SCIENCE CHINA
Information Sciences

• **Supplementary File** •

BiTGAN: bilateral generative adversarial networks for Chinese ink wash painting style transfer

Xiao HE[1], Mingrui ZHU[1*], Nannan WANG[1*], Xiaoyu WANG[2] & Xinbo GAO[3]

[1]*State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an* 710071*, China;*
[2]*The Chinese University of Hong Kong (Shenzhen), Shenzhen* 518172*, China;*
[3]*Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing* 400065*, China*

Appendix A Inspiration

Inspired by the wide usage of UNet in the field of image segmentation, we note that the U-shape structure (a contracting path and a symmetric expanding path) can obtain and combine multi-scale features so that the features can contain context information and better maintain the image structure. However, using UNet directly to complete the neural style transfer task will make the image have a poor style because the skip connections may transfer low-level information (e.g., color) from the source image to the output image and produce visual artifacts. According to an earlier style transfer study [1], we have the following intuition: the shallow features are mostly color and texture information, and the higher-level features capture more complex structures. The arbitrary style transfer method [2] [3] [4] usually selects the $Relu\_3\_1$ or $Relu\_4\_1$ layer in the VGG network for style transfer. The domain transfer method [5] usually uses ResNet and other architectures with fewer down-sampling layers as generators, but it will inevitably lose depth feature information.

Based on the above observation and inspired by [6], we propose a novel GAN-based model called BiTGAN that is designed based on image-to-image translation architecture and combined with the Adaptive Instance Normalization (AdaIN) module. It can achieve domain-level style transfer. Specifically, based on CycleGAN [4], we present a new bilateral generator for Chinese ink wash painting style transfer. It consists of two paths: ResNet Path (RP) and UNet Path (UP). As their names imply, ResNet Path is equipped with stacked residual blocks (ResBlocks) and UNet Path consists of a contracting path and a symmetric expanding path. To the best of our knowledge, the U-shape structure can perform multi-scale feature extraction and fusion. So the UNet Path can obtain context information and maintain the structure of the content image under the constraint of cycle consistency loss and adversarial loss. On the other hand, inspired by the AdaIN module [1], we know that instance normalization performs style normalization by normalizing the feature statistics, which have been found to carry the style information of an image [7] [8]. So when the features of UP are up-sampled to the same size as the features derived from the corresponding layer of RP, we use AdaIN to align their mean and variance, which allows UP to obtain the global style information from RP. As a result, when adversarial loss pushes the output image to conform to the distribution of the Chinese ink wash painting domain, UP will receive the global style information (e.g., textures, colors) from RP through adaptive instance normalization layers. In this way, our method could achieve satisfactory performance in the stylization of Chinese ink wash painting.

Appendix B Loss function

Our goal is to learn mapping functions between the photo domain X and the Chinese ink wash painting domain Y without paired training samples. So based on CycleGAN, we propose to train our model with adversarial loss and cycle consistency loss.

Appendix B.1 Adversarial loss

Given unpaired training samples x_i where $x_i \in X$ and y_i where $y_i \in Y$. we denote the data distribution as $x \sim p_{\text{data}}\ (x)$ and $y \sim p_{\text{data}}\ (y)$. We have two mapping functions, $G : X \to Y, F : Y \to X$ and two corresponding discriminators D_Y, D_X. For the mapping function $G : X \to Y$, we express the adversarial loss as:

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}\ (y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}\ (x)}[\log(1 - D_Y(G(x)))], \tag{B1}$$

where G attempts to generate the images $G(x)$ that similar to the target domain Y, while D_Y is aims at distinguishing between translated samples $G(x)$ and real samples y. The objective of mapping F is minimized over G and maximized over D_Y, i.e., $\min_G \max_{D_Y} \mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right)$. The other mapping F has the similar objective: $\min_F \max_{D_X} \mathcal{L}_{\text{GAN}}\left(F, D_X, Y, X\right)$.

