Style Guided Face to Anime Translation using Novel GAN

Dr. Vijay Kumar Bohat, Aryan Sehgal, Sourabh, Aditya Divyang, M Aman Chaudhry
Department of Computer Science and Engineering
Netaji Subhas University of Technology (formerly NSIT)
New Delhi, India
vijay.bohat@nsut.ac.in, aryan.sehgal.cs19@nsut.ac.in,
sourabh.cs19@nsut.ac.in aditya.divyang.cs19@nsut.ac.in, maman.cs19@nsut.ac.in

***Abstract*—We propose a novel architecture to translate a portrait photo-face into an anime appearance. Our aim is to synthesize anime-faces which are style-consistent with a given reference anime-face. Existing methods often fail to transfer the styles of reference anime-faces, or introduce noticeable artifacts/distortions in the local shapes of their generated faces. We propose a novel *Generative Adversarial Network (GAN)* in an attempt to solve this problem. the proposed model as a generator for style guided image-to-image translation and a discriminator that captures domain-shared distributions and domain-specific distributions of face images and anime images.**

***Index Terms*—anime-face, anime-face generation, generative adversarial network, style transfer, deep learning**

I. Introduction

Animations play an important role in our daily life and have been widely used in entertainment, social, and educational applications. Recently, anime, aka Japan-animation, has been popular in social media platforms. Many people would like to transfer their profile photos into anime images, whose styles are similar to that of the roles in their favorite animations

However, commercial image editing software fails to do this transfer, while manually producing an anime image in specific styles needs professional skills.

In this paper, we aim to automatically translate a photoface into an anime-face based on the styles of a reference anime-face. We refer to such a task as Style-Guided Face-to-Anime Translation (StyleFAT). Inspired by the advances of generative adversarial networks (GANs) [13], many GAN-based methods (e.g., [19,24,45,51]) have been proposed to automatically translate images between two domains. However, these methods [30,51] focused on learning one-to-one mapping between the source and target images, which does not transfer the information of the reference image into a generated image. Consequently, the styles of their generated anime-faces [30, 51] are usually dissimilar from that of the reference ones.

Recently, a few reference-guided were proposed for multi-modal translation which generates diverse results by additionally taking reference images from the target domain as input. These methods, however, usually fail to fulfill the StyleFAT task and generate low-quality anime images.

Different from the image translation tasks of referenc eguided methods, StyleFAT poses new challenges in two aspects

- An anime-face usually has large eyes, a tiny nose, and a small mouth which are dissimilar from natural ones. The significant variations of shapes/appearances between anime-faces and photo-faces require translation methods to largely overdraw the local structures (e.g., eyes and mouth) of a photo-face, different from caricature translation [5] and makeup-face transfer [6] which preserve the identity of a photo-face.,
- Anime-faces involve various appearances and styles (e.g. various hair textures and drawing styles). Such large intra-domain variations poses challenges in devising a generator to translate a photo-face into a specificstyle anime-face, as well as in training a discriminator to capture the distributions of anime-faces.

To address the above problems, we propose a novel GAN-based model. Since it is difficult to collect pairs of photo-faces and anime-faces,we train our model with unpaired data in an unsupervised manner.

We propose a new generator architecture that preserves the global information (e.g., pose) of a source photo-face, while transforming local facial shapes into anime-like ones and transferring colors/textures based on the style of a reference anime-face. The proposed generator does not rely on face landmark detection or face parsing. Our insight is that the local shapes (e.g., large and round eyes) can be treated as a kind of styles like color/texture.

In this way, transforming a face's local shapes can be achieved via style transfer. To transform local facial shapes via style transfer, we explore where to inject the style information into the generator. In particular, the multi-layer feature maps extracted by the decoder represent multi-level semantics (i.e., from high-level structural information to low-level textural information). Our generator therefore injects the style information into the multi-level feature maps of the decoder.

In addition to the generator, we propose a novel discriminator, that explicitly considers large appearance variations between photo-faces and anime-faces as well as variations among anime images. The discriminator learns domain-specific distributions and intra-domain distributions, so as to mitigate artifacts in generated faces.

