Computers, Materials & Continua
DOI:10.32604/cmc.2021.019305
Article

Tech Science Press

Generating Cartoon Images from Face Photos with Cycle-Consistent Adversarial Networks

Tao Zhang[1,2], Zhanjie Zhang[1,2,*], Wenjing Jia[3], Xiangjian He[3] and Jie Yang[4]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214000, China
[2]Key Laboratory of Artificial Intelligence, Jiangsu, 214000, China
[3]The Global Big Data Technologies Centre, University of Technology Sydney, Ultimo, NSW, 2007, Australia
[4]The Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, 201100, China
[*]Corresponding Author: Zhanjie Zhang. Email: zhangzhanj@126.com
Received: 08 April 2021; Accepted: 09 May 2021

Abstract: The generative adversarial network (GAN) is first proposed in 2014, and this kind of network model is machine learning systems that can learn to measure a given distribution of data, one of the most important applications is style transfer. Style transfer is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image. CYCLE-GAN is a classic GAN model, which has a wide range of scenarios in style transfer. Considering its unsupervised learning characteristics, the mapping is easy to be learned between an input image and an output image. However, it is difficult for CYCLE-GAN to converge and generate high-quality images. In order to solve this problem, spectral normalization is introduced into each convolutional kernel of the discriminator. Every convolutional kernel reaches Lipschitz stability constraint with adding spectral normalization and the value of the convolutional kernel is limited to [0, 1], which promotes the training process of the proposed model. Besides, we use pretrained model (VGG16) to control the loss of image content in the position of l1 regularization. To avoid overfitting, l1 regularization term and l2 regularization term are both used in the object loss function. In terms of Frechet Inception Distance (FID) score evaluation, our proposed model achieves outstanding performance and preserves more discriminative features. Experimental results show that the proposed model converges faster and achieves better FID scores than the state of the art.

Keywords: Generative adversarial network; spectral normalization; Lipschitz stability constraint; VGG16; l1 regularization term; l2 regularization term; Frechet inception distance

1 Introduction

Goodfellow et al. [1,2] proposed a new neural network model in 2014, and named it as generative adversarial network (GAN). Nowadays, the GAN develops rapidly and promotes the development of the whole neural network. The GAN is composed of two parts: one is a generator

and the other is a discriminator. Although the generator and discriminator are two separate parts, they need to be trained at the same time. Then, the generator generates false data to deceive the discriminator and the discriminator identifies the generated data. At present, the GAN is widely used due to its unsupervised learning characteristics [3], it's worth noting that the GAN is prone to collapse during model convergence. Arjovsky et al. [4–7] tried to solve this problem by adjusting the object loss function. They performed rigorous mathematical derivation to find out why the model was prone to collapse and introduced Wasserstein distance. With the development of GAN, the convolutional neural network is firstly used in DCGAN [8]. After that, GAN is gradually applied in the field of style transfer [9–11]. Style transfer means that we can transfer the style of one image to another image. Since 2016, many style transfer methods have been proposed, such as PIX2PIX [12], CO-GAN [13], CYCLE-GAN [14], STAR-GAN [15,16], CARTOON-GAN [17], DISCO-GAN [18], DUAL-GAN [19], etc.

Style transfer has been widely applied in diverse scenes [20–26]. Up to now, one of the most important is to solve the problem of high resolution [27] and multi-target style transfer. To achieve this goal, some researchers propose new methods. Besides, it is difficult for GAN to achieve Nash equilibrium. In the process of style transfer, the model may collapse. The model collapse means that the model cannot generate a better result and the object loss function value doesn't keep descending. For the style transfer of face images, the distorted distribution of facial features often occurs. Thus, it is necessary to add key point constraints for an effective style transfer.

For style transfer, the main methods are based on the GAN. On this basis, adjusting the architecture [28] and reconstructing the object loss function is necessary. Some researchers try to use the classical GAN to replace facial features, expressions [29], clothing [30], etc. And other researchers even break the limitations of one generator to achieve style transfer between multiple fields, such as STAR-GAN. However, the extra hardware resources and time resources are wasted to generate high-quality images in style transfer. Our proposed model achieves style transfer based on unpaired images and uses lightweight neural networks to generate better results. In our proposed model, the discriminator relies on the method of embedded normalization [31–38], and it reduces the oscillation of the object loss function during model convergence. In order to achieve the spectral normalization of each convolutional kernel, it is necessary to obtain the singular value of the weight matrix, and the iteration method is used to get it. After obtaining the spectral normalization, each parameter matrix of the discriminator is subject to Lipschitz constraint [39] and each change of the parameter matrix is limited to a certain range.

At present, many problems remain to be solved in style transfer. Many researchers try to solve these problems by optimizing neural network structure. Besides, some measures are proposed, such as constructing loss function term, adding normalization and implementing attention mechanism [40,41]. In this paper, a novel discriminator of the generative adversarial network is proposed. At the same time, by extracting the high-dimensional features of the generated images, the pretrained model (VGG16 [42]) is used to reduce the loss of image content. In the object loss function, l1 regularization item and l2 regularization item [43,44] are both used to avoid overfitting.

In the process of style transfer, feature extraction and feature reconstruction are very important when treating face images with high dimensional features. Besides, it is beneficial for GAN to learn the style when the cartoon image has clear outlines. In order to obtain better experimental results, there is no fixed learning rate algorithm and object loss function in style transfer. Also, multiple attempts to construct different neural network structures are necessary. It's worth noting that the different methods may conflict with each other and make the object loss function not

converge. In this paper, we accelerated style transfer by training an end-to-end neural network with a lightweight structure.