Appendix B.2 Cycle consistency loss

Adversarial training can learn mappings G and F that generate images identically distributed as target domain Y and X. However, the network may map the same set to any arbitrary permutation of images in the target domain with a large enough capacity. Any of the learned mappings can induce an output distribution that matches the target distribution. Therefore, we introduce the cycle consistency loss to constraint the mappings. Given the image x from domain X, the image translation should be able to bring x back to the original domain, i.e., $x \to G(x) \to F(G(x)) \approx x$. And for the image y from domain Y also has the similar constraint, i.e., $y \to F(y) \to G(F(y)) \approx y$:

* Corresponding author (email: mrzhu@xidian.edu.cn, nnwang@xidian.edu.cn)

$$\mathcal{L}_{\text{cyc}}\ (G, F) = \mathbb{E}_{x\sim p_{\text{data}}\ (x)}\left[\|F(G(x)) - x\|_1\right] + \mathbb{E}_{y\sim p_{\text{data}}\ (y)}\left[\|G(F(y)) - y\|_1\right]. \tag{B2}$$

This cycle consistency loss can push the generated image to preserve the structural information of the source domain, so the generated image can preserve the content while resembling the style from Chinese ink wash painting domain Y.

Appendix B.3 Full objective

Consequently, we combine all the above loss functions as our full objective as follows:

$$\mathcal{L}\left(G, F, D_X, D_Y\right) = \mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) + \mathcal{L}_{\text{GAN}}\left(F, D_X, Y, X\right) + \lambda\mathcal{L}_{\text{cyc}}(G, F), \tag{B3}$$

where λ is the hyper-parameters controlling the relative importance of the two objectives.

Appendix C Experiments

Table C1 Average inference time of different methods under 256×256 resolution. The best results are in **bold** while the second-best results are marked with an underline.

Method	BiTGAN	ChipGAN	CycleGAN	UGATIT	AdaIN	MST
Inference time(s)	0.0078	0.0055	0.0054	0.0532	**0.0030**	0.7307

Table C2 Quantitative comparison with other methods. The best results are in **bold** while the second-best results are marked with an underline.

Method	BiTGAN	ChipGAN	CycleGAN	UGATIT	AdaIN	MST
FID ↓	221.09	225.19	**220.92**	311.39	279.74	259.22
MS-SSIM ↑	**0.67**	0.65	0.62	0.10	0.13	0.11

Table C3 User study results. The percentage of votes for six style transfer methods on "ChipPhi" dataset. The best results are in **bold** while the second-best results are marked with an underline.

Method	BiTGAN	ChipGAN	CycleGAN	UGATIT	AdaIN	MST
Percentage(%)	**39.19**	25.37	18.63	6.60	6.97	3.24

Appendix C.1 Baseline

We compare our method with image style transfer methods including AdaIN and MST [9], which are reference-guided image style transfer schemes. In addition, we compare our approach with CycleGAN, ChipGAN, UGATIT [10], the state-of-the-arts in reference-free image translation. They are most relevant to our method.

AdaIN presents a simple yet effective method that enables arbitrary style transfer in real-time for the first time. The core ingredient of the method is a novel adaptive instance normalization layer, which aligns the mean and variance of the content features with the style features to extract the style information of target images.

MST explicitly considers the matching of semantic patterns in content and style images. It clusters style features into sub-style pattern components and matches them with the content features according to the graph cutting formulation. In this way, it can transfer complex style patterns and achieve satisfactory local style patterns.

CycleGAN learns a mapping $G : X \rightarrow Y$ to generate the output image $G(x)$, which has the same distribution as the target domain Y. In order to reduce the possible mappings, cycle consistency constraints are proposed, which add an inverse mapping $F : Y \rightarrow X$ and push $F(G(x))$ to be equal to x. This method can generate a realistic image with unpaired data.

UGATIT proposes a new attention module and a new learnable normalization in an end-to-end manner. The attention module guides the model to focus on more important regions between the source and target domains. The AdaLIN function helps the model flexibly control the changes of shape and texture according to the parameters learned from the dataset.

ChipGAN is based on CycleGAN architecture. It considers the unique painting skills of Chinese ink wash painting and put forward two losses: edge loss to emphasize vigorous lines and ink wash loss to mimic the ink wash diffusion and tone.

Appendix C.2 Inference Speed

Our model is trained on a Quadro RTX 8000 GPU. We compare the inference speed of different style transfer methods under 256×256 resolutions using RTX 8000. As shown in Table C1, the reference time of our method is in line with other methods, both in the order of single-digit milliseconds, which means it can well meet the needs of real-time applications.

