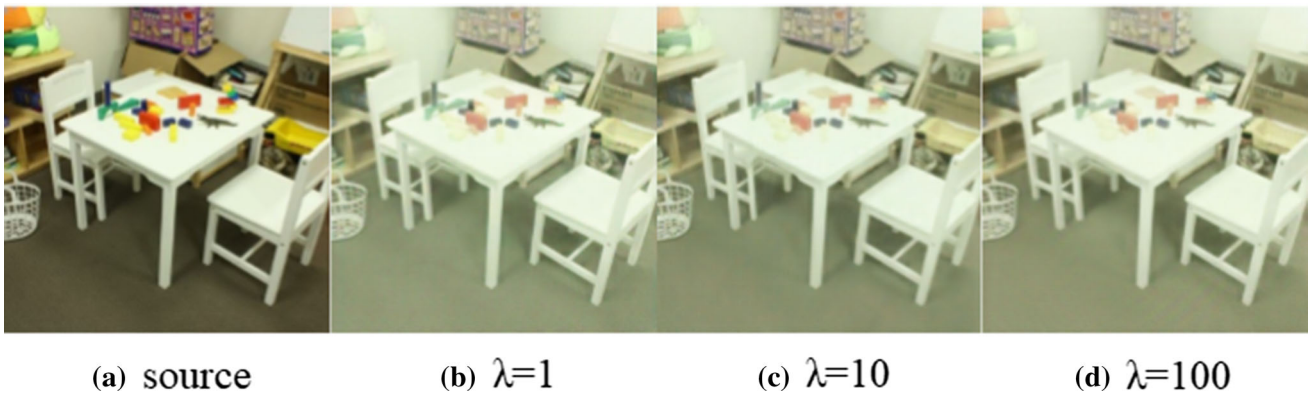

Fig. 9 Comparison of experimental results of different hyperparameters

4.3 Image hazing results analysis

4.3.1 Subjective results analysis

We use the five state-of-the-art algorithms: Pix2pix [3], Pix2pix-HD [19], CycleGAN [1], DRPAN [6], MUNIT [2] and Adobe lightroom CC to compare with the propose in the field of image hazing. The results are shown in Fig. 10.

As is shown in Fig 10, the hazing effect of Pix2pix is more realistic, which is very similar to the hazing effect of the Adobe lightroom CC, but the Pix2pix algorithm has obvious blocking effect in the blue ocean in the upper right corner of the image. The CycleGAN hazing image is very bad, and the content is blurry. Apart from the detail and blurring of the image, the color of the DRPAN does not resemble the atomization effect, and MUNIT has significant distortion in the hazy image. The details and content of the images hazed by the proposed will not be lost, nor will it produce blocking effects.

4.3.2 Objective index analysis

This paper uses three objective indicators to evaluate the performance of each algorithm, namely FADE (Fog Aware Density Evaluator) [40], PSNR (Peak Signal to Noise Ratio) and SSIM (structural similarity index measure). FADE is an index describing the density of image haze. PSNR is an index measuring the ratio of the energy of the peak signal to the average energy of the noise. SSIM is an indicator that measures the similarity of the two images. Table 1 shows the mean and mean square error of the three indicators in test charts, and compares the fog-free image, CycleGAN, Pix2pix-HD, Pix2pix, DRPAN, MUNIT and Adobe lightroom CC fogging effects.

Through the processing of Pix2pix, Pix2pix-HD, CycleGAN, DRPAN, MUNIT, Adobe lightroom CC and the proposed, the image's fog density index was significantly increased, with the lowest DRPAN degree and the highest CycleGAN degree. Pix2pix and Pixpix-HD are similar to the hazing degree of the proposed and Adobe lightroom CC. CycleGAN has a high score because the image has fewer colors, and the entire image is yellow-green. After image hazing, the PSNR and SSIM values of images basically keep a slight fluctuation within a certain range, while DRPAN, MUNIT and CycleGAN have a large fluctuation, which proves that their atomization effect is poor. In addition, the PSNR value of CycleGAN is basically the lowest among several hazing algorithms, which is

Fig. 10 Comparison of fogging results

also due to the errors and deficiencies of the image content generated by this algorithm. In contrast, the overall PSNR and SSIM values of the DRPAN are relatively high, due to the retention of good structural content, but the fogging effect is not apparent. The FADE of the proposed method has a great increase. The PSRN and SSIM of the PSRN and SSIM of the proposed method are relatively high, and the image produced by the proposed method is more in line with the real situation. In general, the proposed method performs well in all aspects.