Our major contributions are summarized as follows:

- We propose a new generator to simultaneously transfer color/texture styles and transform the local facial shapes of a source photo-face into their anime-like counterparts based on the style of a reference anime-face, while preserving the global structure of the source photo-face.
- We devise a novel discriminator to help synthesize high-quality anime-faces, while effectively avoiding noticeable distortions in generated faces via learning cross-domain shared distributions between anime-faces and photofaces.

II. Related Work

Generative Adversarial Networks(GANs)[13] have achieved impressive performance for various image generation and translation tasks [4, 7, 8, 9, 15]. The key to the success of GANs is the adversarial training between the generator and discriminator. In the training stage, networks are trained with an adversarial loss, which constrains the distribution of the generated images to be similar to that of the real images in the training data. In our work, we also utilize an adversarial loss to constrain the image generation. Our model uses GANs to learn the transformation from a source domain to a significantly different target domain, given unpaired training data.

Image-to-Image Translation. With the popularization of GANs, GAN-based image-to-image translation techniques have been widely explored in recent years. For example, trained with paired data, Pix2Pix [18] uses a cGAN framework with an L1 loss to learn a mapping function from input to output images. Wang et al. proposed an improved version of Pix2Pix [19] with a feature matching loss for high-resolution image-toimage translation.

For unpaired data, recent efforts [10] have greatly improved the quality of generated images. Cycle- GAN proposes a cycle-consistency loss to get rid of the dependency on paired data. UNIT maps sourcedomain and target-domain images into a shared-latent space to learn the joint distribution between the source and target domains in an unsupervised manner. MUNIT extends UNIT to multi-modal contexts by incorporating AdaIN [17] into a content and style decomposition structure. However, the style controllability of the above methods is limited due to the fact that the instance-level style features are not explicitly encoded.

Neural Style Transfer. Our problem statement can also be regarded as a kind of the neural style transfer (NST) [11, 12]. In the field of NST, many approaches have been developed to generate paintings with different styles.For example, CartoonGAN [8] devises several losses suitable for general photo cartoonization. ChipGAN [15] enforces voids, brush strokes, and ink wash tone constraints to a GAN loss for Chinese ink wash painting style transfer.CariGANs [5] design special modules for geometric transformation to generate caricatures. However, the above methods either are designed for a specific art field which is completely different from animation, or rely on additional annotations (such as facial landmarks).

III. Proposed Work

A. Datasets

face2anime: A dataset of anime and human faces The dataset contains diverse anime styles (e.g., face poses, drawing styles, colors, hairstyles, eye shapes, strokes, facial contours). The face2anime dataset contains 17,796 images in total, where the number of both anime-faces and natural photo-faces is 8,898. The anime-faces are collected from the Danbooru2019 dataset, which contains many anime characters with various anime styles. For natural-faces, we randomly select 8,898 female faces from the CelebA-HQ dataset. All images are aligned with facial landmarks and are cropped to size 256x256. We separate images from each domain into a training set with 8,000 images and a test set with 898 images.

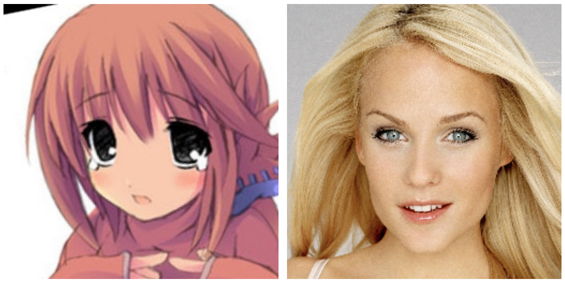

Fig. 1. Sample images from the face2anime Dataset

B. Generator

The generator of StyleFAT model consists of a content encoder Ec, a style encoder Es and a decoder F, as shown in Fig. 2

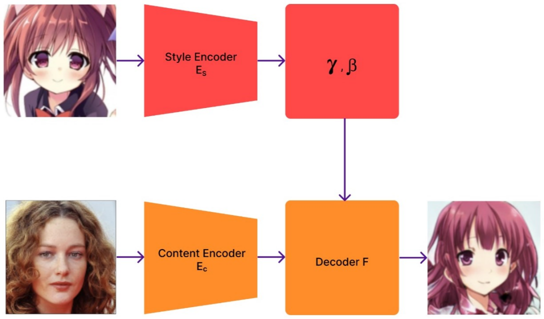

Fig. 2. Architecture of the proposed generator. It consists of a content encoder and a style encoder to translate the source image into the output image reflecting the style of the reference image.

