Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241144

Exploration of classical neural network architecture in cycleGAN framework with face photo-sketch synthesis

Xinyu Wang

Faculty of Science and Technology, University of Macau, 999078, Macau

dc02785@um.edu.mo

Abstract. CycleGAN has been a benchmark in the style transfer field and various extensions with wide applications and excellent performance have been introduced in recent years, however, discussion about its architecture exploration which could enable us to further understand the concept of generative model is scarce. In this paper, several architectures referenced from classical convolutional neural networks are implemented into the generator and discriminator of the cycleGAN model, including AlexNet, DenseNet, GoogLeNet, and ResNet. Their feature extraction modes are imitated and modified into blocks to embed into the encoder part of the generator while the discriminator directly uses their model except it outputs a patch classification. In advance to mitigate the possible imbalance between generator and discriminator ability, a self-adjusting learning rate strategy based on the discriminator confidence is introduced. Multiple evaluation metrics are utilized to measure the performance of each model. Experimental results indicate an AlexNet-like architecture model could achieve a competitive performance than the baseline cycleGAN and present better fine details and high-frequency information.

Keywords: Generative Adversarial Network, Convolutional Neural Network, Style Transfer.

1. Introduction

Style transfer is a conspicuous topic in the computer vision and artificial intelligence field, which enables combining art and humanities with technological advances in neural networks and applying them to reality. It reconstructs the source image that keeps geometry content and converts texture appearance style into target one [1]. Human face style transfer has always been a popular topic among style transfer ever since its origin [1], style transfer between reality and animation style or painting style is relatively mature in various applications. Style transfer between photo and sketch could have a variety of applications, ranging from criminal investigation where hand-drawn suspect portraits based on description could be converted into realistic style photos which provide more appearance information so that police and the public could identify them more efficiently, to entertainment use for example to create a sketch style self-portrait as social media avatar, to virtue reality scenarios synthesis, where creator could draw only laconic lines of face to get a realistic looking portrait and greatly reduce workload.

In the fundamental research of style transfer, content and style are demonstrated to be distinguishable in the neural network, which then was used to reconstruct images constrained by content loss and style loss, the former one from a feature map in a deep layer while the later one from each layer [1]. Then the network was extended to Convolutional Neural Network (CNN) [2], and perceptual loss was introduced,

DOI: 10.54254/2755-2721/50/20241144

computed by comparing features from the target and generated image extracted with a VGG [3]. CycleGAN inherent its feature-extraction network architecture and implement the adversarial architecture and break the limitation of requiring paired dataset of other works [4, 5], where dual generators and discriminators in both directions are trained to raise cycle-consistency loss, that is the style-transferred image could be inversely converted back to the original source image, ensuring the generator would transfer appearance style while preserving geometry content [6]. Later research on style transfer such as StarGAN explored style transfer in multiple domains [7], and StyleGAN proposed an alternative generator architecture that enables intuitive control of synthesis [8]. These later researches ameliorated the architecture and promote the generated quality at a higher level, though little research has been done on architecture based on CycleGAN. Therefore, this paper hopes to explore different network architectures in both generator and discriminator of CycleGAN based on existing classical convolutional neural network architectures.

This research would focus on the effect of different network structures on portrait style transfer. Various convolutional neural network architectures in CycleGAN [6] would be implemented, including AlexNet [9], DenseNet [10], GoogLeNet [11], and ResNet [12]. These classical architectures would be implemented in discriminator while down-sampling convolutional generator layers are replaced by blocks designed to contain characteristics from corresponding classical architecture. Additionally, a self-adjusting learning rate strategy was attempted to solve the training stability problem caused by the difference in generator and discriminator strengths. These CycleGAN could convert images between photo and sketch domains in both directions, and the approach could also be generalized to other domains if data is available. Synthesis images are evaluated by multiple metrics to statistically measure the generative quality of each architecture.

2. Method

2.1. Dataset preparation

This paper uses CUHK Student Face Sketch dataset published in 2009 [13], in which contains in total of 188 faces with blue backgrounds. For each face, there is a corresponding sketch drawn by an artist based on a photo taken in a frontal pose, under normal lighting conditions, and with a neutral expression. Image data is in RGB format and has a relatively high resolution of 200×250. This paper will employ 100 pairs of images as the training set, and 88 pairs as the testing set. Some sample images of the collected dataset are shown in Figure 1.

