Unpaired Portrait Drawing Generation via Asymmetric Cycle Mapping

Ran Yi, Yong-Jin Liu*
CS Dept, BNRist
Tsinghua University, China
{yr16,liuyongjin}@tsinghua.edu.cn

Yu-Kun Lai, Paul L. Rosin
School of Computer Science and Informatics
Cardiff University, UK
{LaiY4,RosinPL}@cardiff.ac.uk

Abstract

Portrait drawing is a common form of art with high abstraction and expressiveness. Due to its unique characteristics, existing methods achieve decent results only with paired training data, which is costly and time-consuming to obtain. In this paper, we address the problem of automatic transfer from face photos to portrait drawings with unpaired training data. We observe that due to the significant imbalance of information richness between photos and drawings, existing unpaired transfer methods such as CycleGAN tend to embed invisible reconstruction information indiscriminately in the whole drawings, leading to important facial features partially missing in drawings. To address this problem, we propose a novel asymmetric cycle mapping that enforces the reconstruction information to be visible (by a truncation loss) and only embedded in selective facial regions (by a relaxed forward cycle-consistency loss). Along with localized discriminators for the eyes, nose and lips, our method well preserves all important facial features in the generated portrait drawings. By introducing a style classifier and taking the style vector into account, our method can learn to generate portrait drawings in multiple styles using a single network. Extensive experiments show that our model outperforms state-of-the-art methods.

1. Introduction

Portrait drawing is a unique style of art which is highly abstract and expressive. However, drawing a delicate portrait drawing is time consuming and needs to be carried out by skilled artists. Therefore, automatic generation of portrait drawings is very desirable.

Image style transfer has been a longstanding topic in computer vision. In recent years, inspired by the effectiveness of deep learning, Gatys et al. [4] introduced convolutional neural networks (CNNs) to transfer style from a style image to a content image, and opened up the field of neural style transfer. Subsequently, generative adversarial networks (GANs) have achieved much success in solving image style transfer problems [10, 25]. However, existing methods are mainly applied to cluttered styles (e.g., oil painting style) where a stylized image is full of fragmented brush strokes and the requirement for the quality of each individual element is low.

Artistic portrait line drawings (APDrawings) are completely different from the previously tackled painting styles. Generating them is very challenging because the style is highly abstract: it only contains a sparse set of graphical elements, is line-stroke-based, disables shading, and has high semantic constraints. Therefore, previous texture-based style transfer methods and general image-to-image translation methods fail to generate good results on the APDrawing style (Fig. 1). To the best of our knowledge, APDrawingGAN [20] is the only method that explicitly deals with APDrawing by using a hierarchical structure and a distance transform loss. However, this method requires *paired* training data that is costly to obtain. Due to the limited availability of paired data, this method cannot adapt well to face photos with unconstrained lighting in the wild.

Compared to paired training data, APDrawing generation learned from *unpaired* data is much more challenging. Previous methods for unpaired image-to-image translation [25, 21] use a cycle structure to regularize training. Although cycle consistency loss enables learning from unpaired data, we observe that when applying them to face photo to APDrawing translation, due to significant imbalance of information richness in these two data types, these methods tend to embed invisible reconstruction information indiscriminately in the whole APDrawing, causing a deterioration in the quality of the generated APDrawings, such as important facial features partially missing (Figs. 1(f-g)).

In this paper, we propose an *asymmetric cycle structure* to tolerate certain reconstruction quality issues. We argue that the network does not need to reconstruct an *accurate* face photo from a generated APDrawing due to information imbalance. Accordingly, we introduce a relaxed cycle consistency loss between the reconstructed face photo

