MangaGAN: Unpaired Photo-to-Manga Translation Based on the Methodology of Manga Drawing

Hao Su[1], **Jianwei Niu**[1,2,3*], **Xuefeng Liu**[1], **Qingfeng Li**[1], **Jiahe Cui**[1], **Ji Wan**[1]

[1]State Key Lab of VR Technology and System, School of Computer Science and Engineering, Beihang University
[2]Industrial Technology Research Institute, School of Information Engineering, Zhengzhou University
[3]Hangzhou Innovation Institute, Beihang University
{bhsuhao, niujianwei, liu_xuefeng, liqingfeng, cuijiahe, wanji}@buaa.edu.cn

Abstract

Manga is a world popular comic form originated in Japan, which typically employs black-and-white stroke lines and geometric exaggeration to describe humans' appearances, poses, and actions. In this paper, we propose MangaGAN, the first method based on Generative Adversarial Network (GAN) for unpaired photo-to-manga translation. Inspired by the drawing process of experienced manga artists, MangaGAN generates geometric features and converts each facial region into the manga domain with a tailored multi-GANs architecture. For training MangaGAN, we collect a new dataset from a popular manga work with extensive features. To produce high-quality manga faces, we propose a structural smoothing loss to smooth stroke-lines and avoid noisy pixels, and a similarity preserving module to improve the similarity between domains of photo and manga. Extensive experiments show that MangaGAN can produce high-quality manga faces preserving both the facial similarity and manga style, and outperforms other reference methods.

1 Introduction

Manga, originated in Japan, is a worldwide popular comic form of drawing on serialized pages to present long stories. Typical manga is printed in black-and-white (as shown in Fig. 1 left), which employs abstract stroke lines and geometric exaggeration to illustrate humans' appearances, poses, and actions. Drawing manga is a time-consuming process, and even a professional manga artist requires several hours to finish one page of high-quality work. Therefore, automatically translating a face photo to manga in an attractive style is desirable. This task can be described as image translation that is a hot topic in the computer vision area. In recent years, deep learning-based image translation has undergone remarkable progress and a series of systematic methods have been proposed. For instance, the *Neural Style Transfer* (NST) methods (e.g.,(Gatys, Ecker, and Bethge 2016; Li and Wand 2016a; Chen et al. 2017b)) use tailored CNNs and objective functions to stylize images, and the methods based on *Generative Adversarial Network* (GAN) (e.g., (Goodfellow et al. 2014; Isola et al. 2017; Zhu et al. 2017a; Liu, Breuel, and Kautz 2017)) map paired or unpaired images from the original domain to the stylized domain.

Although these recent researches have achieved good performances in their applications, they have difficulties to generate a high-quality manga face due to the following four *challenges*. First, in the manga domain, humans' faces are abstract, geometrically exaggerated, and are far from those in the photo domain. The facial correspondences between the two domains are hard to be established by networks. Second, for different facial features, manga artists always use different drawing styles. These independent features (i.e., appearance, location, size, style) are unable to be extracted and concluded by a network simultaneously. Third, the generated manga face has to preserve the identity of a user without compromising the abstract manga style. It is a challenge to keep both of them with high performances. Therefore, the existing studies of image stylization (e.g., (Gatys, Ecker, and Bethge 2016; Isola et al. 2017; Zhu et al. 2017a; Yi et al. 2019)) are unable to generate high-quality manga face images.

To address these challenges, we present MangaGAN, the first GAN-based method for translating frontal face photos to the manga domain with preserving the attractive style of a popular manga work Bleach[1]. We observed the following important steps of an experienced artist in drawing manga: first outlining the exaggerated face and locating the geometric distributions of facial features, and then fine-drawing each of them. MangaGAN follows the above process and employs a multi-GANs architecture to translate different facial features, and maps their geometric features by another developed GAN model. Furthermore, we present a Similarity-Preserving (SP) module to improve the similarity between domains of photo and manga, and leverage a structural smoothing loss to avoid artifacts.

To summarize, our main contributions are three-fold:

- We propose MangaGAN, the first GAN-based method for unpaired photo-to-manga translation. MangaGAN adopts a novel network architecture to simulate the drawing process of manga artists, and can produce attractive manga faces preserving facial similarity of the input photo.
- We propose a similarity-preserving module that substantially improves the performances on preserving both the facial similarity and manga style. We also propose a structural smoothing loss to generate smooth stroke-lines and

*Corresponding author.

[1]Written and illustrated by Tite Kubo, 2001-2016.

Figure 1: *Left:* the combination of manga faces we generated and body components we collected from the target manga work Bleach, which shows a unified style with a strong attraction. *Right:* the input face photos, the output manga faces, and some samples of the target manga work. Our method can effectively transform an input facial photo into a high-quality manga face with preserving both the facial similarity and the target manga style.

less messy pixels.

- We construct a new dataset called MangaGAN-BL (including manga facial features, landmarks, bodies, etc.), collected from a popular manga work named *Bleach*. And each training sample is manually processed by cropping, angle-correction, and repairing of disturbing elements (e.g, hair covering, shadows). The MangeGAN-BL dataset is available for academic use.

2 Related Work

Recent literature suggests two main directions with the ability to generate manga-like results: neural style transfer, and GAN-based cross-domain translation.

2.1 Neural Style Transfer

The goal of neural style transfer (NST) is to transfer the style from an art image to another content target image. Inspired by the progress of CNN, Gatys et al. (Gatys, Ecker, and Bethge 2016) propose the pioneering NST work by utilizing CNN's power of extracting abstract features, and the style capture ability of Gram matrices (Gatys, Ecker, and Bethge 2015). Then, Li and Wand (Li and Wand 2016a) use the Markov Random Field (MRF) to encode styles, and present an MRF-based method (CNNMRF) for image stylization. Afterward, various follow-up works have been presented to improve their performances on visual quality (Liao et al. 2017; Zhang, Zhang, and Cai 2018; Gu et al. 2018; Jing et al. 2018; Men et al. 2018), generating speed (Chen et al. 2017b; Johnson, Alahi, and Li 2016; Huang and Belongie 2017; Shen, Yan, and Zeng 2018; Li et al. 2017), and multimedia extension (Chen et al. 2017a; Huang et al. 2017; Xu et al. 2019; Chen et al. 2018).