Some basic image-augmentation methods are also implemented on the dataset for data preprocessing, including resizing images to 256×256 to fit with models input, images have 0.5 possibilities to be horizontally flipped so the trained models could be more robust and normalization that map data input into a normal distribution with mean=0 and standard deviation=1 to accelerate the convergence of model training.

DOI: 10.54254/2755-2721/50/20241144

Figure 1. Example of CUHK Student Face Sketch dataset.

2.2. CycleGAN model

CycleGAN is an adversarial generative network for style transfer proposed by Zhu et al. in 2017, style transfer methods like pix2pix before it requires paired data which is a heavy limitation. This burden is solved by cycleGAN by dual generator and dual discriminators and introducing cycle-consistency loss. Therefore, based on classical GAN architecture, that input image x into the generator G synthesis $y' = G(x)$, and put into discriminator D_y with real target image y, to calculate the loss function below.

$$\min_{D_y} \max_{G} \mathcal{L}_{GAN} = \mathbb{E}_{y \sim P_{data}(y)}[\log D_y(y)] + \mathbb{E}_{x \sim P_{data}(x)}[\log (1 - D_y(G(x))] \tag{1}$$

This loss function would restrain the discriminator D_y to classify between the image from target training data y and the image generated by the generator $G(x)$ and generator G to synthesize image similar to the target image to deceive the discriminator D_y. The problem is that if training data x and y are not paired images, that is they do not share a common geometry content, the generator G might map the same set of input images to any random permutation of image in the target domain while ignoring the source content because discriminator D_y only classifies if the input image having the target style and that is enough to confuse D_y. That is why cycleGAN introduce cycle-consistency loss showed below.

$$\mathcal{L}_{cyc} = \mathbb{E}_{y \sim P_{data}(y)}[\|G(F(y)) - y\|] + \mathbb{E}_{x \sim P_{data}(x)}[\|F(G(x)) - x\|] \tag{2}$$

The cycle-consistency loss restricts the $G(F(y))$ to be similar with y, and vice versa, to let $F(G(x))$ to be similar with x. By letting G convert the source style image to the target style, then convert it back to the source style by F to output the original image, generators would be constrained to only modify the style of the image and preserve content.

Besides cycle-consistency loss, cycleGAN also introduces identity loss though it is not contained in paper. It aims to preserve the original color style, to prevent from generator modifying color.

$$\mathcal{L}_{identity} = \mathbb{E}_{y \sim P_{data}(y)}[\|F(y) - y\|] + \mathbb{E}_{x \sim P_{data}(x)}[\|G(x) - x\|] \tag{3}$$

The full objective would be:

$$\mathcal{L} = \mathcal{L}_{GAN} + \mathcal{L}_{cyc} + \mathcal{L}_{identity} \tag{4}$$

DOI: 10.54254/2755-2721/50/20241144

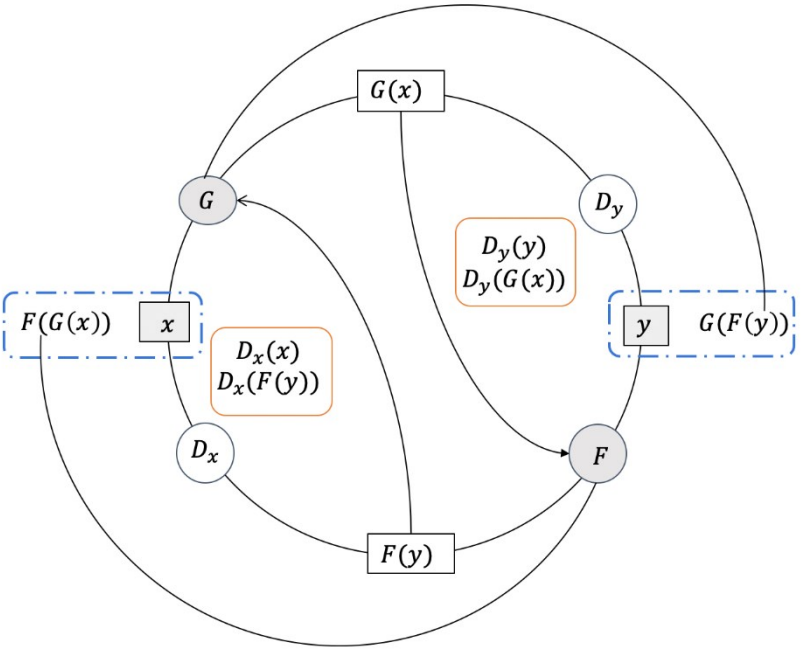

Figure 2. Demonstration of CycleGAN workflow.