*Corresponding author

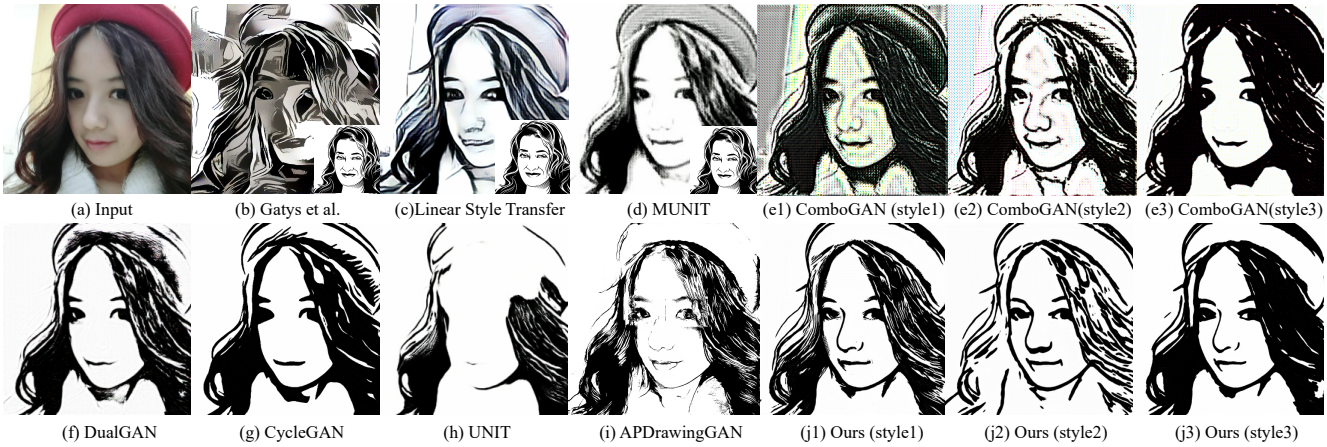


Figure 1. Comparison with state-of-the-art methods: (a) input face photo; (b)-(c) style transfer methods: Gatys [4] and Linear Style Transfer [14]; (f)-(h) single-modal image-to-image translation methods: DualGAN [21], CycleGAN [25], UNIT [15]; (d)-(e) multi-modal image-to-image translation methods MUNIT [9] and ComboGAN [1]; (i) a portrait generation method APDrawingGAN [20]; (j) our method. Note that APDrawingGAN requires *paired* data for training, so unlike other work, it is trained using the paired APDrawing dataset. Due to this essential difference, we do not compare with this method in the follow-up evaluation.

and the input photo. By doing so the unnecessarily detailed photo information does not need to be fully embedded in APDrawings. Along with localized discriminators for the eyes, nose and lips, our method can generate high-quality APDrawings in which all important facial features are preserved.

Learning from unpaired data makes our method able to utilize APDrawings from web data for training and include more challenging photos into the training set. To exploit the natural diversity of styles from web training images (see Fig. 2 for some examples), our method[1] further learns APDrawings in *multiple styles* from mixed web data and can control the output style using a simple style code.

The main contributions of our work are:

- We propose a novel asymmetric cycle-structure GAN model to avoid indiscriminately embedding reconstruction information in the whole APDrawing that is often caused by cycle consistency loss.
- We use multiple local discriminators to enforce the existence and ensure quality for facial feature drawing.
- We learn multi-style APDrawings from unpaired, mixed web data such that the user can switch between multiple styles using a simple style code.

2. Related Work

2.1. Neural Style Transfer

The power of CNNs has been validated by many visual perception tasks. Inspired by this, Gatys et al. [4] proposed to use a pretrained CNN to extract content features and style features from images and achieve style transfer by optimizing an image such that it maintains the content from the content image and matches the style features from the style image, where the Gram matrix is used to measure style similarity. This method opens up the field of neural style transfer and many follow-up methods are proposed based on this.

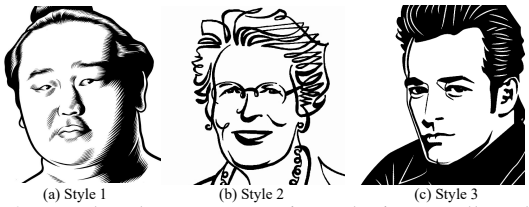
Li and Wand [13] proposed to maintain local patterns by using a Markov Random Field (MRF) regularizer instead of Gram matrix to model the style, and combined MRF with CNN to synthesize stylized images. To speed up the slow optimization process of [4], some methods (e.g., [11, 17]) use a feed-forward neural network to replace the optimization process and minimize the same objective function. However, these methods still suffer from the problem that each model is restricted to a single style. To speed up optimization and allow style flexibility as [4], Huang and Belongie [8] proposed adaptive instance normalization (AdaIN) to align the mean and variance of content features to those of style features. In these example-guided style transfer methods, the style is extracted from a single image, which is not as convincing as learning from a set of images to synthesize style (refer to Section 2.2). Moreover, these methods model style as texture, and thus are not suitable for our portrait line drawing style that has little texture.

2.2. GAN-based image-to-image translation

GANs [5] have achieved much progress in many computer vision tasks, including image super-resolution [12], text-to-image synthesis [16, 22], facial attribute manipulation [23], etc. Among these works, two unified GAN frameworks, Pix2Pix [10] and CycleGAN [25], have enabled much progress in image-to-image translation.

Pix2Pix [10] was the first general image-to-image translation framework based on conditional GANs, and was later

[1]The code is available at https://github.com/yiranran/Unpaired-Portrait-Drawing



(a) Style 1 (b) Style 2 (c) Style 3

Figure 2. We select three representative styles in our collected web portrait line drawing data. The first style is from Yann Legendre and Charles Burns where parallel lines are used to draw shadows. The second style is from Kathryn Rathke where few dark regions are used and facial features are drawn using simple flowing lines. The third style is from vectorportal.com where continuous thick lines and large dark regions are utilized.