Although these methods work well on translating images into some typical artistic styles, e.g., oil painting, watercolor, they are not good at producing black-and-white manga with exaggerated geometry and discrete stroke lines, since they tend to translate textures and colors features of a target style and preserve the structure of the content image.

2.2 GAN-Based Cross-Domain Translation

Many GAN-based cross-domain translation methods work well on image stylization, whose goal is to learn a mapping from a source domain to a stylized domain. There are a series of works based on GAN (Goodfellow et al. 2014) presented and applied for image stylization. Pix2Pix (Isola et al. 2017) first presents a unified framework for image-to-image translation based on conditional GANs (Mirza and Osindero 2014). BicycleGAN (Zhu et al. 2017b) extends it to multi-modal translation. Some methods including CycleGAN (Zhu et al. 2017a), DualGAN (Yi et al. 2017), DiscoGAN (Kim et al. 2017), UNIT (Liu, Breuel, and Kautz 2017), DTN (Taigman, Polyak, and Wolf 2016) etc. are presented for unpaired one-to-one translation. MNUIT (Huang et al. 2018b), startGAN (Choi et al. 2018) etc. are presented for unpaired many-to-many translation.

These methods succeed in translation tasks that are mainly characterized by color or texture changes only (e.g., summer to winter, apples to oranges). For photo-to-manga translation, they fail to capture the correspondences between two domains due to the abstract structure, colorless appearance, and geometric deformation of manga drawing.

Besides the above two main directions, there are also some works specially designed for creating artistic facial images. They employ techniques of Non-photorealistic rendering (NPR), data-driven synthesizing, computer graphics, etc., and have achieved much progress in many typical art forms, e.g., caricature and cartoon (Cao, Liao, and Yuan 2018; Zhang et al. 2016; Li et al. 2018), portrait and sketching (Yi et al. 2019; Rosin and Lai 2015; Wang, Sindagi, and Patel 2018; Wang et al. 2017). However, none of them involve the generation of manga face.

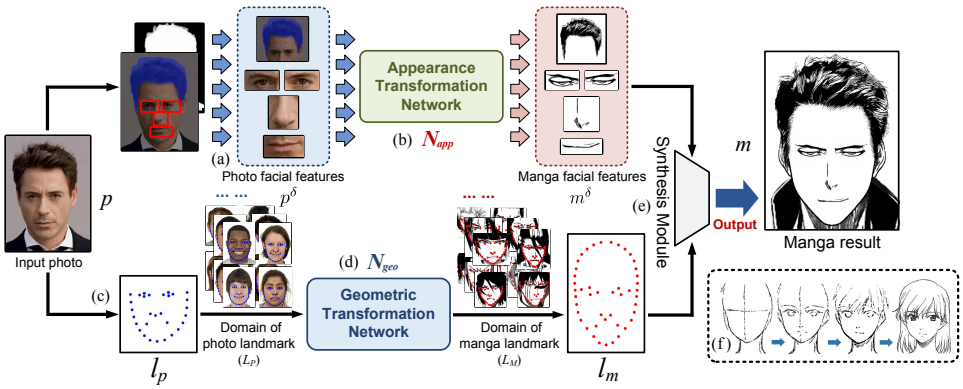

Figure 2: Overall pipeline of MangaGAN. MangaGAN consists of two branches: one branch learns the geometric mapping by a Geometric Transformation Network (GTN), and the other branch learns the appearance mapping by an Appearance Transformation Network (ATN). On the end, a Synthesis Module is designed to fuse them and produce the manga face.

3 Method

3.1 Overview

Let P and M indicate the face photo domain and the manga domain respectively, where no pairing exists between them. Given an input photo $p \in P$, MangaGAN learns a mapping Ψ: $P \rightarrow M$ that can transfer p to a sample $m = \Psi(p)$, $m \in M$, while endowing m with manga style and facial similarity.

We observe that the experienced manga artist first outline the exaggerated face and locate the geometric distributions of facial features, and then do the fine-drawing [Fig. 2(f)]. Accordingly, our MangaGAN consists of two branches: one branch learns a geometric mapping Ψ_{geo} by a Geometric Transformation Network (GTN) N_{geo} to translate the facial geometry from P to M [Fig. 2(d)], and the other branch learns an appearance mapping Ψ_{app} by an Appearance Transformation Network (ATN) N_{app} [Fig. 2(b)] to produce components of all facial features. Finally, a Synthesis Module is designed to fuse facial geometry and all components to output the manga face $m \in M$ [Fig. 2(e)]. Then, we will detail the ATN, the GTN, and the Synthesis Module in Section 3.2, Section 3.3, and Section 3.4 respectively.

3.2 Appearance Transformation Network

ATN N_{app} is a network consisting of a set of local GANs, $N_{app} = \{N^{eye}, N^{nose}, N^{mouth}, N^{hair}\}$, where each local GAN is trained for translating a target facial region, from the input $p \in P$ to the output $m \in M$.