Appendix C.3 Qualitative evaluations

In this section, we will quantitatively compare our method with the approach introduced above. As shown in Figure C1, AdaIN directly adjusts second-order statics of content features globally, which can not convert the style well. MST explicitly considers the matching of semantic patterns in content and style images. But it results in damaged content structures and unreasonable textures (e.g., the sixth row and the seventh row in Figure C1). Reference-free image translation method CycleGAN loses some brush strokes and generates some unnecessary ones, which damages the content structure (e.g., the first row and the fourth row in Figure C1). UGATIT proposes a new attention mechanism that guides the model to focus on more important regions distinguishing between

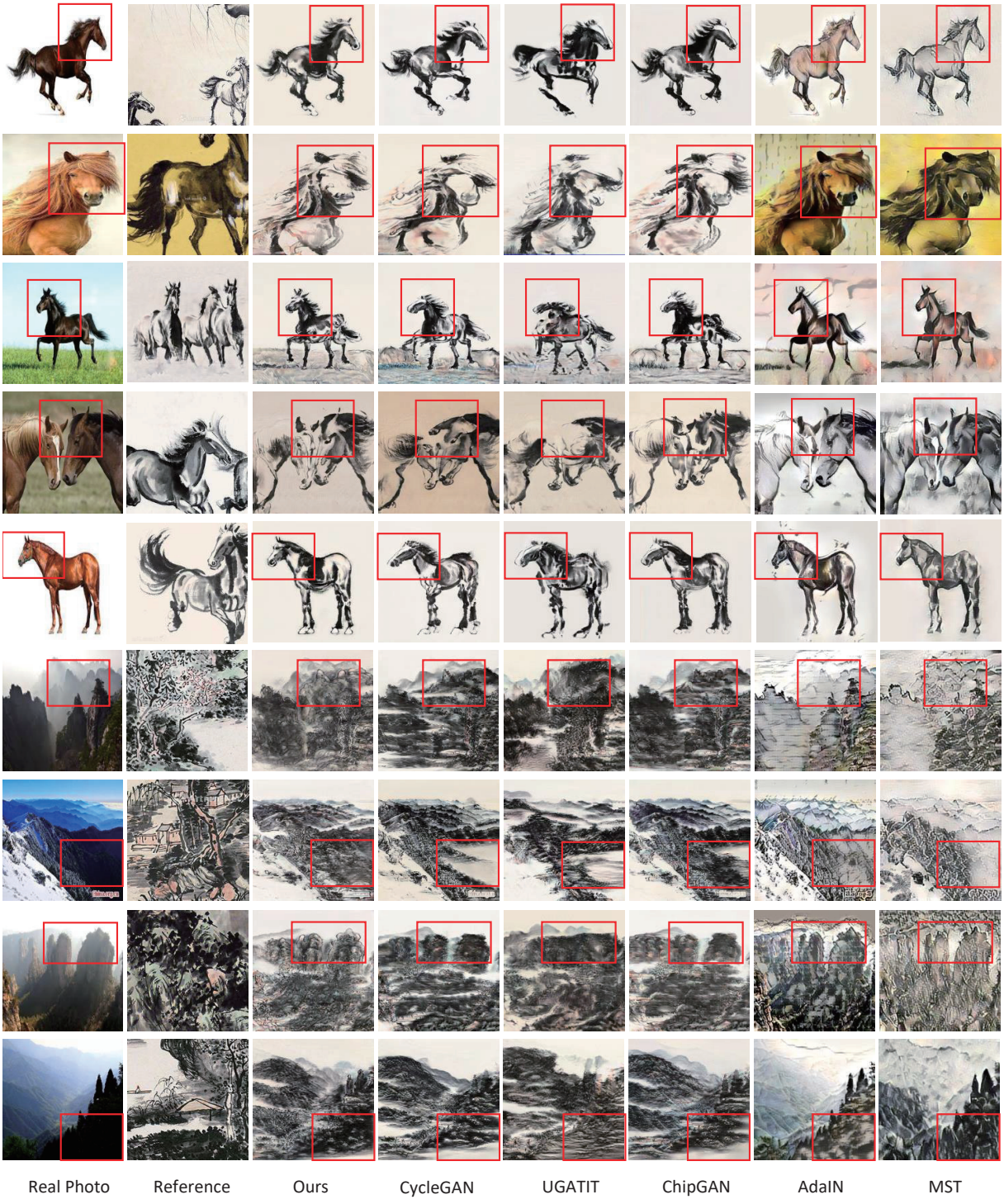

Figure C1 Visual results comparison with other methods on "ChipPhi" dataset. Because AdaIN and MST are reference-guided method, we give the corresponding reference images. Other methods do not need extra input.