4.4 Satellite image to map conversion results analysis

4.4.1 Subjective results analysis

At the same time, this paper tests the conversion of satellite image to map. The training set and test set adopt the public training set of Pix2pix, and test the conversion effect of this algorithm and other algorithms based on GAN. The comparison algorithms include Pix2pix, Pix2pix-HD, CycleGAN, DRPAN and MUNIT. The results are shown in Fig. 11.

In Fig. 11, five methods for satellite imagery of Street View can produce similar effects to Google Maps. We can see from the red box on the picture that the road synthesized by CycleGAN, DRPAN, and the proposed is straighter, while the road synthesized by Pix2pix algorithm is slightly curved. In terms of color, DRPAN composite pictures are too rich in color. In terms of content, the upper left corner of the picture synthesized by CycleGAN does not match the real picture, and it pays too much attention to the change in style. MUNIT produced the worst results, neither the color nor the content was successfully converted. In general, the proposed algorith considers the changes in content and style, and therefore performs well.

4.4.2 Objective index analysis

In the algorithms such as Pix2pix, observing and evaluating the image generated after image conversion is adopted. Therefore, this paper conducts a questionnaire survey on map image conversion and map conversion, compares various algorithms and scores the true degree of the image (the real night image is not given). The more realistic the generated image is, the higher the true and false is difficult to distinguish. The highest score of 5 points, the final scores of 21 questionnaires collected in the random population are as follows:

It can be seen from Table 2. Pix2pix, Pix2pix-HD and CycleGAN scores are similar, which make the image conversion remarkable. The road and grassland conversions are also accurate. Because the color is too rich and not similar to the target image, DRPAN has a relatively

Table 1 Comparison of objective index of fogging

	Fog-free image	CycleGAN	Pix2pix	Pix2pix-HD	DRPAN	MUNIT	Adobe lightroom CC fogging	Proposed
FADE	0.230 ± 0.116	**0.736 ± 0.431**	0.689 ± 0.348	0.651 ± 0.322	0.459 ± 0.250	0.518 ± 0.365	0.670 ± 0.412	0.646 ± 0.331
PSNR	19.231 ± 2.157	12.703 ± 1.708	13.830 ± 0.690	14.206 ± 1.103	**17.211 ± 2.541**	13.187 ± 1.541	13.611 ± 1.147	14.864 ± 0.814
SSIM	0.789 ± 0.052	0.349 ± 0.084	0.752 ± 0.074	0.726 ± 0.094	**0.809 ± 0.082**	0.517 ± 0.092	0.782 ± 0.075	0.738 ± 0.105

The best results are highlighted in boldface

low score. MUNIT gets the lowest score for its terrible translation. The segmentation of the proposed is complete, and the lines are smooth.

4.5 Optic image to SAR conversion results analysis

4.5.1 Subjective results analysis

At the same time, this paper tests the conversion of optic image to SAR. The training set and test set adopt the public training set of Pix2pix, and test the conversion effect of this algorithm and other algorithms based on generating against network. The comparison algorithms include Pix2pix,Pix2pix-HD, CycleGAN, DRPAN and MUNIT. The results are shown in Fig. 12.

See Fig. 12 for the details. For example, vertical stripes in the lower-left corner of the image, Pix2pix and DRPAN can generate SAR-like images as a whole, but vertical stripes cannot be recovered, and CycleGAN is just the opposite. DRPAN and the algorithm in this article can synthesize content well and maintain the SAR image style. What's more, the image content synthesized by the proposed is richer, and the texture is clearer.

4.5.2 Objective index analysis

The same questionnaire survey was used for map image conversion. The 21 recovery results were analyzed, the mean and the mean square error. The results are shown in Table 3:

It can be seen from the results that the CycleGAN score is only slightly lower than the Pix2pix, because the CycleGAN can retain the sharper and more accurate image edges, and the image as a whole also shows gray-white color, but for The feature conversion of SAR images is not accurate enough, the algorithm score is slightly higher, because the effect of the algorithm is outstanding regardless of color or edge, and there is no block effect.

4.6 Day to night conversion results analysis

4.6.1 Subjective results analysis

This paper trains and tests the scene transition from day to night. The results are shown in Fig. 13.