Translating eyes and mouths. Eyes and mouths are the critical components of manga faces but are the hardest parts to translate, since they are most noticed, error sensitive, and vary with different facial expressions. For N^{eye} and N^{mouth}, for better mapping the unpaired data, we couple it with a reverse mapping, inspired by the network architecture of CycleGAN (Zhu et al. 2017a). Accordingly, the baseline architecture of $N^\delta (\delta \in \{\text{eye, mouth}\})$ includes the forward / backward generator G_M^δ / G_P^δ and the corresponding discriminator D_P^δ / D_M^δ. G_M^δ learns the mapping Ψ_{app}^δ: $p^\delta \rightarrow \widehat{m}^\delta$, and G_P^δ learns the reverse mapping $\Psi_{app}^{\delta'}$: $m^\delta \rightarrow \widehat{p}^\delta$, where $\widehat{m}_i^\delta$ and $\widehat{p}_i^\delta$ are the generated fake samples; the discriminator D_P^δ / D_M^δ learn to distinguish real samples p^δ / m^δ and fake samples $\widehat{p}^\delta$ / $\widehat{m}^\delta$. Our generators G_P^δ, G_M^δ use the Resnet 6 blocks (He et al. 2016), and D_P^δ, D_M^δ use the Markovian discriminator of 70×70 patchGANs (Isola et al. 2017; Li and Wand 2016b; Ledig et al. 2017).

We adopt the stable least-squares losses (Mao et al. 2017) instead of negative log-likelihood objective (Goodfellow et al. 2014) as our adversarial losses L_{adv}, defined as

$$\begin{aligned}\mathcal{L}_{adv}^\delta(G_M^\delta, D_M^\delta) = \mathbb{E}_{m^\delta \sim M^\delta}[(D_M^\delta(m^\delta) - 1)^2] \\ + \mathbb{E}_{p^\delta \sim P^\delta}[D_M^\delta(G_M^\delta(p^\delta))^2]\end{aligned}, \quad (1)$$

while $\mathcal{L}_{adv}^\delta(G_P^\delta, D_P^\delta)$ is defined in a similar manner.

$\mathcal{L}_{cyc}$ is the cycle-consistency loss (Zhu et al. 2017a) that is used to constrain the mapping solution between the input and the output domain, defined as

$$\begin{aligned}\mathcal{L}_{cyc}^\delta(G_P^\delta, G_M^\delta) = \mathbb{E}_{p^\delta \sim P^\delta}[\|G_P^\delta(G_M^\delta(p^\delta)) - p^\delta\|_1] \\ + \mathbb{E}_{m^\delta \sim M^\delta}[\|G_M^\delta(G_P^\delta(m^\delta)) - m^\delta\|_1]\end{aligned}. \quad (2)$$

However, we find that the baseline architectures of N^{eye} and N^{mouth} with $\mathcal{L}_{adv}$ and $\mathcal{L}_{cyc}$ still fail to preserve the similarity between two domains. Specifically, for regions of eyes and mouths, it always produces messy results since networks are unable to match colored photos and discrete black lines of mangas. Therefore, we further make the following three improvements to optimize their performances.

First, we design a *Similarity Preserving (SP)* module with an SP loss $\mathcal{L}_{SP}$ to enhance the similarity. Second, we train an encoder E^{eye} that can extract the main backbone of p^{eye} to binary results, as the input of N^{eye}, and an encoder E^{mouth} that encodes p^{mouth} to binary edge-lines, used to guide the shape of manga mouth. Third, a structural smoothing loss $\mathcal{L}_{SS}$ is designed for encouraging networks to produce manga with smooth stroke-lines, defined as $\mathcal{L}_{SS}(G_P^\delta, G_M^\delta) = \frac{1}{\sqrt{2\pi}\sigma}\left[\sum_{j\in\{1,2,\ldots,N\}} \exp\left(\frac{-(G_P^\delta(m^\delta)_j - \mu)^2}{2\sigma^2}\right) + \sum_{k\in\{1,2,\ldots,N\}} \exp\left(-\frac{(G_M^\delta(p^\delta)_k - \mu)^2}{2\sigma^2}\right)\right]$, where $\mathcal{L}_{SS}$ based uses a Gaussian model with $\mu = \frac{255}{2}$, j and k are the indexes of pixels. The

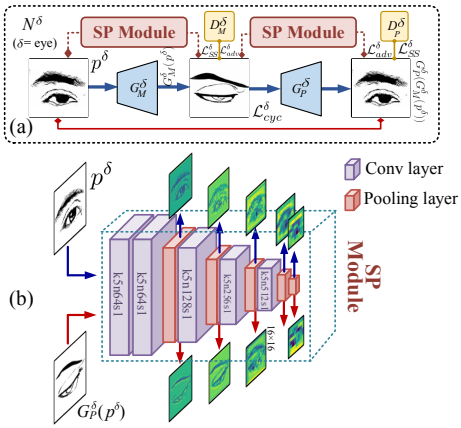

Figure 3: (a) We append two SP modules on both forward and backward mappings. (b) SP module extracts feature maps with different resolutions and measures the similarities between two inputs in different latent spaces.

underlying idea is that producing unnecessary gray areas will distract and mess the manga results since manga mainly consists of black and white stroke lines. Thus we give a pixel smaller loss when its gray value closer to black or white, to smooth the gradient edges of black stroke lines and produce clean results.

Similarity-Preserving Module. The main idea of SP module is to keep the similarity between two images at a lower resolution, so that they will have similar spatial distributions and different pixel details when they are up-sampled to a higher resolution. As shown in Fig. 3(a), we append two SP modules on both forward and backward mappings of N^δ. SP module is inspired by VGG-Net (Simonyan and Zisserman 2014) and pre-trained network ϕ on MSCOCO (Lin et al. 2014), which is we designed to extract feature maps in different latent spaces and resolutions. The architecture of ϕ as shown in Fig. 3(b), it only uses few convolutional layers since we consider the correspondences of encoded features are relatively clear. For the forward mapping Ψ^δ_{app}: $\widehat{m}^\delta = G^\delta_M(p^\delta)$, we input p^δ and $G^\delta_M(p^\delta)$ to SP module, and optimize G^δ_M by minimizing the loss functions $\mathcal{L}_{SP}(G^\delta_M, p^\delta)$ defined as