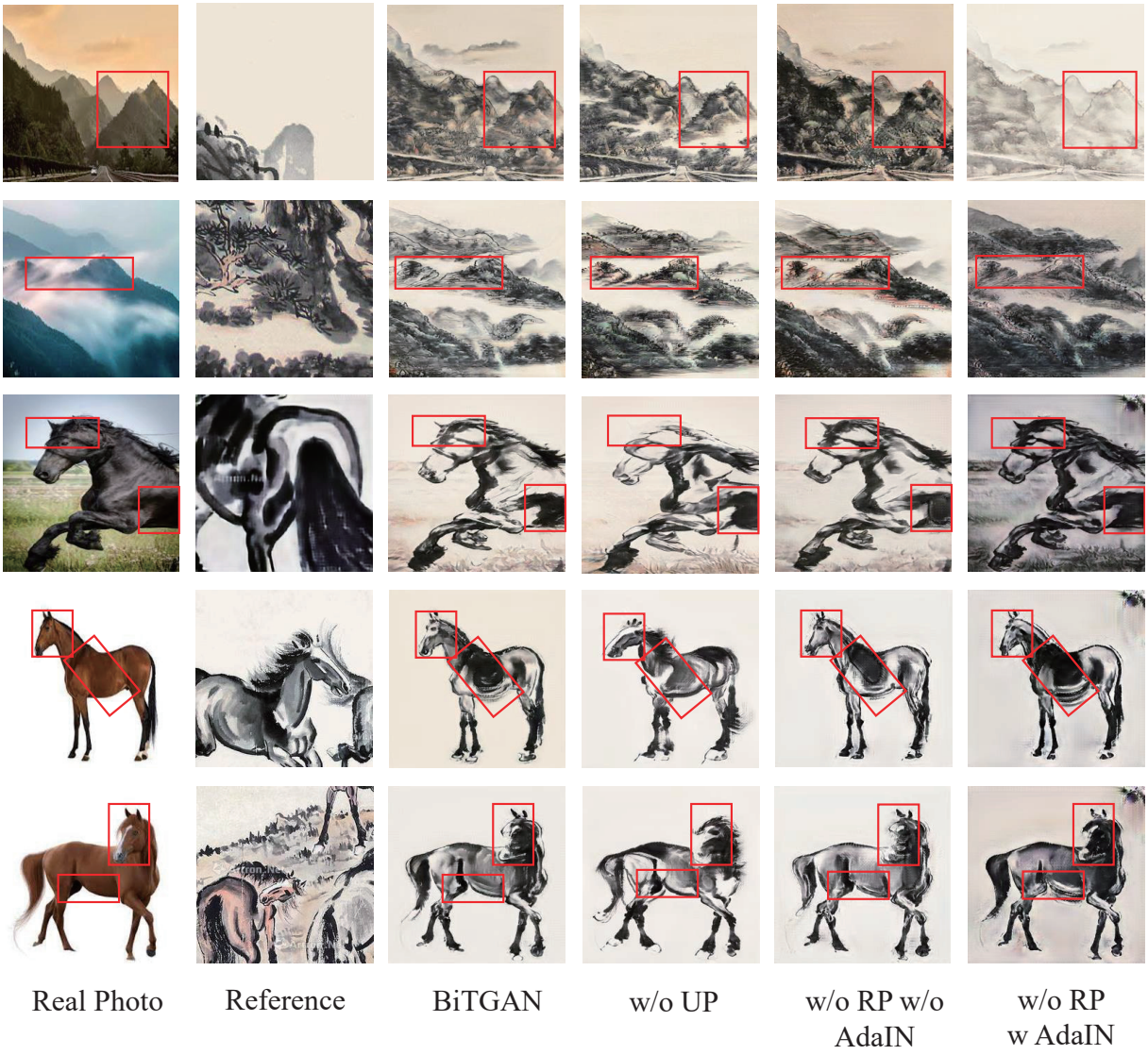

Figure C2 Visual quality comparison of different variants of our method. From left to right : Real Photo, Reference, BiTGAN (with full objective), w/o UNet Path (w/o UP), w/o ResNet Path w/o AdaIN (w/o RP w/o AdaIN), w/o ResNet Path w AdaIN (w/o RP w AdaIN). The reference image is used to normalize the UNet Path in the variants of w/o RP w AdaIN. The red box indicates the most recognizable regions of the generated image. It includes the change of content structure, the differences in ink wash diffusion and tone, and the different brush strokes.