As can be seen in Fig. 13, the content of the proposed is basically unchanged. But for the sky area, it turns black, and the Eiffel Tower lights up. The night visions of images produced by algorithms other than MUNIT are closer to real night images, i.e., similarities in color and brightness. But Pix2pix and DRPAN have blocks, especially at the junction of the sky and the ground. The image converted

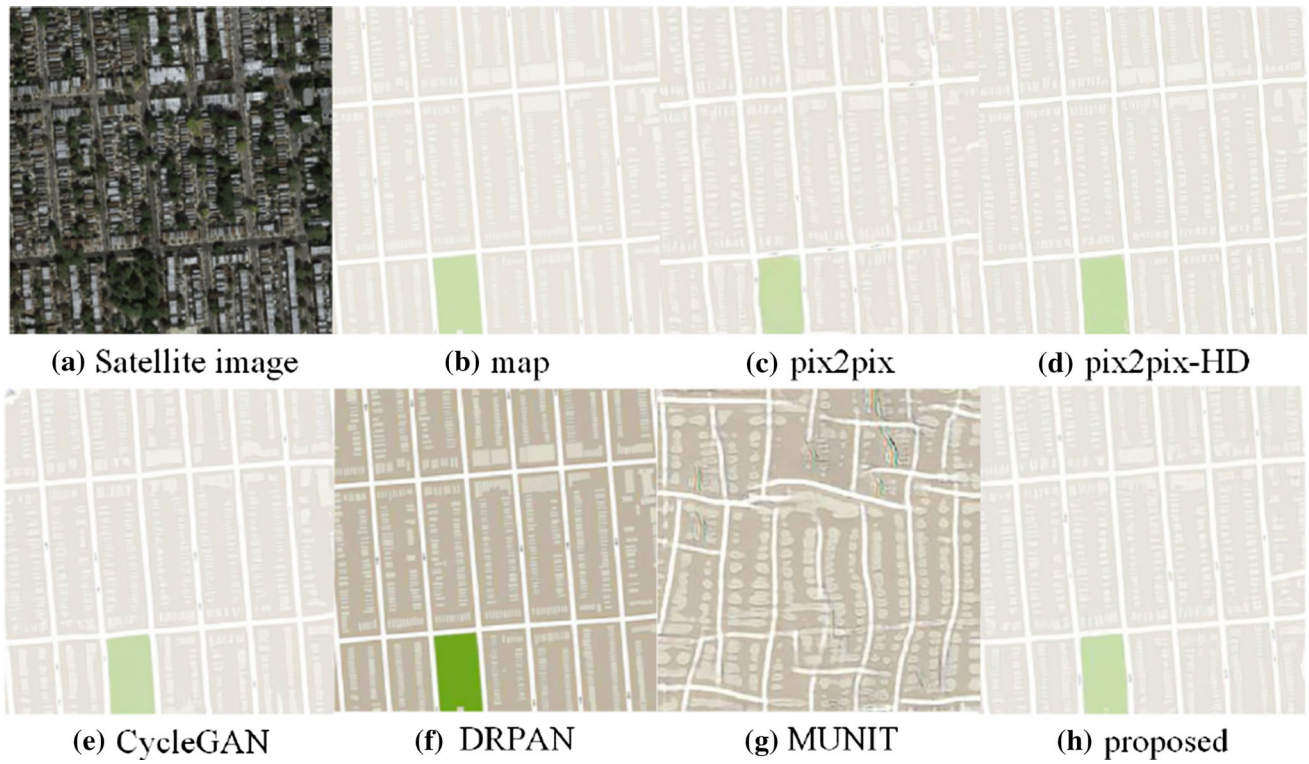

Fig. 11 Map scene synthesis results comparison

Table 2 Map image conversion score comparison

Algorithm	Pix2pix	Pix2pix-HD	CycleGAN	DRPAN	MUNIT	Proposed
Average scorce	3.65	3.64	3.52	2.39	1.13	4.04

by MUNIT is seriously distorted directly below the image. Although the image synthesized by the algorithm in this article still has some distortion directly below, however, compared with other algorithms, the proposed reduces the block effect and is more realistic.

4.6.2 Objective index analysis

The same questionnaire survey was used for SAR and map image conversion. The 21 recovery results were analyzed, the mean and the mean square error. The results are shown in Table 4.

It can be seen from Table 4. Pix2pix, Pix2pix-HD, DRPAN and the proposed method have similar scores, and the fluctuation range is large. The night image conversion effect is remarkable. Pix2pix and DRPAN do better in color and saturation. The algorithm in this paper has less image distortion. CycleGAN and MUNIT perform poorly on this conversion task, the color is dim, and the texture structure of the image is mostly lost.