$$\mathcal{L}_{SP}(G^\delta_M, p^\delta) = \sum_{i\in\phi} \lambda_i \mathcal{L}^{\phi,i}_{feat}\left[f^\phi_i(p^\delta), f^\phi_i(G^\delta_M(p^\delta))\right] + \lambda_I \mathcal{L}^I_{pixel}\left[p^\delta, G^\delta_M(p^\delta)\right], \quad (3)$$

where λ_i, λ_I controls the relative importance of each objective, $\mathcal{L}^I_{pixel}$ and $\mathcal{L}^{\phi,i}_{feat}$ are used to keep the similarity on pixel-wise and different feature-wise respectively. $\mathcal{L}^I_{pixel}$ and $\mathcal{L}^{\phi,i}_{feat}$ defined as

$$\mathcal{L}^{\phi,i}_{feat}[f^\phi_i(p^\delta), f^\phi_i(G^\delta_M(p^\delta))] = \left\|f^\phi_i(p^\delta) - f^\phi_i(G^\delta_M(p^\delta))\right\|^2_2, \quad \mathcal{L}^I_{pixel}[p^\delta, G^\delta_M(p^\delta)] = \left\|p^\delta - G^\delta_M(p^\delta)\right\|^2_2 \quad (4)$$

where $f^\phi_i(x)$ is a feature map extracted from i-th layer of network ϕ when x as the input. Note that we only extract feature maps after pooling layers.

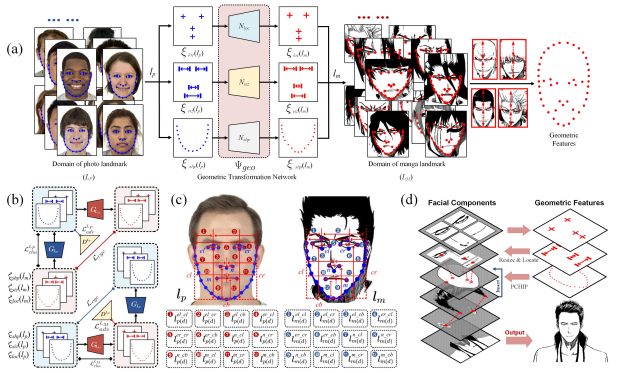

Figure 4: (a) Pipeline of GTN. (b) Architectures of N_{loc}, N_{siz}, and N_{sha}. (c) Definitions of relative locations in $\xi_{loc}(l_p), \xi_{loc}(l_m)$. (d) In synthesis module, we generate manga by fusing facial components and their geometric features.

Combining Eq.(1)-(4), the full objective for learning the appearance mappings of $N^\delta(\delta\in\{\text{eye, mouth}\})$ is:

$$\mathcal{L}^\delta_{app} = \mathcal{L}^\delta_{adv}(G^\delta_M, D^\delta_M) + \mathcal{L}^\delta_{adv}(G^\delta_P, D^\delta_P) + \alpha_1\mathcal{L}^\delta_{cyc}(G^\delta_P, G^\delta_M) + \alpha_2\mathcal{L}^\delta_{SP}(G^\delta_M, p^\delta) + \alpha_3\mathcal{L}^\delta_{SP}(G^\delta_P, m^\delta) + \alpha_4\mathcal{L}_{SS}(G^\delta_M, G^\delta_P), \quad (5)$$

where α_1 to α_4 used to balance the multiple objectives.

Translating nose and hair. Noses are insignificant to manga faces since almost all characters have a similar nose in the target manga style. Therefore, N^{nose} adopts a generating method instead of a translating one, which follows the architecture of progressive growing GANs (Karras et al. 2017) that can produce a large number of high-quality results similar to training data. We first train a variational autoencoder (Kingma and Welling 2013) to encode the nose region of the input photo into a feature vector, then make the vector as a seed to generate a default manga nose, and we also allow users to change it according to their preferences.

N^{hair} employs a pre-trained generator of APDdrawingGAN (Yi et al. 2019) that can produce binary portrait hair with the style similar to manga. In addition, the coordinates of generated portraits can accurately correspond to the input photos. We first extract the rough hair region by a hair segmentation method (Muhammad et al. 2018) with a fine-tune of expanding the segmented area, and then remove the extra background area by a portrait segmentation method (Shen et al. 2016).

3.3 Geometric Transformation Network

The goal of GTN is to translate the geometric features of faces from the photo domain to the manga domain, where we represent these features with facial landmarks. Let L_P and L_M express the domain of landmarks corresponding to photo and manga. GTN learns a geometric mapping Ψ_{geo}: $l_p \in L_P \rightarrow l_m \in L_M$, where l_m must be similar to l_p and follow manga's geometric style. For training data, each landmark l_p can be extracted by an existing face landmark de-

tector (King 2009), and 106 facial landmarks of manga data l_m are manually marked by us.

When translating facial landmarks, an issue is that the collocation mode of facial features constrains the variety of results. For example, people with the same face shape may have different sizes or locations of eyes, nose, or mouth. However, GAN may generate them in a fixed or similar collocation mode when it is trained by the landmarks of global faces. Accordingly, as shown in Fig. 4(a), we divide the geometric features into three attributions (face shape, facial features' locations and sizes) and employ three sub-GANs N_{sha}, N_{loc}, N_{siz} to translate them respectively.