2.3. Implementation details

2.3.1. Architecture. Models in this paper will follow the principle and structure including the generator and discriminator of cycleGAN, The generator is adapted from the network proposed by Johnson et al. [5] which contains two stride-2 convolutions to extract features, nine residual blocks [14] to restore and augmentation images, and two 2-stride inverse-convolutions to reconstruct the image back to its original size. In both the down-sampling stage encoder and up-sampling stage decoder, instance normalization [15] accelerates model convergence and maintains independence between image instances, and ReLU as activation function. Patch-GANs [4] is implemented for discriminator, similar to cycleGAN. However, depending on the various models embedded, the size of the overlapping image from Path-GANs will be different.

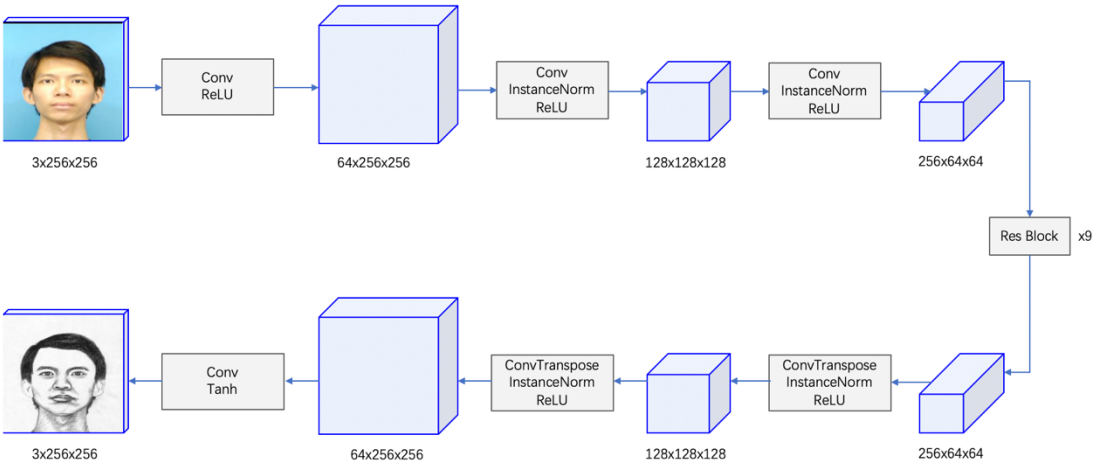

Figure 3. Demonstration of cycleGAN architecture.

DOI: 10.54254/2755-2721/50/20241144

2.3.2. Generator. The down-sampling stage is extracting abstract features from the input image, with the price of geometry content. Therefore, if the down-sampling is too much, that is extracting features and down-size image to less than $1/4$ of its original length, the feature preserved would be so abstract that it is even impossible for the up-sampling process to reconstruct its geometry content. Then it would be reasonable to explore how to extract the features more efficiently within a limited down-sampling step. This paper will present some architecture designed based on the principle of some classical convolutional neural networks, that are utilized in the down-sampling block. Including AlexNet [9], DenseNet [10], GoogLeNet [11], and ResNet [12], would be both imitated in the down-sampling stage of the generator and partially used in the discriminator. The various types of neural networks used endeavor to comprise their primary concepts while modifications are carried out to fit the purpose of style transformation. These down-sampling blocks have a common initial layer, and share exactly the same input and output size of $64 \times 256 \times 256$ from preprocessed image data and output $256 \times 64 \times 64$ feature into the residual blocks.

Encoder architectures are illustrated in Figure 4. The self-defined convolutional block is also from cycleGAN implementation, containing a convolution layer, an instance normalization layer, and a ReLU activation, this convolution block is able to extract features from image identity efficiently and will be utilized frequently in architectures. In terms of the original cycleGAN, two such 3×3 convolutional blocks directly compose the down-sample module. For the AlexNet-like architecture, each convolutional block is followed by a 2×2 max pooling layer to further draw abstract features. DenseNet-like architecture is composed of a dense network and a transpose network, in the former, outputs of two convolutional blocks would concatenate with all previous inputs, and the latter takes the responsibility to map feature channels and size to the output standard and further extract information. The GoogLeNet-like architecture is an imitation of the Inception block [11], where four separate networks extract features in different levels and concatenate by a max pooling layer. The weights for different branches are from GoogLeNet, sequence from left to right in Figure 4 (e), the first block is 2 :4 :1 :1, and the second block is 4 :6 :3 :2. The last one is the ResNet-like architecture, outputs of two 3×3 convolutional blocks is concatenated with its own input, and a 1×1 convolutional block to map feature into standard output format.