The remaining of this paper is organized as follows. In Section 2, we introduce our related work. We introduce our proposed GAN model in Section 3. The experimental result of style transfer is presented in Section 4. At last, the conclusions are summarized in Section 5.

2 Related Works

Traditional non-parametric image style transfer methods are mainly based on the physical model by image rendering and texture synthesis. The non-parametric image style transfer method can only extract the low-level features of the image. When processing the images with complex colors and textures, the final image synthesis result is relatively rough, therefore, it doesn't meet the actual needs.

In deep learning, image style transfer methods mainly include image iteration method and model iteration method. The image iteration method presents many advantages, such as high-quality composite image, good controllability, convenient parameter adjustment and no training data. However, the additional computing resources are consumed. More specifically, image iteration methods can be divided into maximum mean difference methods [45–47], Markov Random Field methods [48] and depth image analogy methods [49]. At present, the model iteration method is the mainstream technology of industrial application software, which can be used for fast video programming. However, the quality of image generation needs to be further improved, and lots of images are needed for training the model. The model iteration method can be divided into generative model method [50–52] and image reconstruction decoder method [53].

Many generative model methods were proposed, such as CYCLE-GAN, DISCO-GAN and DUAL-GAN. CYCLE-GAN is based on the cycle consistency method, while DISCO-GAN and DUAL-GAN are based on machine translation. These excellent models break the limit of paired training data and successfully realize unsupervised transfer learning. At present, GAN is quite unstable in model convergence, and discriminator makes it difficult to implement the style transfer in clear direction. In addition, GAN is an iterative optimization based on the distribution of image divergence, rather than based on the content, texture and color of the image, so it is difficult to control the process of style transfer. In order to solve the model instability and improve the quality of the generated images, the spectral normalization and pretrained model (VGG16) are introduced in proposed model.

In this paper, our proposed model is different from the traditional GAN model. Instead of using one generator and one discriminator, two generators and two discriminators are both used. Besides, the pretrained model (VGG16) is also added to reduce the loss of image content. The proposed model is based on CYCLE-GAN and we make some improvements. In the generator, U-NET method is used to extract features and reconstruct features. To get more discriminative facial features of generated images, we use a pretrained model (VGG16) to extract high dimensional features from generated images and face images. Based on that, the generated image could preserve the image content well. In our constructed discriminator, convolutional neural networks are used instead of fully connected neural networks. The proposed discriminator is designed according to human visual characteristics. In object loss function, the loss function of image content is introduced in this paper.

The open-source dataset is collected including face images and cartoon images. All the images are divided into two parts: training set and test set. To reduce computation resources, all the

images are normalized when the proposed model loads the image dataset. In this paper, the proposed model is constructed to transfer style between face images and cartoon images. And then the loss function value of the proposed model is recorded, as well as the generated images. On the basis of the original GAN model, the spectral normalization, l2 regularization item and pretrained model (VGG16) are introduced. The object loss function is recorded for comparison. After training the proposed model, the size of the image is resized to the original image. During the convergence process of the proposed model, the images in the training set are randomly selected. The learning rate changes at each epoch.

The open-source cartoon dataset is selected which has simple features and small image sizes. For datasets with more complex features, it is better to construct deeper neural networks [54–56]. We use TensorFlow to save the checkpoints during the convergence process, and use the built-in TensorBoard to plot the loss function graph. The generator adopts the classical architecture of the convolutional network, residual network and deconvolutional network respectively. In order to stabilize the training process of GAN and prevent image distortion, it's effective to adjust the angle of face images and cartoon images. In addition, in order to get better training results, adjusting the learning rate is effective. The designed generator firstly extracts the image features through a three-layer convolutional neural network, then learns the image features through a nine-layer residual network, and finally reconstructs the image features through a three-layer deconvolutional network. We make some improvements on the traditional GAN and propose a new discriminator. However, due to the many unstable factors in the convergence process of the GAN, the model often breaks down. The main reason is the mismatch between the abilities of the generator and the discriminator. Besides, it's possible that the neural network is not deep enough to learn the complex features of the image dataset. So, many attempts to adjust the model structure and manually adjust the parameters are very necessary. Thus, we worked hard to reconstruct the architecture of GAN.

Generally, the main contributions of this paper are summarized as follows:

- The key point is to obtain the parameter matrix of the convolutional neural network and calculate the singular value. Extra computing resources are consumed during singular value calculation. Thus, the power iteration method is firstly used to approximate singular values. The singular value is used to conduct spectral normalization. After adding spectral normalization, the parameter matrix of the convolutional neural network meets the Lipschitz constraint. It replaces the parts of the parameter matrix is greater than 1 set for less than 1.
- The proposed discriminator is designed according to human visual characteristics. It has six layers of convolutional neural networks. Each convolution kernel of the convolutional neural network is designed to satisfy the Lipschitz stability constraint, and the value of it is limited to [0, 1].
- We use pretrained model (VGG16) to extract the high dimensional features of the face images and the generated images. In this way, the generator can learn the style features from the style image and preserve the content features of the face images well. Then, high dimensional features are extracted from the face images and generated images. Besides, l1 loss is used. In the object loss function, l1 regularization term and l2 regularization term are used to avoid overfitting.
- We collected more images of faces and cartoons. When loading the total length of the image list and the image index of the dataset, the image is normalized to reduce computation. After style transfer, the generated image is restored to a 128×128 pixel image.