1) Content Encoder: The encoder includes a content encoder Ec and a style encoder Es. Given a source photo-face x the content encoder Ec is used to encode the content of the source image x. The mathematical formulation for the same is as follows:

$$\alpha = Ec(x) \tag{1}$$

where, α is the content code encoded by the content encoder Ec

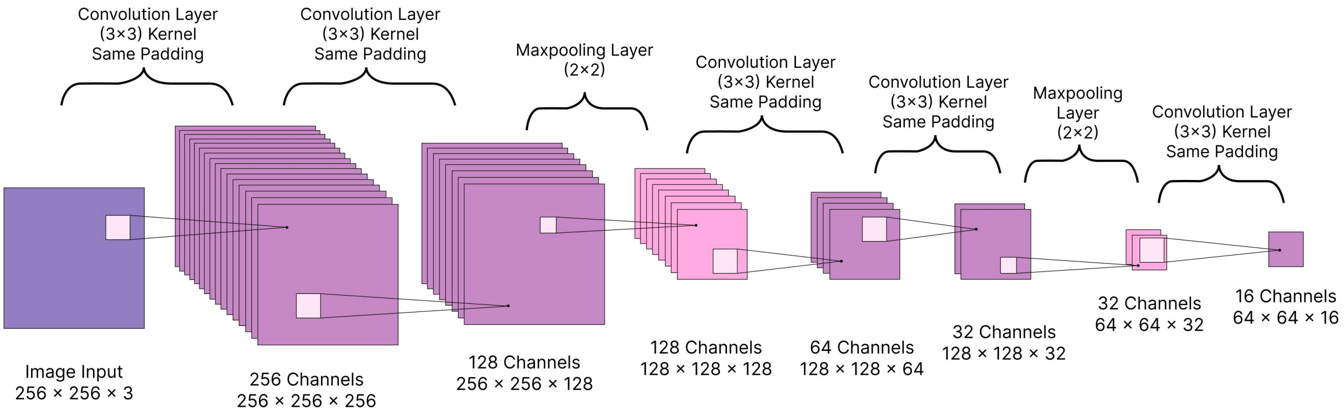

Fig. 3. Architecture of the content encoder. It takes a source image as input and gives an encoding that caputres all the important features of the input human-face.

It produces a content encoding for the human-face. The content encoding has all the important features of the given input. The devised architecture given in Fig. 3

The model consists of convolution layers and maxpooling layer. The input image when passed through the proposed model and produces a feature map of 64x64x16.

2) Style Encoder: This is the second part of the encoder represented by Es. Given a reference anime-face y, the style encoder is employed to extract the style information from the reference y. The mathematical formulation for the same is as follows:

$$(\gamma s, \beta s) = Es(y) \tag{2}$$

It produces a style parameters for the anime-face. The style parameters have all the important style information of the given input. The devised architecture is given in Fig. 5

The model consists of multiple convolution layers, maxpooling layers and batch normalization layer. The style parameters are injected into the content-encoding during decoding process to obtain the resultant image which has content from the human-face image and style from anime-face image.

3) Decoder: The decoder F constructs an image from a content code and style codes. However, different from typical image translations that transfer styles while preserving both local and global structures of source image, our decoder aims to transform the local shapes of facial parts and preserve the global structure of the source photo-face during style transfer.

Recent image translation methods [17, 18] transfer the style of the target domain by equipping residual blocks (resblocks) in the bottleneck of the generator with style codes. However, we observe that such decoder architectures cannot well handle the StyleFAT task. Specifically, they are either insufficient to elaborately transfer an anime style, or introduce visually annoying artifacts in generated faces. To decode high-quality anime-faces for StyleFAT, we propose a novel architecture with no skip connection for the decoder.