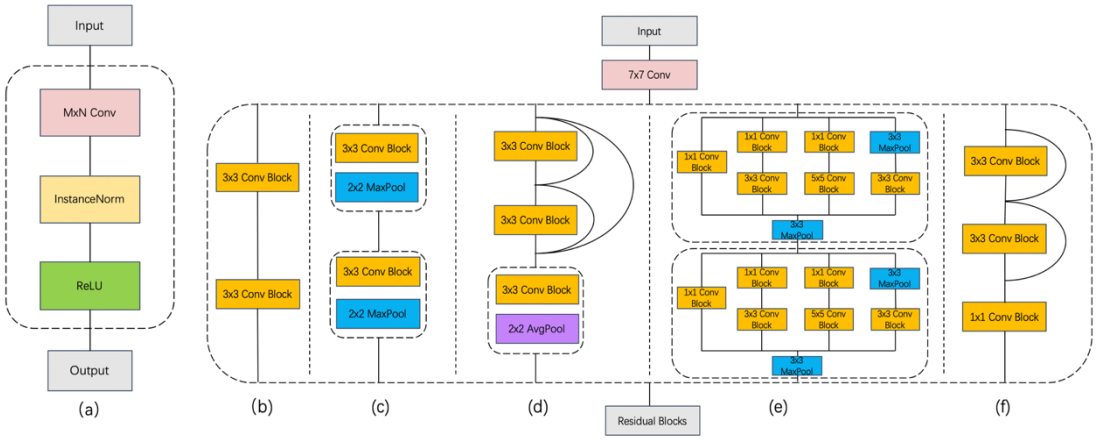

Figure 4. Demonstration of down-sampling model architectures. (a) a self-defined $M \times N$ Convolutional Block, used in other architectures as $M \times N$ Conv Block. (b) cycleGAN baseline architecture. (c) AlexNet-like architecture. (d) DenseNet-like architecture. (e) GoogLeNet-like architecture. (f) ResNet-like architecture.

2.3.3. Discriminator. The application of these architectures on discriminator is more straightforward, this paper will directly use models from torchvision library with modifications to fit with Patch-GAN.

DOI: 10.54254/2755-2721/50/20241144

For a better comparison of these network structures, discriminators should output feature maps in the same, or at least close channels.

2.3.4. Self-adjusting learning rate strategy. This experiment faces the challenge that, all of these convolutional neural networks are designed for classification, that is for discriminator, and their performance on generator has yet to be proven. Therefore, there might be an ability gap between the generator and discriminator [16], this would lead to the phenomenon that the discriminator converges at a speed far more rapid than the generator, then whatever generator synthesis the discriminator could always identify it as fake and thus loss the function of directing generator to synthesis image similar to target domain in adversarial generative network. To prevent this, a self-adjusting learning rate strategy is implemented, which will automatically coordinate the learning rate of the generator and discriminator, in order to maintain a relative symmetric ability between them. It is achieved by supervising the confidence of discriminators to classify real and fake image inputs as real, with the difference between confidence in real input and fake input increasing, which implies that the convergent speed of the discriminator is beyond that of the generator, learning rate of discriminator will decrease as it increases for generator. In this way, the convergent speed difference will be constrained so the discriminator could play the role of director. The performance of this strategy is shown in Figure 5.

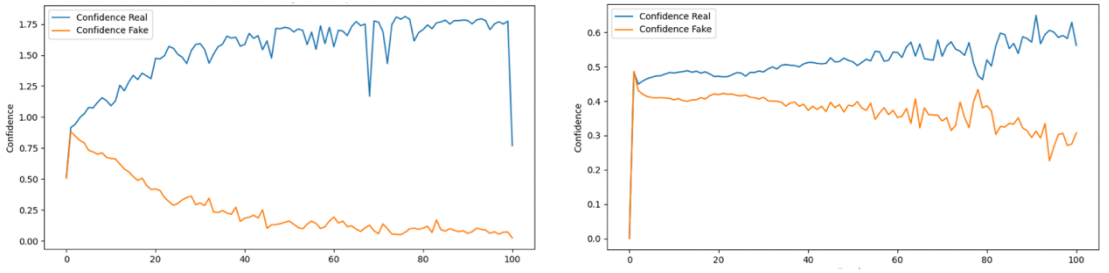

Figure 5. Demonstration of discriminator confidence of predicting real image as real and predicting fake image as real when ability of generator and discriminator are unbalanced, without (a) and with (b) self-adjusting learning rate strategy.