4.7 Ablation study

The main purpose of this section is to analyze whether the two added modules (attention module and compromised normalization module) have improved the network based on the original network architecture. The following are introduced from the subjective results and objective indicators. Fig. 14. is a comparison of the subjective results of the four conversion tasks. The first line is Haze-free to Haze, the second line is Satellite to Map, the third line is Day to Night, and the fourth line is Optical to SAR image. For each specific image, (a) is the original image, (b) is the converted image of the original network structure (without attention and compromised normalization module), (c) is the original network The architecture adds the attention module (the compromised normalization module is not used) after conversion, (d) is the original network architecture with attention and compromised normalization module after conversion, and (e) is the target image, that is, the fitted image in the conversion process.

In generative adversarial networks, the discriminator and the generator's objective function are usually used to

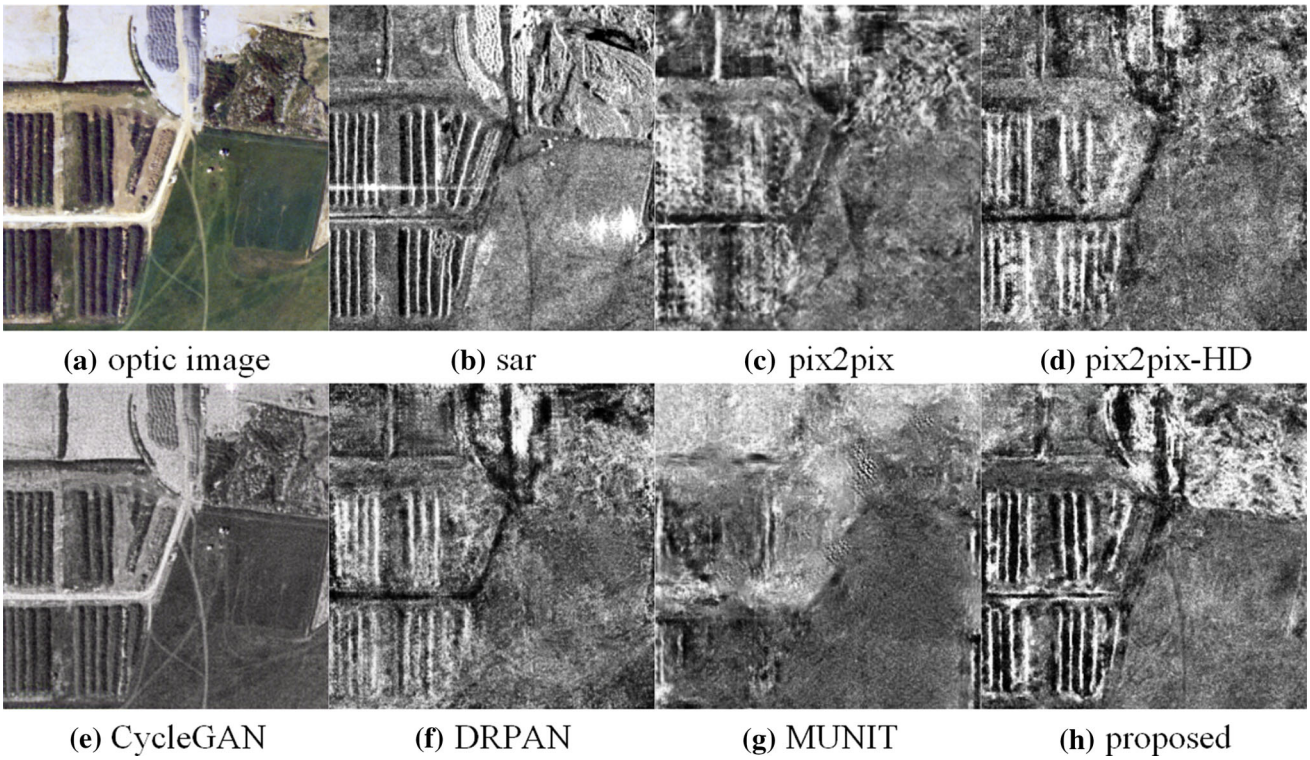

Fig. 12 SAR scene synthesis results comparison

Table 3 SAR image conversion score comparison

Algorithm	Pix2pix	Pix2pix-HD	CycleGAN	DRPAN	MUNIT	Proposed
Average score	2.96	3.85	2.80	3.18	1.26	4.39

measure how they each do. For example, the objective function of the generator is used to measure the performance of the generated images can fool the discriminator. But this is not a good measure of the quality and diversity of the generated images. Usually, three indicators, IS (inception score), FID (Frechet Inception Distance) and KID [41] (kernel Inception Distance) are used to evaluate different GAN models. They can only describe the quality of the generated results to a certain extent. This paper uses the IS, FID and KID to evaluate the role of the attention module and the compromised normalization module.