Our decoder model takes the content encoding and decodes it, injecting the style parameters to it in multiple layers. The decoder model can be seen in Fig. 4

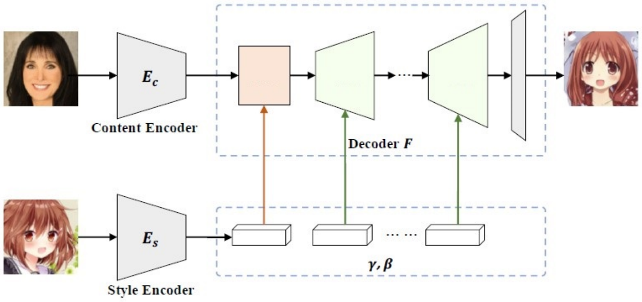

Fig. 4. Decoder model to be used to generate anime faces from content encoding and injecting style parameters to it at various levels.

C. Discriminator

Discriminator constitutes the second part of our GAN that discriminates between real and generated human-faces, and real and generated anime faces. In particular, we assume that anime-faces and photo-faces partially share common distributions and such cross-domain shared distributions constitute meaningful face information, since these two domains are both about faces.

The model takes the output of generator as its input and gives an output probability for real/fake human face and real/fake anime face. The discriminator model learns inter-domain and intra-domain distribution of anime-faces and human-faces.

In other words, by learning and utilizing the cross-domain shared distributions, the discriminator can help reduce distortions and artifacts in translated anime-faces.

The generator and discriminator are trained in an alternate fashion as gradually improving them both.

The discriminator is as shown in the Fig. 6

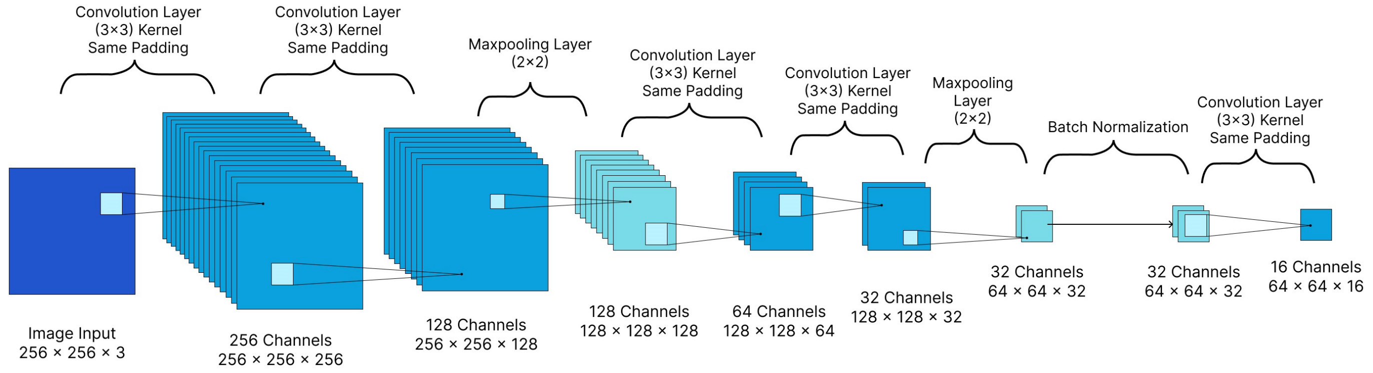

Fig. 5. Architecture of the style encoder. It takes a reference image as input and gives style parameters that has style information for the reference anime-face

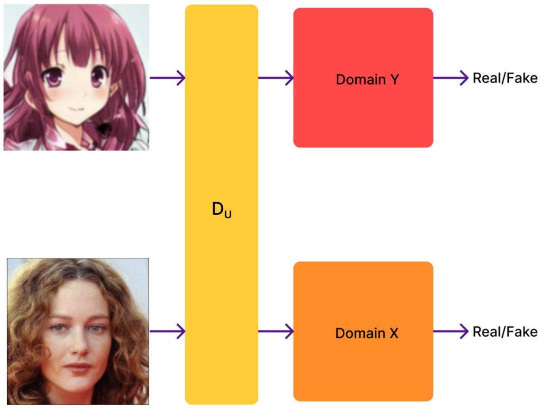

Fig. 6. Diagram for Discriminator

D. Training

GANs have two separate networks(generator and discriminator), its training is also carried out in two phases. Training of each network is carried out one at a time. Generator and discriminator both improve gradually.

However, the model convergence is hard to identify. As the generator improves with training, the discriminator performance gets worse.