3. Results and discussion

Table 1. FID, PSNR, and SSIM scores of each synthesis method. For these models, Source is the synthesized photo-style image and Target is the synthesized sketch-style image.

Method	**Original**		**AlexNet-like**		**DenseNet-like**		**GoogLeNet-like**		**ResNet-like**	
	Source	Target	Source	Target	Source	Target	Source	Target	Source	Target
FID↓	*34.0495*	*28.2799*	***22.2442***	***17.5150***	*31.2464*	*52.9721*	*41.9954*	*41.0583*	*42.5291*	*34.9678*
PSNR↑	***20.0200***	***20.1956***	*19.1518*	*13.1724*	*18.9084*	*17.8109*	*17.6106*	*18.4159*	*19.5422*	*18.5674*
SSIM↑	*0.5653*	*0.5825*	***0.6436***	***0.6963***	*0.4970*	*0.4449*	*0.5588*	*0.6296*	*0.5829*	*0.6417*
Sec per Epoch	*10.1*		*11.3*		*25.1*		*16.8*		*16.3*	

Synthesis result samples of each architecture are illustrated in Figure 6, and several generation evaluation metrics are implemented and demonstrated in Table 1. Including Fréchet Inception Distance

DOI: 10.54254/2755-2721/50/20241144

using CLIP feature [17], PSNR, and SSIM. Results are the average value of 10 independent training with 20 epochs on each architecture, completed on a V-100 GPU.

These results illustrated that the AlexNet-like architecture achieved a better result than the original cycleGAN on FID and SSIM measurements, however, for other more complex architectures it turned out to be worse. One of the most significant reasons for this phenomenon is, that although a self-adjusting learning rate strategy is used and mitigates the issue, these models still suffer from the unbalance of the generator and discriminator, that their classify ability as discriminator is too powerful and convergent speed is too rapid.

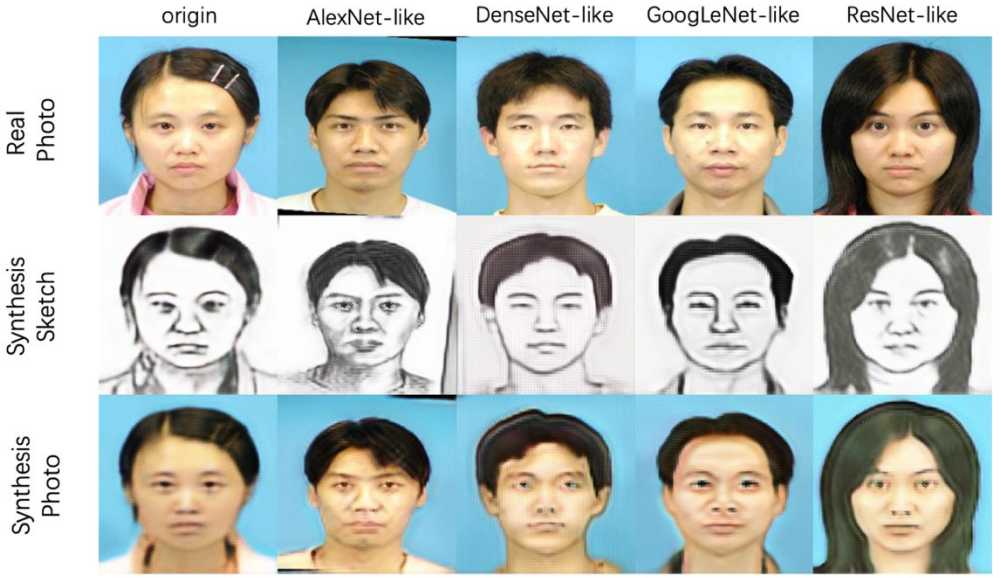

Figure 6. Generate results of each architecture.

Apart from the asymmetric of the generator and discriminator, poor performances of DenseNet-like, GoogLeNet-like, and ResNet-like architectures could also be caused by the encoder extracting features to a too abstract level, that the decoder could not reconstruct fine detail from remaining information. Therefore, the results of these architectures lack the ability to express fine details, this is evident in the facial lighting and hair. From the synthesis sketch-style images, compared with sketch-style images synthesized by the original cycleGAN and AlexNet-like architecture model, they neglect the shadow part of images with only simple facial lines. And for hair, they turn to express hair like a whole entity, withdrawing detailed hair line. These fine details might be preserved in the output of the original and AlexNet-like architecture model's encoders, but discarded from that of other models.