Input of sub-GANs. For N_{loc}, we employ relative locations instead of absolute coordinates, since directly generating coordinates may incur few facial features beyond the face profile. As shown in Fig. 4(c), for l_p, relative locations are represented as a vector $\xi_{loc}(l_p)$. $\xi_{loc}(l_p) = \{l_p^{el}, l_p^{er}, l_p^{n}, l_p^{m}\}$ and $\xi_{loc}(l_m)$ is represented similarly, where l_p^{el}, l_p^{er}, l_p^{n}, l_p^{m} represent regions of left eye, right eye, nose, and mouth respectively. Take l_p^n as an example, its relative location is represented as three scalars $l_{p(d)}^{n\_cl}$, $l_{p(d)}^{n\_cr}$, $l_{p(d)}^{n\_cb}$, corresponding to distances of nose's center to cheek's left edge, right edge, and bottom edge respectively, and l_p^{el}, l_p^{er}, l_p^{m} are defined similarly. N_{siz} only learns the mapping of facial features' widths, since the length-width ratio of the generated manga facial regions are fixed. Then, the size features of l_p is represent as $\xi_{siz}(l_p) = \{l_{p(w)}^{el}, l_{p(w)}^{er}, l_{p(w)}^{n}, l_{p(w)}^{m}\}$, where $l_{p(w)}^{el}$, $l_{p(w)}^{er}$, $l_{p(w)}^{n}$, $l_{p(w)}^{m}$ represent the width of left eye, right eye, nose, and mouth respectively. N_{sha} learns to translate the face shape, where the face shape is represented as the landmark of cheek region containing 17 points.

Network architecture. As shown in Fig. 4(b), N_{loc}, N_{siz}, and N_{shp} roughly follow the structure of CycleGAN (Zhu et al. 2017a) with adversarial loss $\mathcal{L}_{adv}$ as eq(1) and cycle loss $\mathcal{L}_{cyc}$ as eq(2). Moreover, we replace all convolutional layer in generators with the fully connected layers, and add the characteristic loss $\mathcal{L}_{cha}$ (Cao, Liao, and Yuan 2018) that leverages the differences between a face and the mean face to measure the distinctive features after exaggeration. Let $\mathcal{L}_{cha}^{L_M}(G_{L_M})$ indicates the characteristic loss on the forward mapping, defined as

$$\mathcal{L}_{cha}^{L_M}(G_{L_M}) = \mathbb{E}_{\xi_*(l_p)\sim\xi_*(L_P)}\big\{1 - \cos[\xi_*(l_p) - \xi_*(\overline{L_P}), G_{L_M}(\xi_*(l_p)) - \xi_*(\overline{L_M})]\big\}, \quad (6)$$

where $\xi_*(\overline{L_P})$ or $\xi_*(\overline{L_M})$ denotes the averages of vector $\xi_*(L_P)$ or $\xi_*(L_M)$ whose format defined by network N_*, $*\in\{loc, siz, shp\}$, while the reverse loss $\mathcal{L}_{cha}^{L_P}$ is defined similarly. We let $\underset{loc}{\mathcal{L}}$ denotes the loss of N_{loc}, and losses of N_{siz} and N_{sha} are represented in a similar manner. The objective function $\mathcal{L}_{geo}$ to optimize GTN is $\mathcal{L}_{geo} = \underset{loc}{\mathcal{L}}_{adv}^{L_P} + \underset{loc}{\mathcal{L}}_{adv}^{L_M} + \beta_1\mathcal{L}_{cyc} + \beta_2(\underset{loc}{\mathcal{L}}_{cha}^{L_P} + \underset{loc}{\mathcal{L}}_{cha}^{L_M}) + \underset{siz}{\mathcal{L}}_{adv}^{L_P} + \underset{siz}{\mathcal{L}}_{adv}^{L_M} + \beta_3\mathcal{L}_{cyc} + \beta_4(\underset{siz}{\mathcal{L}}_{cha}^{L_P} + \underset{siz}{\mathcal{L}}_{cha}^{L_M}) + \underset{sha}{\mathcal{L}}_{adv}^{L_P} + \underset{sha}{\mathcal{L}}_{adv}^{L_M} + \beta_5\underset{sha}{\mathcal{L}}_{cyc} + \beta_6(\underset{sha}{\mathcal{L}}_{cha}^{L_P} + \underset{shp}{\mathcal{L}}_{cha}^{L_M})$, where β_1 to β_6 used to balance the multiple objectives.

Finally, as shown in Fig. 4(b), according to the pre-defined proportion of cheek and forehead, we produce the geometric features of the whole manga face.

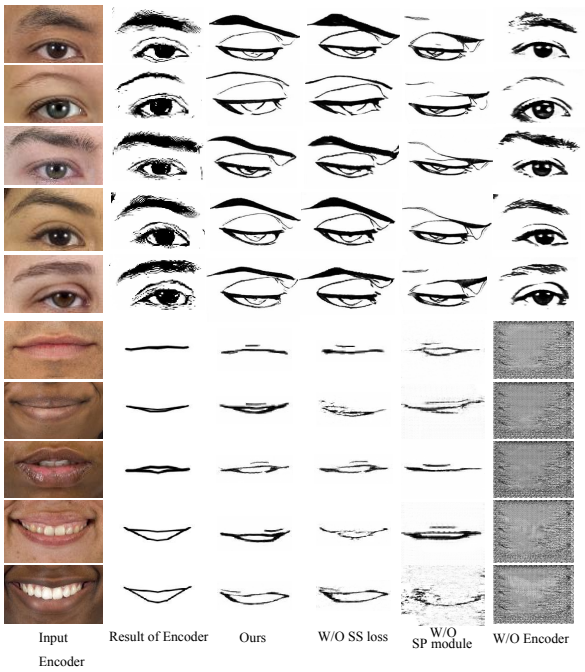

Figure 5: Ablation experiments on our improvements.

3.4 Synthesis Module

The goal of this module is to synthesis an attractive manga face by combining facial components and their geometric features. As mentioned above, facial components are generated by ATN in Section 3.2, and the geometric features are generated by GTN in Section 3.3.