This progression poses a problem for convergence of the GAN as a whole: the discriminator feedback gets less meaningful over time. If the GAN continues training past the point when the discriminator is giving completely random feedback, then the generator starts to train on junk feedback, and its own quality may collapse.

We train and evaluate our approach using the face2anime datasets. We use the network architecture mentioned earlier as our backbone. For fast training, the batch size is set to 4 and the model is trained for 10K iterations.

Training our GAN is a very long and computation expensive task, hence the training is carried out is intermediate stages. The model is trained for few iteration and the intermediate results are saved. This creates a checkpoint for model training. The model training continues from the last checkpoint. We use Adam optimizer for this purpose.

The discriminator and generator losses for 50 epochs can be seen in Fig. 7

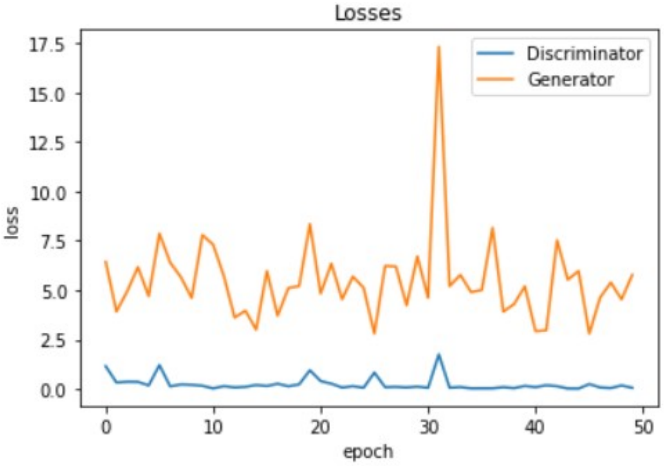

Fig. 7. Lineplot for discriminator and generator loss versus epochs

IV. Results and Discussion

A. Qualitative Analysis

Given a source photo-face and a reference anime-face, a good translation result for this task should share similar/ consistent anime-styles (e.g., color and texture) with the reference without introducing noticeable artifacts, while facial features are anime-like and the global information from the source is preserved.

The results show that CycleGAN introduces visible artifacts in their generated anime-faces (see the forehead in the fourth row). MUINT also leads to visible artifacts in some generated anime-faces, as shown in Fig.8. UGATIT better performs than CycleGAN and MUNIT. However, the anime styles of generated animefaces by CycleGAN, UGATIT and MUNIT are dissimilar to that of the references. DRIT++ is designed for few-shot reference-guided translation, and hence is not suitable