Compared with the original result, AlexNet-like architecture appears to synthesize images with more exact fine details. Focus on the area near the eyes, where the lines are denser and information is expressed at a higher frequency, it is conspicuous that eyes synthesized by the original are fuzzier than those generated by the AlexNet-like one, it is hard to distinguish between black pupil and white area of the eye in the former, while the latter present this distinguish clearly. Within its encoder blocks, the convolutional layer is used to extract features into higher dimensions, followed by the max pooling layer to down-sample feature maps to a smaller size. This information-extracting strategy better clusters global meaningful information and presents a relatively better performance in the synthesis of high-frequency information.

This paper only experiments with paired-architecture encoder and decoder and the encoder with AlexNet-like architecture only experimented with the decoder with AlexNet-like architecture. More combinations should be available, to group different encoder and decoder architectures to test for their performance. For example, use the original decoder to work with each encoder, so the result would better represent the generative ability of each encoder, with less risk of model asymmetric.

DOI: 10.54254/2755-2721/50/20241144

4. Conclusion

In this research, different architectures of cycleGAN are explored on human face style transfer between photo and sketch style. Structures from classical convolutional neural networks are implemented in the cycleGAN framework to seek for a better synthesis model architecture. Experiments with a self-adjusting learning rate strategy are conducted to evaluate the generative ability of each architecture. The result showed that cycleGAN with an AlexNet-like architecture model can achieve a better performance in both human-vision perception and various synthesis metrics, compared with baseline cycleGAN architecture. This paper only experiments with generators and discriminators with the same architectural style. In the future, more architecture combinations will further explore architecture's effectiveness on generative models and find architecture designs with better performance.

References

[1] Gatys L. A. and Ecker A. S. and Bethge M. (2015) A neural algorithm of artistic style arXiv preprint arXiv:1508.06576.

[2] Gatys L. A. and Ecker A. S. and 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).

[3] Johnson J. and Alahi A. and Fei-Fei L. (2016) Perceptual losses for real-time style transfer and super-resolution In Computer Vision–ECCV 2016: 14th European Conference Amsterdam The Netherlands October 11-14 2016 Proceedings Part II 14 (pp. 694-711) Springer International Publishing.

[4] P. Isola J.-Y. Zhu T. Zhou and A. A. Efros Image-toimage translation with conditional adversarial networks In CVPR 2017.

[5] J. Johnson A. Alahi and L. Fei-Fei Perceptual losses for real-time style transfer and super-resolution In ECCV pages 694–711 Springer 2016.

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

[7] Choi Y. and Choi M. and Kim M. and Ha J. W. and Kim S. and Choo J. (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).

[8] Karras T. and Laine S. and Aila T. (2019) A style-based generator architecture for generative adversarial networks In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[9] Krizhevsky A. and Sutskever I. and Hinton G. E. (2012) Imagenet classification with deep convolutional neural networks Advances in neural information processing systems 25.

[10] Huang G. and Liu Z. and Van Der Maaten L. and Weinberger K. Q. (2017) Densely connected convolutional networks In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[11] Szegedy C. and Liu W. and Jia Y. and Sermanet P. and Reed S. and Anguelov D. and Rabinovich A. (2015) Going deeper with convolutions In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[12] He K. and Zhang X. and Ren S. and Sun J. (2016) Deep residual learning for image recognition In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[13] Wang X. and Tang X. (2008) Face photo-sketch synthesis and recognition IEEE transactions on pattern analysis and machine intelligence 31(11) 1955-1967.

[14] He K. and Zhang X. and Ren S. and Sun J. (2016) Deep residual learning for image recognition In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

DOI: 10.54254/2755-2721/50/20241144

[15] Ulyanov D. and Vedaldi A. and Lempitsky V. (2016) Instance normalization: The missing ingredient for fast stylization arXiv preprint arXiv:1607.08022.
[16] Gilbert A. C. and Zhang Y. and Lee K. and Zhang Y. and Lee H. (2017) Towards understanding the invertibility of convolutional neural networks arXiv preprint arXiv:1705.08664.
[17] Parmar G. and Zhang R. and Zhu J. Y. (2022) On aliased resizing and surprising subtleties in gan evaluation In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11410-11420).