3 Proposed Model

3.1 Basic Model

The model of the GAN can be divided into two parts: generator and discriminator. The generator and discriminator constantly learn from each other when the model converges. In the original GAN theory, it is unnecessary to define generator and discriminator as a neural network. At present, the deep neural network is generally used as a generator and discriminator. In order to obtain better generated images, the training sequence of the generator and discriminator needs to be adjusted asynchronously during the training process. The best training strategy is to ensure that the loss function values of the generator and discriminator are close to each other. Among them, the generator is trained to generate realistic images to deceive the discriminator, and the discriminator is trained to not be deceived by the generated image. In the first epoch, the generated image looks very messy. Subsequently, the discriminator receives false and real images and learns to distinguish between them. The generator receives "feedback" from the discriminator through a backpropagation step to generate better images. In this paper, the face image undergoes feature extraction, feature mapping and feature reconstruction respectively. Through constant learning of the generator, the false image is generated which is similar to the real face. The discriminator is used to identify the generated image. The discriminator continuously improves its discriminating ability and furtherly guides the generator. And the generator generates false images to deceive the discriminator continuously. Finally, the discriminator cannot distinguish whether the generated image is true or false. After the generator and the discriminator learning from each other over a period of time, the training process of GAN is finished. The architecture of the GAN model is shown in Fig. 1.

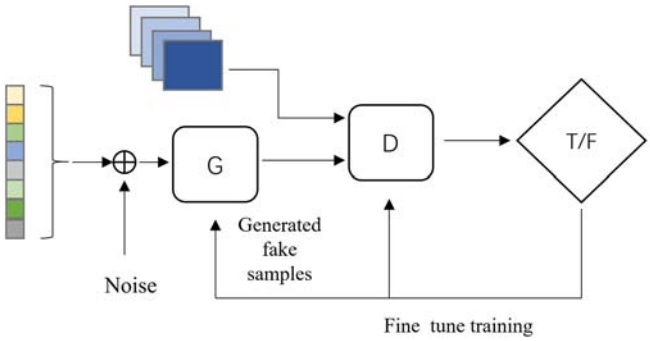


Figure 1: The model of generative adversarial network (I. J. Goodfellow)

Different from the traditional neural network model, the loss function of GAN can be divided into two parts: generator loss function and discriminator loss function. The loss function is described as Eqs. (1)–(3). G represents generator, D represents discriminator, $P_z(z)$ represents false data distribution, $P_{data}(x)$ represents true data distribution

Generator loss function:

$$\min_G V(D,G) = E_{z\sim P_z(z)}[\log(1-(D(z)))] \tag{1}$$

Discriminator loss function:

$$\max_D V(D,G) = E_{x\sim P_{data}(x)}[\log(D(x))] + E_{z\sim P_z(z)}[\log(1-(D(z)))] \tag{2}$$

Object loss function:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim P_{data}(x)}[\log(D(x))] + E_{z\sim P_z(z)}[\log(1-(D(z)))] \tag{3}$$

As shown in the Eqs. (1) and (2), the loss functions of generator and discriminator are actually treated as two separate parts. Different object loss functions are required for different models. Based on the BP [57] approach, the parameter matrix of the convolutional neural network is continuously optimized to minimize the object loss function. For GAN, there are no specific restrictions on the architecture of generators and discriminators. It only represents a network model, and any neural network can be used to implement the GAN.

3.2 Improved Model

The traditional GAN only allows the generator to generate data to deceive the discriminator and the generated data contains similar features to the original data. However, traditional GAN cannot achieve style transfer from one domain to another. On the basis of CYCLE-GAN, the content consistency loss function is proposed with the help of pretrained model (VGG16). In this paper, the pretrained model (VGG16) is used to extract the high dimensional features of the generated image and the face image. In this way, the content of the generated image is determined. The proposed model needn't be trained by paired images. The proposed model is easy to train and fast to converge. The proposed model has a very wide range of applications. Its purpose is to form a general mapping from image domain X to image domain Y. Overview of the proposed model is shown as Fig. 2.