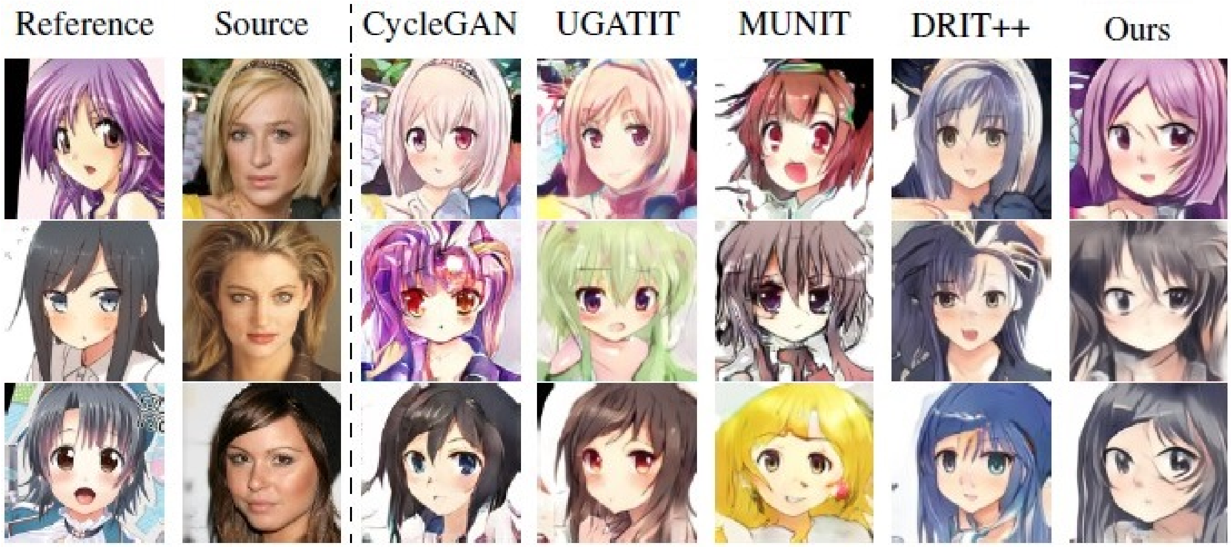

Fig. 8. Comparison of various image translation methods on the face2anime dataset. From left to right: source photo-face, reference anime-face, the results by CycleGAN, UGATIT, MUNIT, DRIT++, and our Model.

for the StyleFAT task. However, it introduces distortion and artifacts to the generated faces.

In contrast, our method generates anime-faces reflecting the various styles of the reference images. In other words, our method achieves the most consistent styles with those of reference anime-faces over the other methods. In addition, our method generates high-quality faces which preserve the poses of source photo-faces, despite a photo-face is partially occluded by other objects.

B. Quantitative Analysis

In addition to qualitative evaluation, we quantitatively evaluate the performance of our method through the visual quality of generated images.

Visual quality. We evaluate the the quality of our results with Frechet Inception Distance (FID) metric [16] which has been popularly used to evaluate the quality of synthetic images in image translation works e.g., [10, 16].

TABLE I
COMPARISON OF FID SCORES ON THE FACE2ANIME DATASET: LOWER IS BETTER

Model	FID Score
CycleGAN	50.09
UGATIT	42.84
MUNIT	43.75
DRIT++	39.98
Ours	**38.45**

The FID score evaluates the distribution discrepancy between the real faces and synthetic anime-faces. A lower FID score indicates that the distribution of generated images is more similar to that of real anime-faces. That is, those generated images with lower FID scores are more plausible as real anime-faces. Following the steps in [16], we compute a feature vector by a pretrained network for each real/generated anime-face, and then calculate FID scores for individual compared methods, as shown in Table I. The FID scores in Table I demonstrate that our model achieves the best score on the face2anime dataset, meaning that the anime-faces generated by our approach have the closest distribution with real anime-faces, thereby making them look similar visually.

V. Conclusion

In this paper, we propose a novel GAN-based method, for style-guided face-to-anime translation. A new generator architecture is proposed, which effectively transfer styles from the reference anime-face, preserve global information from the source photo-face and convert local facial shapes into anime-like ones. We also propose a novel discriminator to assist the generator to produce high-quality anime-faces. Extensive experiments demonstrate that our method achieves superior performance compared with state-of-the art methods.

References

[1] AnimeFace2009, https://github.com/nagadomi/animeface-2009
[2] Danbooru2019, https://www.gwern.net/Danbooru2019
[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016
[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In Proc. Int. Conf. Learn. Rep., 2019.
[5] Kaidi Cao, Jing Liao, and Lu Yuan. Carigans: Unpaired photo-to-caricature translation. ACM Trans. Graphics, 2018.
[6] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: Ondemand makeup transfer framework with reversible generative network. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pages 10042–10050, 2019.
[7] Lei Chen, Le Wu, Zhenzhen Hu, and Meng Wang. Qualityaware unpaired image-to-image translation. IEEE Trans. on Multimedia, 21(10):2664–2674,

2019.

[8] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pages 9465–9474, 2018

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018.

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020.

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576, 2015

[12] Gatys, Leon A and Ecker, Alexander S and Bethge, Matthias. Image style transfer using convolutional neural networks. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pages 2414–2423, 2016

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. pages 2672– 2680, 2014

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In Proc. Adv. Neural Inf. Process. Syst., pages 5767–5777, 2017

[15] Bin He, Feng Gao, Daiqian Ma, Boxin Shi, and Ling-Yu Duan. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In Proc. ACM Int. Conf. Multimedia, pages 1172–1180, 2018

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proc. Adv. Neural Inf. Process. Syst., pages 6626– 6637, 2017

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pages 1501–1510, 2017

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit

[19] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018.

[20] face2anime Dataset: https://github.com/bing-liai/face2anime