The pipeline of fusing components is shown in Fig. 4(d). First, we resize and locate facial components following the geometric features. Second, the face shape is drawn by the fitting curve of generated landmarks, based on the method of *Piecewise Cubic Hermite Interpolating Polynomial (PCHIP)* (Fritsch and Carlson 1980), where PCHIP can obtain a smooth curve and effectively preserving the face shape. Then, for ear regions, we provide 10 components of manga ears instead of generating them, since they are stereotyped and unimportant for facial expression. Moreover, we collect 8 manga bodies in our dataset, 5 for male, and 3 for female, that mainly used for decorating faces. In the end, we output a default manga result, and provide a toolkit that allows users to fast fine-tune the size and location of each manga component, and to switch components that insignificant for facial expression (i.e., noses, ears, and bodies) following their preferences.

4 Experiment

In the following experiments, we first introduce our dataset and training details in Section 4.1 and then evaluate the effectiveness of our improvements in Section 4.2. Finally, in Section 4.3, we compare our MangaGAN with other state-of-the-art works. We implemented MangaGAN in PyTorch

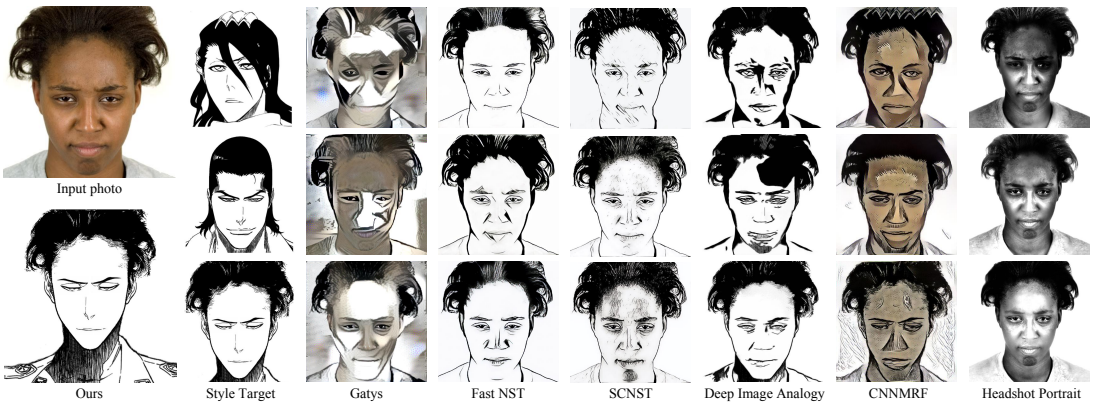

Figure 6: Comparison results with NST methods, containing Gatys et al., Fast NST, SCNST, Deep Image Analogy, CNNMRF, and Headshot Portrait.

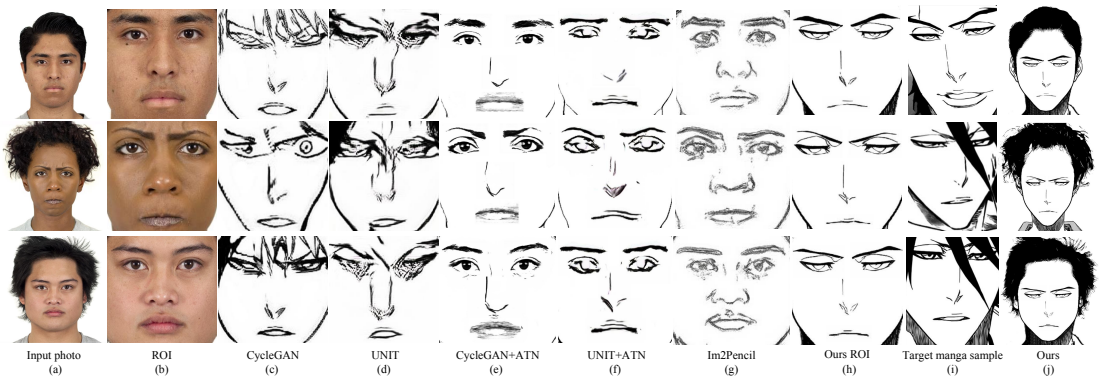

Figure 7: Comparison results with cross-domain translation methods. (a) Input photo. (b) ROI of (a). (c)-(h) Results of CycleGAN, UNIT, Im2Pencil, APDrawingGAN, and our method, respectively. (i) Samples of target manga work.

(Paszke et al. 2017) and all experiments are performed on a computer with an NVIDIA Tesla V100 GPU.

4.1 Training

Dataset. The datasets we used in experiments are divided into three parts, i.e., the manga dataset $\mathcal{D}_m$, the photo dataset $\mathcal{D}_p$, and the portrait dataset $\mathcal{D}_b$. $\mathcal{D}_m$, named MangaGAN-BL, is a new dataset we collected from a popular manga work *Bleach*. It contains manga facial features of 448 eyes, 109 noses, 179 mouths, and 106 frontal view of manga faces whose landmarks have been marked manually. Each sample of $\mathcal{D}_m$ is normalized to 256×256 and is optimized by cropping, angle-correction, and repairing of disturbing elements (e.g, hair covering, glasses, shadows); $\mathcal{D}_p$ contains 1197 front view of face photos collected from CFD (Ma, Correll, and Wittenbrink 2015), and $\mathcal{D}_b$ contains 1197 black-and-white portraits generated by APDrawingGAN (Yi et al. 2019) when $\mathcal{D}_p$ as input.

Training details. For training MangaGAN, each training data is converted to grayscale with 1 channel, and each landmark of manga face is pre-processed by symmetric processing. For all experiments, we set α_1=10, $\alpha_{\{2,3\}}$=5, α_4=1 in Eq.(5); $\beta_{\{1,3,5\}}$=10, $\beta_{\{2,4,6\}}$=1, and parameters of $\mathcal{L}_{SP}$ in Eq.(3) are fixed at λ_I=1, λ_{pool5}=1, λ_i=0, $i \in \{pool1, pool2, pool3, pool4\}$ with the output resolution of 256×256. Moreover, we employ the Adam solver (Kingma and Ba 2014) with a batch size of 5. All networks use the learning rate of 0.0002 for the first 100 epochs, where the rate is linearly decayed to 0 over the next 100 epochs.