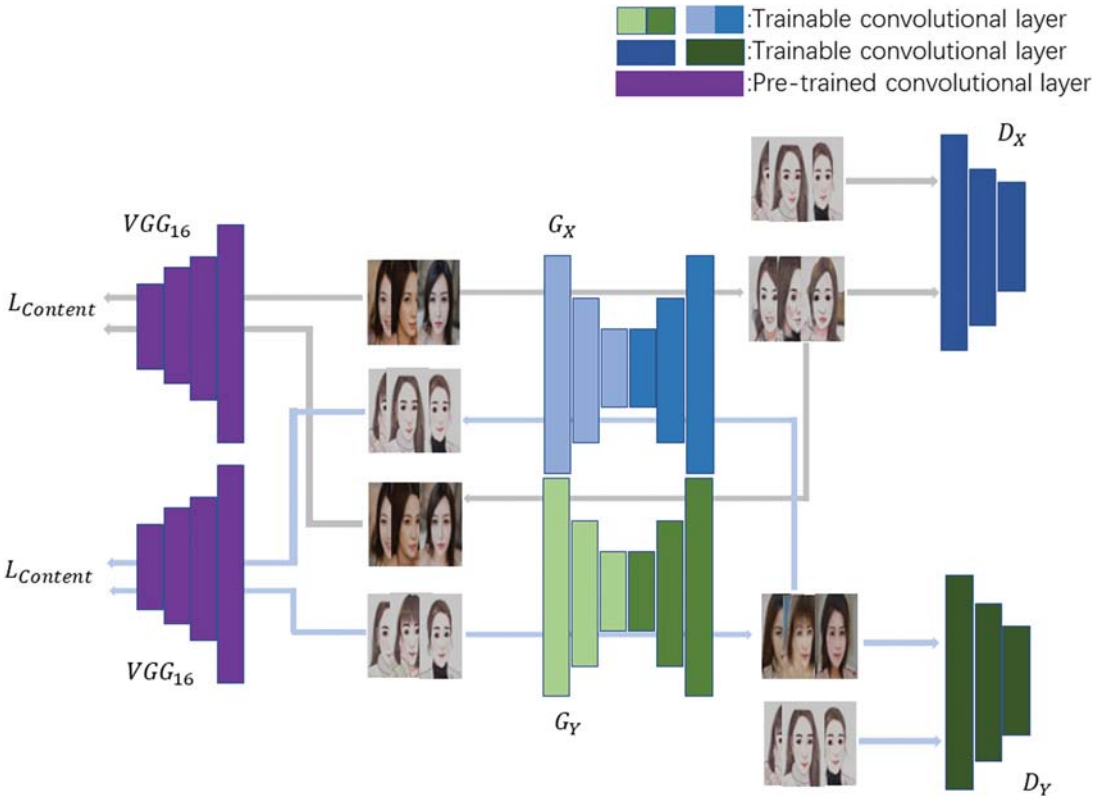


Figure 2: An overview of the proposed model

The proposed model consists of two discriminators and two generators, which control image content loss by the pretrain model. The loss function is shown as Eqs. (4)–(7). L represents the loss function, X represents the image in domain X and Y represents the image in domain Y.

For generator $G_X : X \rightarrow Y$ and its discriminator D_Y, the loss function is defined as Eq. (4):

$$L_{GAN}(G_X, D_Y, X, Y) = E_{y \sim P_{data}(y)}[logD_Y(y)] + E_{x \sim P_{data}(x)}[log(1 - D_Y(G_X(x)))] \tag{4}$$

For generator $G_Y : Y \rightarrow X$, discriminator D_X, the loss function is defined as Eq. (5):

$$L_{GAN}(G_Y, D_X, Y, X) = E_{x \sim P_{data}(x)}[logD_X(x)] + E_{y \sim P_{data}(y)}[\log(1 - D_X(G_Y(y)))] \tag{5}$$

Cycle consistency loss function is defined as Eq. (6):

$$\begin{aligned} L_{Cycle}(G_X, G_Y) = {} & E_{x \sim P_{data}(x)}[||Vgg((G_Y(G_X(x))) - Vgg(X)||_1] \\ & + E_{y \sim P_{data}(y)}[||Vgg(G_Y(G_X(y))) - Vgg(Y)||_1] \end{aligned} \tag{6}$$

For the object loss function, object loss function is defined as Eq. (7):

$$\begin{aligned} L_{Total} = {} & L_{GAN}(G_X, D_Y, X, Y) + L_{GAN}(G_Y, D_X, Y, X) + L_{Cycle}(G_X, G_Y) + L_1(G_X(x), x) \\ & + L_1(G_Y(y), y) + L_2(G_X(x), x) + L_2(G_Y(y), y) \end{aligned} \tag{7}$$

3.3 Spectral Normalization

In 2017, WGAN introduced that the value of each convolution kernel of discriminator must satisfy Lipschitz constraint. In fact, the way to solve this problem is to set the value greater than 1 to 1 in the convolution kernel. Then, the loss function of the discriminator is actually discontinuous. It even leads to difficult optimization problems, and various approaches have been proposed, such as WGAN-GP. In this paper, spectral normalization is conducted to achieve Lipschitz stability constraint. The core idea is the spectral normalization of the weight matrix in convolutional network, and using the maximum singular value to scale the weight matrix. In the proposed model, the generator adopts convolutional neural networks, residual networks, and deconvolutional neural networks in sequence. The discriminator adopts five layers neural network according to human visual characteristics. For the convolutional neural network in the proposed model, the value of the weight matrix is limited to [0, 1]. It is beneficial for the proposed model to reduce the model oscillation. In order to obtain the spectral normalization of each convolutional network parameter matrix, the singular value of the parameter matrix needs to be solved. In this paper, the power iteration method is used to approximate the value. The iteration process is as shown in Algorithm 1.

Algorithm 1: Power iteration method

1. $V_l^0 \leftarrow$ a random Gaussian vector;
2. **WHILE:**
3. $u_l^k \leftarrow W_l V_l^{k-1}$, normalization: $u_L^k \leftarrow \frac{u_l^k}{\|u_l^k\|}$;
4. $v_l^0 \leftarrow (W_l)^T u_l^k$, normalization: $v_l^k \leftarrow \frac{v_l^k}{\|v_l^k\|}$;
5. **END-WHILE**
6. $\sigma_l(W) = (u_l^k)^T W v_l^k$;
7. $WW^T u = \sigma(W) \cdot \mathrm{u} \Rightarrow u^T WW^T \ u = 1 \ \sigma(W)$, as $\|u\| = 1$;
8. $\sigma(W) = u^T W v$, as $v = W^T u$;

The discriminator is mainly composed of a convolutional neural network. It contains a five-layer convolutional neural network to extract the high dimensional features of the image. The size of the convolution kernel is set to 4 at each layer of the convolutional neural network. The proposed discriminator is designed according to human visual characteristics. It's worth noting that every convolutional kernel is subjected to spectral normalization after each iteration. Thus, the value of the convolution kernel is under Lipschitz stability constraint. The architecture of the discriminator is illustrated in Fig. 3.