As can be seen from Table 5, after adding attention and compromised normalization modules, the indicators of other conversion tasks except for map conversion have been significantly improved, which can prove that these two modules are effective and have significant effects. However, in the map conversion task, the IS indicator still declined. After analysis, we believe that this is due to the particularity of map conversion. That is, map conversion converts the real scene into a virtual map, and the roads and most buildings of the virtual map are used. Instead of white pixels, the better the conversion effect, the more the pixel values in the area are closer, and the image is smoother. However, the IS indicator is more inclined to give high scores to colorful images, which may be why the results of improved network conversion in the map conversion task are lower.

As can be seen from Table 6, after adding attention and compromised normalization modules, the indicators of conversion tasks other than map conversion have been significantly reduced, which can prove that these two modules are useful, indicating that the improvement of the network The converted result is closer to the target image, and the fitting effect is better.

As can be seen from Table 7, The performance of the KID value is similar to that of FID. In haze conversion, night conversion and SAR conversion, the addition of HA and CILN modules makes the network more excellent. In the map conversion, the HA module was added separately to obtain the best results. We think this is related to the converted image mode. In map conversion, as shown in Fig 14, the conversion result is often mainly white and gray pixel. On the whole, HA and CILN have a certain effect on the improvement of the network.

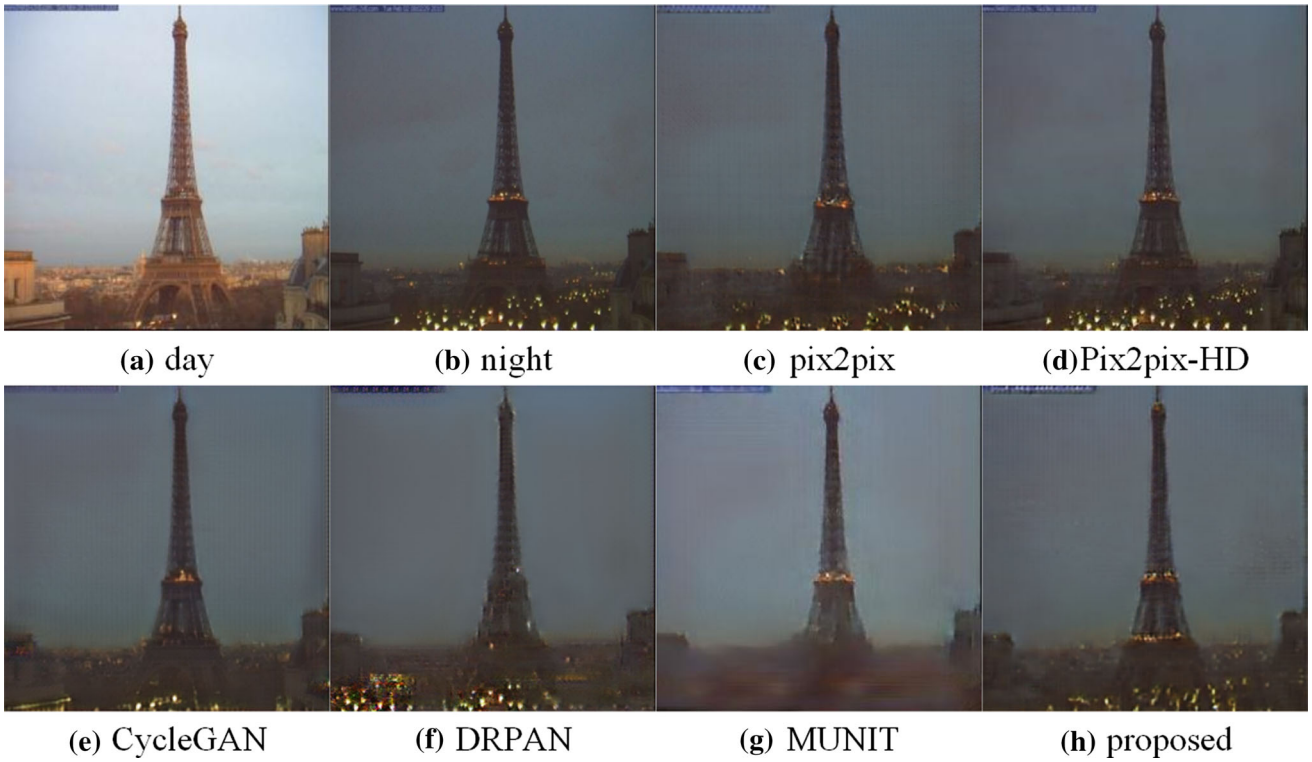

Fig. 13 Night scene synthesis results comparison

Table 4 Night image conversion score comparison

Algorithm	Pix2pix	Pix2pix-HD	CycleGAN	DRPAN	MUNIT	Proposed
Average score	3.57	3.85	3.36	3.68	2.14	3.96

4.8 The cost of implementing the proposed method

We conduct experiments under the same equipment, comparing time consumption of different algorithms. The model size and FPS (Frames Per Second) of different algorithms are shown as Table 8.

It can be seen from Table 8, the models of Pix2pix-HD and MUNIT are relatively large, and the speed is also relatively slow. The models of Pix2pix and DRPAN are relatively lightweight, and the speed is alse relatively fast. In combination, the model size and speed of the proposed have a certain advantage.

5 Conclusions

In this paper, we propose a generative adversarial network with hybrid attention and compromised normalization for multi-scene image conversion. Our generative adversarial network introduces a hybrid attention mechanism as well as a compromised instance and layer normalization method. With a fixed network architecture and hyperparameters, our method produces better results on different datasets. A detailed analysis of experimental results and ablation studies on different datasets supports our conclusion that the attentional mechanism can effectively control and enhance image transformation. The compromised instance and layer normalization approach can also help the accuracy of image transformation in different scenarios. However, there are still some limitations to our work, one being that the resolution of the images is not high, but rather the degree of image conversion is uncontrollable. Therefore, the direction of our future work is to improve the resolution of network convertible images and to control the degree of image conversion. Results reported in this paper can be downloaded from the url https://github.com/xiaojs18/Image-GAN/tree/master/AttGAN.