4.2 Ablation Experiment

In Section 3.2, E^{eye} is a conditional GAN model basically following (Isola et al. 2017), and is pretrained by paired eye regions of photos from dataset $\mathcal{D}_p$ and their binary result from dataset $\mathcal{D}_b$; E^{mouth} includes a landmark detector (King 2009) and a pre-processed program that smoothly connects landmarks of mouth to the black edge-lines to guide the shape of a manga mouth. With the help of E^{eye} and E^{mouth}, our method can effectively preserve the shape of eyebrows (red arrows), eyes, and mouths, and further abstract them into manga style. Without E^{eye} or E^{mouth}, the network cannot capture the correspondences or generated messy results, as shown in the 6^{th} columns in Fig. 5. SP module is essential to keep the similarities between photos and mangas. As shown in the 5^{th} columns of Fig. 5, without the SP module, neither the manga style nor the similarity

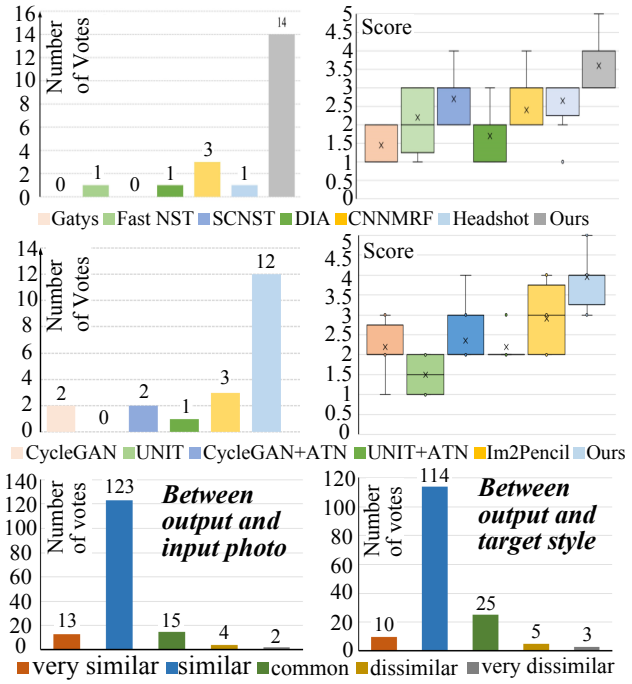

Figure 8: Comparison results with NST methods (upper) and domains translation methods (middle), and results of the online user study on a professional manga forum (bottom).

between input and output can be well preserved. Structural Smoothing (SS) loss is also a key to produce mangas with clean appearances and smooth stroke-lines. As shown in the 4^{th} columns in Fig. 5, when training with SS loss, the structure of black stroke lines are effectively smoothed and the gray messy pixels are reduced as well.

4.3 Comparison with State-of-the-Art Methods

We compare MangaGAN with nine state-of-the-art methods that have potentials to produce manga-like results.

NST method. The first class is NST methods, containing Gatys (Gatys, Ecker, and Bethge 2016), Fast NST (Johnson, Alahi, and Li 2016), SCNST (Jing et al. 2018), Deep Image Analogy (Liao et al. 2017), CNNMRF (Li and Wand 2016a), and Headshot Portrait (Shih et al. 2014). For fair comparison, as shown in Fig. 6, we employ three different manga faces (one of which is our result) as the style targets to stylize each input photo respectively.

Domains translation method. The second class is cross-domain translation methods, containing CycleGAN (Zhu et al. 2017a), UNIT (Liu, Breuel, and Kautz 2017), and Im2pencil (Li et al. 2019). For fair comparison, we train CycleGAN and UNIT to translate the whole face region, and translate each facial feature, respectively. For the whole face region translation [Fig. 7(b)-(d)], we only train the ROI to make these methods easier to find the correspondences between photo and manga, where the photo domain trained by 1197 frontal facial photos' ROIs in $\mathcal{D}_b$, and the manga domain trained by 83 frontal manga faces' ROIs in $\mathcal{D}_m$. For each facial feature translation [Fig. 7(e)(f)], we append CycleGAN and UNIT on ATN structure, and train each facial region by the same data as we use. CycleGAN

Performance \ Method		NST	CDT	Ours
Preserve user identity		✓	✓	✓
Preserve manga style	*clean stroke-line*	✗	✗	✓
	abstract facial feat.	✗	✗	✓
	exaggeration	✗	✗	✓

Table 1: Comparison of preserving user ID and manga style.

Discussion. In Table 1, we summarize the observed comparison results. Specifically, the referenced methods work better in preserving user IDs, since they more focus on translating textures and colors and preserving input photos' structures. In fact, the excessive similarity between photos and mangas will compromise the manga style, where the reason is that the manga characters are fictitious, abstracted, and much unlike real people, e.g., manga faces are often designed to own optimum proportions, and facial features are abstracted to several black lines [Fig. 7(i)]. Therefore, our method outperforms the referenced methods on preserving manga style, i.e., producing manga face with clean stroke-line, abstract facial features, and geometric exaggeration.

4.4 User Study

Bellow, we conduct two main user studies to subjectively evaluate the performances of our method.

Comparison of visual effects. We randomly select 10 face photos from $\mathcal{D}_p$ to produce manga faces, and invite 20 volunteers to finish two tasks: the first task is to score 1-5 for each method's result (higher score means more attractive), and another task is to vote for the most attractive results. The study results in Fig. 8 show that our method gets the highest score and the most vote number.