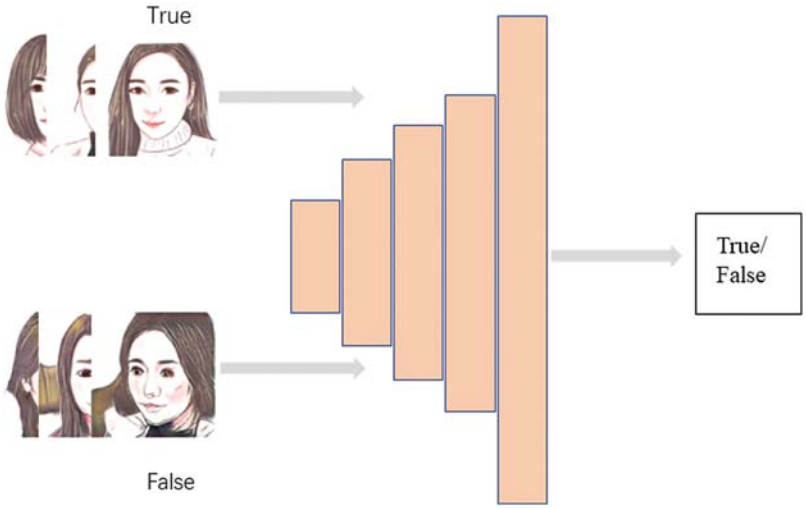


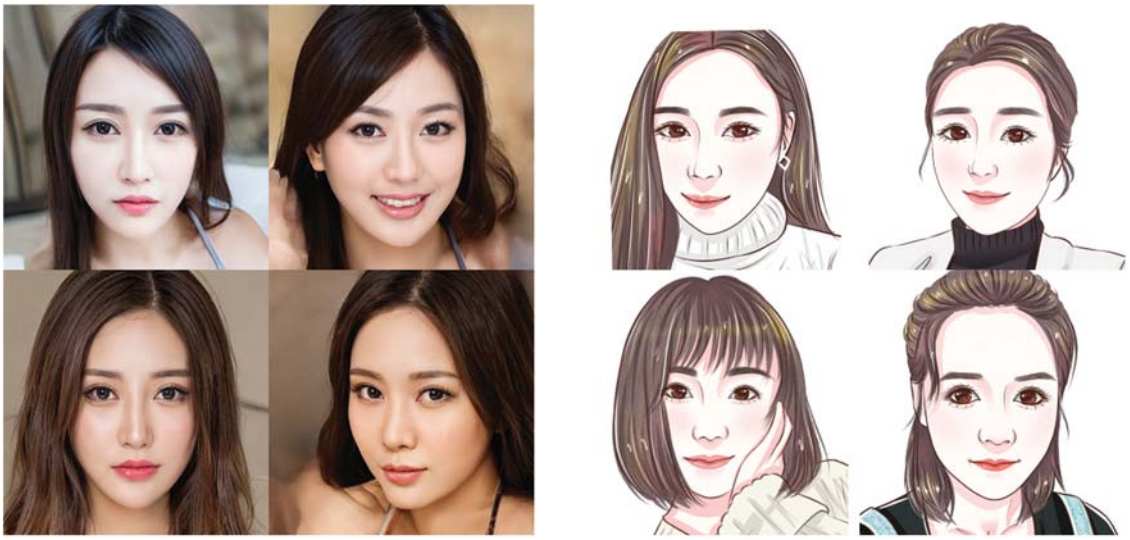
Figure 3: The architecture of the discriminator with adding spectral normalization

4 Results and Analysis

4.1 Datasets

In the experiment, we collect 200 images of faces and 200 cartoon images. Face images are divided into training set X and test set Y. Similarly, cartoon images are divided into training set Y and test set Y. Training set and test set are divided according to the 8 to 2. Besides, increasing the number of training images and improving the quality of training images are very helpful for the proposed model to converge. If the number of images in training set X and training set Y

is not the same, our proposed model selects an equal number of images. The visual examples of training images are shown in Fig. 4.



(a) (b)

Figure 4: Visual examples of the training set. (a) Face images. (b) Cartoon images

4.2 The Training Parameters

The epoch is set to 200, and each epoch contains 200 steps. The learning rate is set to 0.0002 and the random seed is set to 100. Adam optimizer [58] is used, with the default parameter set to 0.5. The loss function value is recorded every 200 steps. The result of the style transfer is saved every 1000 steps. In the convergence process of the proposed model, the image is normalized to reduce computation. In order to improve the experimental result, it's effective to use cartoon images with obvious outlines.

4.3 Experimental Results

In fact, the value of loss function converges more easily and the oscillation is less by adding spectral normalization. This helps to reduce computation time. In the Figs. 5–8, the horizontal axis represents the number of steps, and the vertical axis represents the loss value. The figure shows discriminator loss function value in domain X and domain Y. Besides, the content consistency loss function value and the object loss function value of the proposed model are shown. The number of the training steps for the proposed model is 160,000. The learning rate changes after one epoch in this experiment. Based on the CYCLE-GAN, the loss function value of discriminator in domain X and domain Y is shown in Fig. 5, and the value of content consistent loss function and the object loss function is shown in Fig. 6.

After adding spectral normalization in the discriminator and pretrained model (VGG16), the training set and training parameters were consistent with the original model. Compared with CYCLE-GAN, it is found that the oscillation of the loss function value is significantly reduced in the discriminator. The value of the object loss function converges quickly in the proposed model. As shown in Figs. 7 and 8.