source and target domains, which leads to severe content details being lost in other regions (e.g., the second row in Figure C1). ChipGAN considers the key techniques of Chinese ink wash painting, adds edge loss to show strong lines, and uses eroded and blurred images to mimic ink wash diffusion and tone. However, it can only perform stylization in shallow layers and ignores the context information, which results in some unreasonable voids (e.g., the seventh row and the ninth row in Figure C1). In addition, it destroys the original structure of the image in the process of style transfer. For example, in Figure C1, the posture of the horse's head in the third row has changed, and some content details are lost in the second and fourth rows. BiTGAN not only transforms shallow features through ResBlocks but also obtains context information by using a U-shaped structure, which makes the layout of ink more reasonable. As shown in column 3, our method contains certain areas of voids and can preserve the content structure while resembling the style from the Chinese ink wash painting domain.

Appendix C.4 Quantitative evaluations

We use Fréchet Inception Distance (FID) [11] and Multiscale Structural Similarity (MS-SSIM) [12] to evaluate the quality of generated images. FID measures the distance between two distributions of real images and generated images, and lower indicates the style of the generated image is similar to the real one. MS-SSIM measures the distance between the generated image and the input image, and higher indicates the generated image preserves better content information. We calculated the average FID and MS-SSIM scores of the results of horse and landscape datasets. Table C2 shows the corresponding quantitative results. In terms of general distribution, CycleGAN and BiTGAN outperform the other methods. However, CycleGAN can not maintain the content structure as well as BiTGAN. Our results can effectively preserve both the input content and the reference style simultaneously.

Appendix C.5 User study

To further evaluate the quality of generated images, we conduct a user study under human perception. 50 subjects are invited to participate in the study, whose ages range from 20 to 40. We randomly selected 20 content images, which includes landscape photos and horse photos. And then we showed each subject the stylized results obtained by the six style transfer methods side-by-side in a random order. We asked the subject to choose his/her favorite result for Chinese ink wash painting style and show the percentage of votes for each method in Table C3. The results show that our algorithm is superior to other methods.

Appendix C.6 Ablation study

In this section, We performed an ablation study on some key factors in the proposed model to quantify the contribution of each factor to the overall effectiveness.

UNet Path. In this section, we seek to validate whether the UNet Path effectively extracts the context information of the source image and preserves the content structure well. We hence implement a baseline called "w/o UNet Path" which only uses ResNet Path as the generator architecture. As shown in Figure C2 , "w/o UNet Path" tends to ignore the context information due to the shallow convolution. For example, "w/o UNet Path" loses some brush strokes and generates some unnecessary ones that damage the content structure (e.g., the vertical red box of the fourth row and fifth row in Figure C2). This result demonstrates that UNet Path could extract the context information and help the generator preserve the content structures during the process of stylization.

ResNet Path and AdaIN. We build two baselines to evaluate the effectiveness of ResNet Path and AdaIN layer. The first baseline, named "w/o ResNet Path w/o AdaIN", is constructed independently using UNet Path. The second baseline, named "w/o Resnet Path w/AdaIN", is constructed by using a reference image to normalize the UNet Path instead of the Resnet Path. As shown in Figure C2, "w/o ResNet Path w/o AdaIN" has the wrong color and several black spots, which can not reflect the tone and diffusion effect of ink wash painting (e.g., the red box in the second row and the red box in the lower right corner of the third row in Figure C2). The second baseline "w/o ResNet Path w/AdaIN" can extract the style patterns from the reference image and remove the black spots. However, it can not generates a robust output because of the different reference images. This result shows that our ResNet Path and adaptive instance normalization layer are helpful for generating high-quality Chinese ink wash paintings.

References

1 Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2414-2423
2 Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 1501-1510
3 Li Y, Fang C, Yang J, et al. Universal style transfer via feature transforms. Advances in neural information processing systems, 2017, 30
4 Li X, Liu S, Kautz J, et al. Learning linear transformations for fast image and video style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 3809-3817
5 Chen Y, Lai Y K, Liu Y J. Cartoongan: Generative adversarial networks for photo cartoonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 9465-9474
6 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 4401-4410
7 Li C, Wand M. Combining markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 2479-2486
8 Li Y, Wang N, Liu J, et al. Demystifying neural style transfer. 2017. ArXiv:1701.01036
9 Zhang Y, Fang C, Wang Y, et al. Multimodal style transfer via graph cuts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 5943-5951
10 Kim J, Kim M, Kang H, et al. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. 2019. ArXiv:1907.10830
11 Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in neural information processing systems, 2017. 4,5,13
12 Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment. In: proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003. 2: 1398-1402