Preserve user ID and manga style. We design an online questionnaire and open it to a professional manga forum, and then ask the experienced manga readers to evaluate the preservation of facial similarity and manga style. In a two-week period, 157 participants attended this study, and the results in Fig. 8 show that 86.62 % and 78.98 % participants believe our method preserves the facial similarity and the target manga style respectively, which means our method achieves good performances on both two aspects.

5 Conclusion

In this paper, we propose the first GAN-based method for unpaired photo-to-manga translation, called MangaGAN. It is inspired by the prior-knowledge of drawing manga, and can translate a frontal face photo into manga domain with preserving manga style. Extensive experiments show that MangaGAN can produce high-quality manga faces and outperforms other state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 61772060, 61976012).

References

Cao, K.; Liao, J.; and Yuan, L. 2018. CariGANs: unpaired photo-to-caricature translation. In *SIGGRAPH Asia 2018 Technical Papers*, 244. ACM.

Chen, D.; Liao, J.; Yuan, L.; Yu, N.; and Hua, G. 2017a. Coherent Online Video Style Transfer. In *Proc. Intl. Conf. Computer Vis.*

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2017b. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1897–1906.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2018. Stereoscopic neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6654–6663.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; 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*, 8789–8797.

Fritsch, F. N.; and Carlson, R. E. 1980. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis* 17(2): 238–246.

Gatys, L.; Ecker, A. S.; and Bethge, M. 2015. Texture synthesis using convolutional neural networks. In *Advances in neural information processing systems*, 262–270.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2414–2423.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gu, S.; Chen, C.; Liao, J.; and Yuan, L. 2018. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8222–8231.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, H.; Wang, H.; Luo, W.; Ma, L.; Jiang, W.; Zhu, X.; Li, Z.; and Liu, W. 2017. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 783–791.

Huang, J.; Tan, M.; Yan, Y.; Qing, C.; Wu, Q.; and Yu, Z. 2018a. Cartoon-to-Photo Facial Translation with Generative Adversarial Networks. In *Asian Conference on Machine Learning*, 566–581.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018b. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976. IEEE.

Jing, Y.; Liu, Y.; Yang, Y.; Feng, Z.; Yu, Y.; Tao, D.; and Song, M. 2018. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 238–254.

Johnson, J.; Alahi, A.; and Li, F.-F. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proc. Eur. Conf. Comput. Vis.*, 694–711.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* .

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1857–1865. JMLR. org.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10(Jul): 1755–1758.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, C.; and Wand, M. 2016a. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2479–2486.

Li, C.; and Wand, M. 2016b. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, 702–716. Springer.

Li, W.; Xiong, W.; Liao, H.; Huo, J.; Gao, Y.; and Luo, J. 2018. CariGAN: Caricature Generation through Weakly Paired Adversarial Learning. *arXiv preprint arXiv:1811.00445* .

Li, Y.; Fang, C.; Hertzmann, A.; Shechtman, E.; and Yang, M.-H. 2019. Im2pencil: Controllable pencil illustration from photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1525–1534.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, 386–396.

Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; and Kang, S. B. 2017. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)* 36(4): 120.

Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollr, P.; and Zitnick, C. L. 2014. Microsoft COCO: common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, volume 8693, 740–755.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 700–708.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47(4): 1122–1135.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Men, Y.; Lian, Z.; Tang, Y.; and Xiao, J. 2018. A common framework for interactive texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6353–6362.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Muhammad, U. R.; Svanera, M.; Leonardi, R.; and Benini, S. 2018. Hair detection, segmentation, and hairstyle classification in the wild. *Image and Vision Computing* .

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *In NeurIPS Workshop*.

Rosin, P. L.; and Lai, Y.-K. 2015. Non-photorealistic rendering of portraits. In *Proceedings of the workshop on Computational Aesthetics*, 159–170. Eurographics Association.

Selim, A.; Elgharib, M.; and Doyle, L. 2016. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)* 35(4): 129.

Shen, F.; Yan, S.; and Zeng, G. 2018. Neural style transfer via meta networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8061–8069.

Shen, X.; Hertzmann, A.; Jia, J.; Paris, S.; Price, B.; Shechtman, E.; and Sachs, I. 2016. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, 93–102. Wiley Online Library.

Shi, Y.; Deb, D.; and Jain, A. K. 2019. WarpGAN: Automatic Caricature Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10762–10771.

Shih, Y.; Paris, S.; Barnes, C.; Freeman, W. T.; and Durand, F. 2014. Style transfer for headshot portraits. *ACM Transactions on Graphics (TOG)* 33(4): 148.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for Large-Scale image recognition. *Comput. Sci.* .

Taigman, Y.; Polyak, A.; and Wolf, L. 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200* .

Wang, L.; Sindagi, V.; and Patel, V. 2018. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 83–90. IEEE.

Wang, N.; Gao, X.; Sun, L.; and Li, J. 2017. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing* 26(3): 1264–1274.

Xu, M.; Su, H.; Li, Y.; Li, X.; Liao, J.; Niu, J.; Lv, P.; and Zhou, B. 2019. Stylized Aesthetic QR Code. *IEEE Transactions on Multimedia* .

Yi, R.; Liu, Y.-J.; Lai, Y.-K.; and Rosin, P. L. 2019. APDrawingGAN: Generating Artistic Portrait Drawings from Face Photos with Hierarchical GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10743–10752.

Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, 2849–2857.

Zhang, S.; Gao, X.; Wang, N.; Li, J.; and Zhang, M. 2015. Face sketch synthesis via sparse representation-based greedy search. *IEEE transactions on image processing* 24(8): 2466–2477.

Zhang, Y.; Dong, W.; Ma, C.; Mei, X.; Li, K.; Huang, F.; Hu, B.-G.; and Deussen, O. 2016. Data-driven synthesis of cartoon faces using different styles. *IEEE Transactions on image processing* 26(1): 464–478.

Zhang, Y.; Zhang, Y.; and Cai, W. 2018. Separating style and content for generalized style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8447–8455.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 465–476.