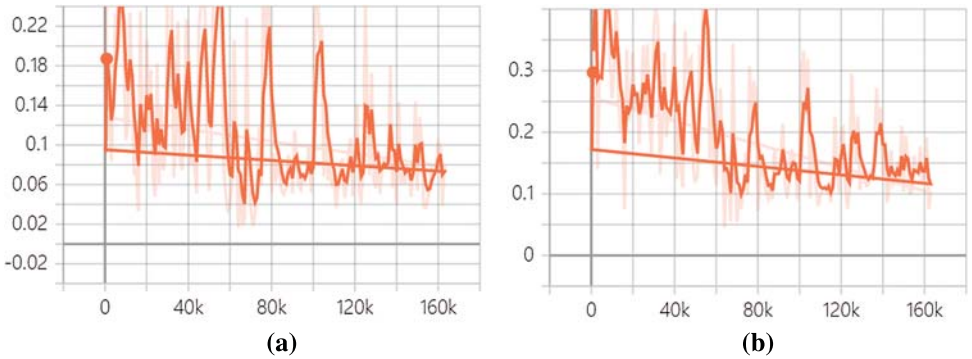


Figure 5: The loss function value of discriminator in CYCLE-GAN. (a) Loss function value in domain X. (b) Loss function value in domain Y

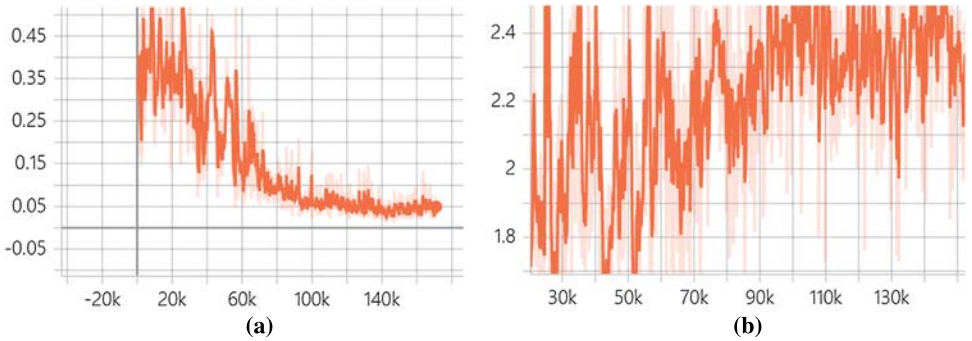


Figure 6: The content consistent loss function value and the object loss function value in CYCLE-GAN. (a) Content consistent loss function value. (b) Object loss function value

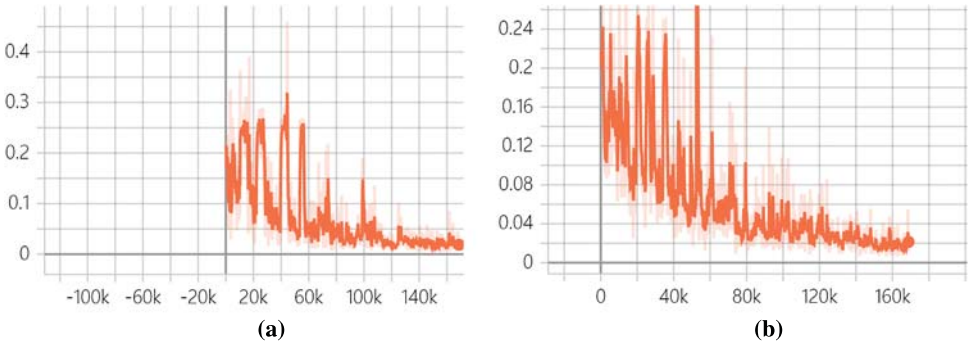


Figure 7: The loss function value of discriminator in proposed model. (a) The loss function value of discriminator in domain X. (b) The loss function value of discriminator in domain Y

As shown in the Fig. 9, the face image is transferred into a cartoon image. The cartoon image generated by CYCLE-GAN and our proposed model respectively is shown in Fig. 9.

In this paper, we evaluate the FID [59] scores between face images and the generated images. The FID scores comparison between CYCLE-GAN and ours is shown in Tab. 1. The image number in the table represents the generated image (the cartoon image) respectively in Fig. 9.

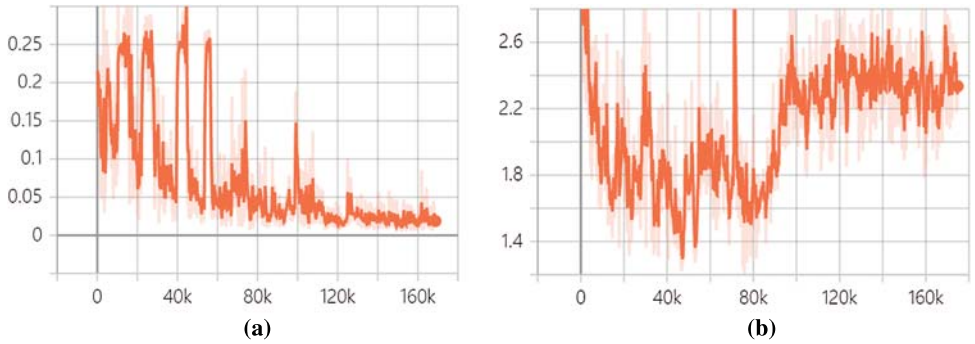


Figure 8: The content consistent loss function value and object loss function value in proposed model. (a) Content consistent loss value. (b) object loss function value