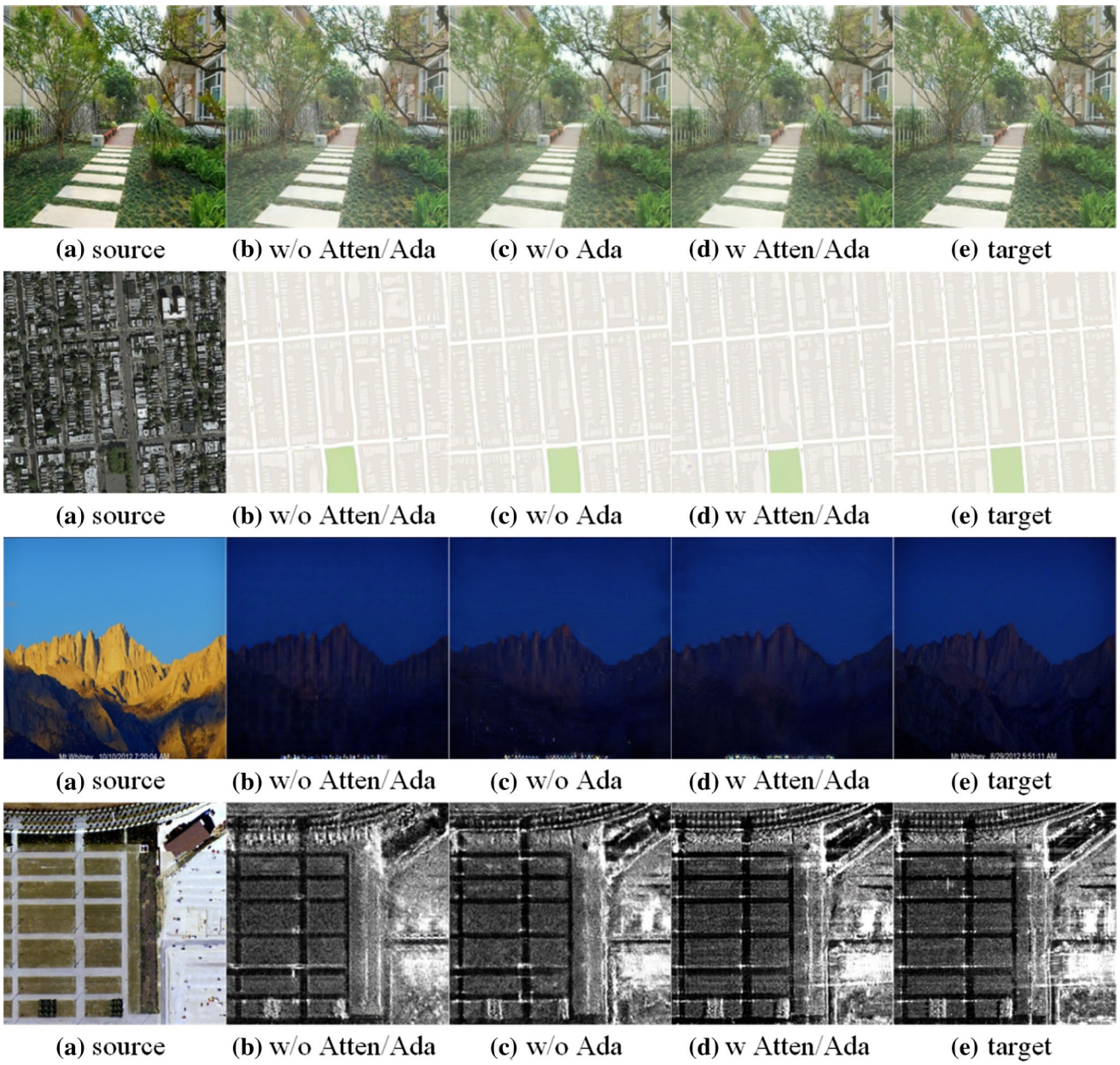

Fig. 14 Ablation study of the generator

Table 5 IS comparison of different tasks (The larger the IS, the better the image quality)

	Haze	Map	Night	SAR
w/o Atten/CILN	2.74 ± 0.31	2.63 ± 0.18	2.83 ± 0.12	2.18 ± 0.37
w/o CILN	2.75 ± 0.36	**2.64 ± 0.18**	2.61 ± 0.15	2.34 ± 0.27
w Atten/CILN	**2.81 ± 0.38**	2.55 ± 0.17	**2.98 ± 0.29**	**2.44 ± 0.28**

The best results are highlighted in boldface

Table 6 FID comparison of different tasks (The smaller the FID, the better the image quality)

	Haze	Map	Night	SAR
w/o Atten/CILN	49.101	137.932	153.740	193.435
w/o CILN	46.890	**93.104**	168.145	155.257
w Atten/CILN	**41.208**	108.539	**145.711**	**117.231**

The best results are highlighted in boldface

Table 7 KID $\times$ 100 $\pm$ std. $\times$ 100 comparison of different tasks (The smaller the KID, the better the image quality)

	Haze	Map	Night	SAR
w/o Atten/CILN	2.29 $\pm$ 0.41	11.08 $\pm$ 0.72	6.62 $\pm$ 0.42	3.89 $\pm$ 0.25
w/o CILN	1.92 $\pm$ 0.38	**6.46 ± 0.39**	7.46 $\pm$ 0.43	2.20 $\pm$ 0.43
w Atten/CILN	**1.82 ± 0.39**	7.63 $\pm$ 0.56	**4.83 ± 0.38**	**2.07 ± 0.55**

The best results are highlighted in boldface

Table 8 The cost of different algorithms

Method	Model size	FPS
Pix2pix	52.7M	12.5
Pix2pix-HD	751.9M	2.7
CycleGAN	113.2M	4.1
DRPAN	101.3M	21.0
MUNIT	559.3M	5.9
Proposed	201.3M	11.8

Acknowledgements This work is funded by the Social Science Foundation of Shaanxi Province (Grant No. 2019H010), the New Star of Youth Science and Technology of Shaanxi Province (Grant No. 2020KJXX-007) and the Open Project Program Foundation of the Key Laboratory of Opto-Electronics Information Processing, Chinese Academy of Sciences (OEIP-O-202009). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Declarations