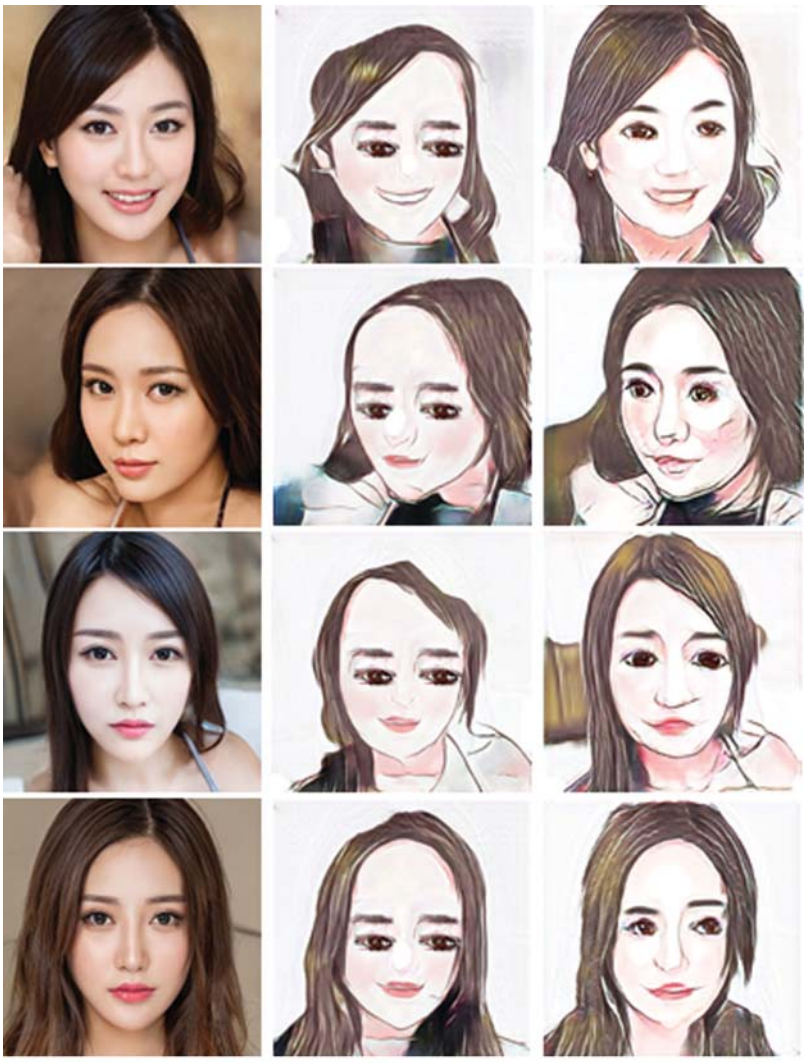


Figure 9: Visual examples after style transfer. (a) Face images. (b) CYCLE-GAN. (c) Ours

Table 1: Performance evaluation based on the FID metric. Lower is better for the FID metric

FID to photo	Image 1	Image 2	Image 3	Image 4	Mean	Total
CYCLE-GAN	253.9	252.4	254.3	249.5	253.5	1010.1
Ours	206.8	210.6	218.9	209.4	213.4	845.7

5 Conclusions

Style transfer mainly relies on the unsupervised learning characteristics of GAN. On this basis, the generator deceives the discriminator by enhancing its ability to fake. The discriminator strengthens its discriminability by constantly discriminating the generated images. In this paper, it is different from CYCLE-GAN, we proposed to add pretrained model (VGG16) to control content loss in the position of l1 loss. Besides, spectral normalization is used to reduce the oscillations of the loss function value. In the convergence process of the proposed model, it is found that the quality of the cartoon image plays an important role in style transfer. Thus, it is very necessary to select high-quality training set. The higher the resolution of the cartoon image, the deeper the neural network needs to be selected. To learn complex image features, it is necessary to increase the depth and width of the neural network. However, the problem of GAN model collapse remains to be solved. Notably, the value of the object loss function drops, and the generated image is very distorted. Therefore, we did many attempts to design a reasonable GAN structure.

Acknowledgement: We thank all the team members for their efforts.

Funding Statement: This work is supported by the National Natural Science Foundation of China (No. 61702226); the 111 Project (B12018); the Natural Science Foundation of Jiangsu Province (No. BK20170200); the Fundamental Research Funds for the Central Universities (No. JUSRP11854).

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

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* "Generative adversarial nets," in *Proc. Advances in Neural Information Processing System*, Montreal, Canada, pp. 2672–2680, 2014.

[2] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow and B. Frey, "Adversarial autoencoders," 2015. [Online]. Available: https://arxiv.org/abs/1511.05644.

[3] J. Kim, M. Kim, K. Hyeonwoo and H. L. Kwang, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-Instance normalization for image-to-image translation," 2020. [Online]. Available: https://arxiv.org/abs/1907.10830.

[4] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein GAN," 2017. [Online]. Available: https://arxiv.org/abs/1701.07875.

[5] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville, "Improved training of wasserstein GANs," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, USA, pp. 5769–5779, 2017.

[6] J. B. Zhao, M. Mathieu and Y. LeCun, "Energy-based generative adversarial networks," 2017. [Online]. Available: https://arxiv.org/abs/1609.03126.

[7] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow," 2018. [Online]. Available: https://arxiv.org/abs/1810.00821.

[8] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06434.

[9] V. Dumoulin, J. Shlens and M. Kudlur, "A learned representation for artistic style," 2016. [Online]. Available: https://arxiv.org/abs/1610.07629.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen *et al.,* "Analyzing and improving the image quality of StyleGAN," in *Proc. 2020 IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8110–8119, 2020.

[11] A. Abarghouei, Amir and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via Image style transfer," in *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 2800–2810, 2018.

[12] P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, Hawwi, USA, pp. 1125–1134, 2017.

[13] M. Y. Liu and T. Oncel, "Coupled generative adversarial networks," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, Spain, pp. 469–477, 2016.

[14] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. 2017 IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2223–2232, 2017.

[15] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim *et al.,* "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8789–8797, 2018.

[16] Y. Choi, Y. Uh, J. Yoo and J. W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. 2020 IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8188–8197, 2020.

[17] C. Yang, Y. K. Lai and Y. J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 9465–9474, 2018.

[18] T. Kim, C. Moonsu, H. Kim, J. K. Lee and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," 2019. [Online]. Available: https://arxiv.org/abs/1703.05192.

[19] Z. L. Yi, H. Zhang, P. Tan and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. 2017 IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 2868–2876, 2017.

[20] Z. He, W. Zuo, M. Kan, S. Shan and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[21] W. Jiang, S. Liu, C. Gao, J. Cao, R. He *et al.,* "PSGAN: Pose and expression robust spatial-aware GAN for customizable makeup transfer," in *Proc. 2020 IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 5194–5202, 2020.

[22] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 2414–2423, 2016.

[23] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 1, 2019.

[24] R. Adbal, Y. P. Qin and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space," in *Proc. 2019 Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 4432–4441, 2019.

[25] J. Gui, Z. N. Sun, Y. G. Wen, D. C. Tao and J. P. Ye, "A review on generative adversarial networks: Algorithms, theory and applications," 2020. [Online]. Available: https://arxiv.org/abs/2001.06937.

[26] A. Brock, J. Donahue and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018. [Online]. Available: https://arxiv.org/abs/1809.11096.

[27] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2018. [Online]. Available: https://arxiv.org/abs/1710.10196.

[28] M. X. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 2019 Int. Conf. on Machine Learning*, California, pp. 6105–6114, 2019.

[29] S. Tulyakov, M. Y. Liu, X. D. Yang and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1526–1535, 2018.

[30] H. Y. Dong, X. D. Liang, Y. X. Zhang, X. J. Zhang, Z. Y. Xie *et al.,* "Fashion editing with adversarial parsing learning," in *Proc. 2020 IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8120–8128, 2020.

[31] T. Miyato, T. Kataoka, M. Koyama and Y. Yoshida, "Spectral norm regularization for improving the generalizability of deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1705.10941.

[32] D. Ulyanov, A. Vedaldi and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016. [Online]. Available: https://arxiv.org/abs/1607.08022.

[33] T. Miyato, T. Kataloka, M. Koyama and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018. [Online]. Available: https://arxiv.org/abs/1802.05957.

[34] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Proc. 2017 Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 1945–1953, 2017.

[35] Y. Wu and K. He, "Group normalization," in *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake, USA, pp. 3–19, 2018.

[36] X. Huang and B. Serge, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. 2017 IEEE Int. Conf. on Computer Vision*, Venice, Italy, pp. 1510–1519, 2017.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 2015 Int. Conf. on Machine Learning*, Lille, France, pp. 448–456, 2015.

[38] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. 2016 Int. Conf. on Neural Information Processing Systems*, Barcelona, Spain, pp. 901–909, 2016.

[39] G. J. Qi, "Loss-sensitive generative adversarial networks on Lipschitz densities," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1118–1140, 2020.

[40] Z. L. Huang, X. G. Wang, L. C. Huang, C. Huang, Y. C. Wei *et al.,* "CCNet: criss-cross attention for semantic segmentation," in *Proc. 2019 IEEE Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 603–612, 2019.

[41] G. Daras, A. Odena, H. Zhang and A. G. Dimakis, "Your local GAN: Designing two dimensional local attention mechanisms for generative models," in *Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 14519–14527, 2019.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556.

[43] K. Roth, A. Lucchi, S. Nowozin and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, USA, pp. 2018–2028, 2017.

[44] L. Mescheder, A. Geiger and S. Nowozin, "Which training methods for GANs do actually converge," in *Proc. 2018 Int. Conf. on Machine Learning*, Stockholm, Sweden, pp. 3481–3490, 2018.

[45] L. A. Gatys, A. S. Ecker and M. Bethge, "A neural algorithm of artistic style," *Journal of Vision*, vol. 16, no. 12, pp. 326, 2016.

[46] E. Risser, W. Pierre and B. Connelly, "Stable and controllable neural texture synthesis and style transfer using histogram losses," 2017. [Online]. Available: https://arxiv.org/abs/1701.08893.

[47] R. Yin, "Content aware neural style transfer," 2016. [Online]. Available: https://arxiv.org/abs/1601.04568.
[48] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 2479–2486, 2016.
[49] J. Liao, Y. Yao, L. Yuan, G. Hua and S. B. Kang, "Visual attribute transfer through deep image analogy," in *Proc. 2017 Int. Conf. on Computer Graphics and Interactive Techniques*, vol. 36, pp. 120, 2017.
[50] J. Johnson, A. Alexandre and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 694–711, 2016.
[51] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang *et al.,* "Real-time neural style transfer for videos," in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, Piscataway, USA, pp. 7044–7052, 2017.
[52] U. Dmitry, V. Lebedev, A. Vedaldi and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *Proc. 2016 Int. Conf. on Machine Learning*, vol. 48, pp. 1349–1357, 2016.
[53] Y. Li, F. Chen, J. Yang, Z. Wang, X. Lu *et al.,* "Universal style transfer via feature transforms," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, USA, vol. 30, pp. 386–396, 2017.
[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh *et al.,* "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.
[55] A. Rahnama, A. T. Nguyen and E. Raff, "Robust design of deep neural networks against adversarial attacks based on lyapunov theory," in *Proc. 2020 IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 8178–8187, 2020.
[56] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 770–778, 2016.
[57] D. E. Rumelhart, E. David and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 696–699, 1988.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015. [Online]. Available: https://arxiv.org/abs/1412.6980.
[59] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Advances in Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6626–6637, 2017.