Conflict of interest We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232
2. Huang X, Liu M-Y, Belongie S, Kautz J (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 172–189
3. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
4. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
5. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
6. Wang C, Zheng H, Yu Z, Zheng Z, Gu Z, Zheng B (2018) Discriminative region proposal adversarial networks for high-quality image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 770–785
7. Wang W, Li Z (2018) Advances in generative adversarial network. Tongxin Xuebao/J Commun 39(2):133–146
8. Xiao J, Zhou J, Lei J, Xu C, Sui H (2020) Image hazing algorithm based on generative adversarial networks. IEEE Access 8:15883–15894
9. Xun Huang SB (2016) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision, pp 1501–1510
10. Goodfellow IJ, Jean P-A, Mehdi M, Bing X, Yoshua B (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
11. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: Proceedings of the 34th international conference on machine learning, pp 214–223
12. Iizuka S, Simo-Serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graph (TOG) 36(4):1–14
13. Bhattacharjee D, Kim S,Vizier G, Salzmann M (2020) Dunit: detection-based unsupervised image-to-image translation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)
14. Chen Y-C, Xu X, Jia J (2020) Domain adaptive image-to-image translation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)
15. Jiang S, Chen Y, Yang J, Zhang C, Zhao T (2019) Mixture variational autoencoders. Lecture Notes in Computer Science 128:263–269
16. Larsen ABL, Boesen, Sønderby SK, Larochelle H, Winther O (2016) Autoencoding beyond pixels using a learned similarity metric. In: 33rd international conference on machine learning, pp 2341–2349
17. Xu W, Shawn K, Wang G (2019) Toward learning a unified many-to-many mapping for diverse image translation. Pattern Recognit 93:570–580
18. Li W, Fan L, Wang Z, Ma C, Cui X (2021) Tackling mode collapse in multi-generator GANs with orthogonal vectors. Pattern Recognit 110:107646
19. Wang T-C, Liu M-Y, Zhu J-Y, Tao A, Kautz J, Catanzaro B (2018) High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE

conference on computer vision and pattern recognition, pp 8798–8807
20. Mao X, Li Q, Xie H, Lau R, Zhen W, Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2813–2821
21. Zhang Y, Xiao J, Peng J, Ding Y, Liu J, Guo Z, Zong X (2018) Kernel wiener filtering model with low-rank approximation for image denoising. Inf Sci 462:402–416
22. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2414–2423
23. Li R, Cao W, Jiao Q, Wu S, Wong H-S (2020) Simplified unsupervised image translation for semantic segmentation adaptation. Pattern Recognit 105:107343
24. Li Y, Fang C, Yang J, Wang Z, Lu X, Yang M-H (2017) Diversified texture synthesis with feed-forward networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3920–3928
25. Chen D, Yuan L, Liao J, Yu N, Hua G (2017) Stylebank: an explicit representation for neural image style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1897–1906
26. Liu M-Y, Tuzel O (2016) Coupled generative adversarial networks. In: Advances in neural information processing systems, pp 469–477
27. Luan F, Paris S, Shechtman E, Bala K (2017) Deep photo style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4990–4998
28. Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D (2017) Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3722–3731
29. Chen Q, Koltun V (2017) Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE international conference on computer vision, pp 1511–1520
30. Dosovitskiy A, Brox T (2016) Generating images with perceptual similarity metrics based on deep networks. In: Advances in neural information processing systems, pp 658–666
31. Dai T, Cai J, Zhang Y, Xia S-T, Zhang L (2019) Second-order attention network for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 11057–11066
32. Zhang Y, Tian Y, Kong Y,Zhong B, Fu Y (2018)Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2472–2481
33. Zhang H, Goodfellow I, Metaxas DN, Odena A. Self-attention generative adversarial networks. Machine Learning. arXiv:1805.08318
34. Miyato T, Koyama M (2018) cGANs with projection discriminator. In: Proceedings of the international conference on learning representations
35. Park T, Liu M-Y, Wang T-C, Zhu J-Y (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2337–2346
36. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023
37. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. Lecture Notes in Computer Science 11211:3–19
38. Li B, Ren W et al (2018) Benchmarking single-image dehazing and beyond. IEEE Trans Image Process 28(1):492–504
39. Scharstein D, Hirschmüller H, Kitajima Y, Krathwohl G, Nešić N, Wang X, Westling P (2014) High-resolution stereo datasets with subpixel-accurate ground truth. In: German conference on pattern recognition, pp 31–42
40. Choi LK, You J, Bovik AC (2015) Referenceless prediction of perceptual fog density and perceptual image defogging. IEEE Trans Image Process 24(11):3888–3901
41. Bi'nkowski M, Sutherland DJ, Arbel M, Gretton A (2018) Demystifying MMD GANs. In: International conference on learning representations