extended to high-resolution image synthesis [18]. More works focus on learning from unpaired data, due to the difficulty of obtaining paired images in two domains. A popular and important observation is the cycle consistency constraint, which is the core of CycleGAN [25] and DualGAN [21]. The cycle consistency constraint enforces that the two mappings from domains A to B and from B to A when applied consecutively to an image revert the image back to itself. Different from enforcing cycle consistency at the image level, UNIT [15] tackles the problem by a shared latent space assumption and enforcing a feature-level cycle consistency. These methods work well for general image-to-image translation tasks. However, in face photo to APDrawing translation, cycle consistency constraints lead to facial features partially missing in APDrawings, because the information between the source and target domains is imbalanced. In this paper, we relax the cycle consistency in the forward (photo $\rightarrow$ drawing $\rightarrow$ photo) cycle and propose additional constraints to avoid this problem. The NIR (near infrared)-to-RGB method in [3] adopts a very different type of asymmetry: it uses the same loss for the forward and backward cycles, and only changes the network complexity. Moreover, it targets a different task from ours.

The aforementioned unpaired translation methods are also limited in the diversity of translation outputs. Unpaired data such as crawled web data often naturally contains multi-modal distributions (i.e. inconsistent styles). When knowing the exact number of modes and the mode each sample belongs to, the multi-modal image-to-image translation could be solved by treating each mode as a separate domain and using a multi-domain translation method [1]. However, in many scenarios including our problem setting, this information is not available. MUNIT [9] deals with multi-modal image-to-image translation without knowing the mode each sample belongs to. It encodes an image into a domain-invariant content code and a domain-specific style code, and recombines the content code with the style code sampled from a target domain. Although MUNIT generates multiple outputs with different styles, it cannot generate satisfactory portrait line drawings with clear lines. Our architecture inserts style features into the generator and uses a soft classification loss to discriminate modes in the training data and produce multi-style outputs, generating better APDrawings than state-of-the-art methods.

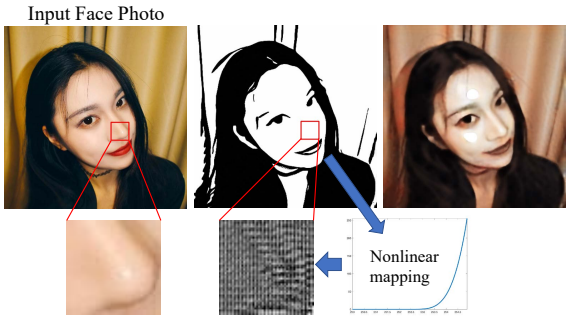


Figure 3. In CycleGAN, to reconstruct the input photo from generated drawings, a strict cycle-consistency loss embeds invisible reconstruction information indiscriminately in the whole drawings. A nonlinear monotonic mapping of the gray values is applied in a local region around the nose to visualize the embedded reconstruction information.

3. Our Method

3.1. Overview

Our proposed method performs face photo to APDrawing translation without paired training data using a novel asymmetric cycle-structure GAN. Let $\mathcal{P}$ and $\mathcal{D}$ be the face photo domain and the APDrawing domain, and no pairings need to exist between these two domains. Our model learns a function Φ that maps from $\mathcal{P}$ to $\mathcal{D}$ using training data $S(p) = \{p_i | i = 1, 2, \cdots, N\} \subset \mathcal{P}$ and $S(d) = \{d_j | j = 1, 2, \cdots, M\} \subset \mathcal{D}$. N and M are the numbers of training photos and APDrawings. Our model consists of two generators — a generator G transforming face photos to portrait drawings, and an inverse generator F transforming drawings back to face photos — and two discriminators, $D_{\mathcal{D}}$ responsible for discriminating generated drawings from real drawings, and $D_{\mathcal{P}}$ responsible for discriminating generated photos from real photos.

The information in the APDrawing domain is much less than in the face photo domain. For example, in the cheek region, there are many color variations in the original photo but the cheek is usually drawn completely white (i.e. no lines are included) in an APDrawing. Enforcing a strict cycle-consistency loss like in CycleGAN [25] on the reconstructed face photo and the input photo will cause the network to embed reconstruction information in very small variations in the generated APDrawings (i.e., color changes that are invisible to the eye but can make a difference in network calculation) [2]. See Fig. 3 for an example. Embedding reconstruction information in very small variations achieves a balance between cycle-consistency loss and GAN loss in CycleGAN; the generated drawing $G(p)$

can successfully reconstruct a face photo similar to the input photo because of small color changes, while at the same time $G(p)$ can be similar to real drawings and be classified as real by the discriminator. Embedding invisible reconstruction information indiscriminately in the whole drawing will put a very strong restriction on the objective function optimization. Moreover, it will allow important facial features to be partially missing in the generated drawings.

We observe that although cycle consistency constraints are useful to regularize training, we are only interested in the one way mapping from photos to portrait drawings. Therefore, different from CycleGAN, we do not expect or require the inverse generator F to reconstruct a face photo exactly as the input photo (which is a near impossible task). Instead, our proposed model is *asymmetric* in that we use a relaxed cycle-consistency loss between $F(G(p))$ and p, where only edge information is enforced to be similar, while a strict cycle-consistency loss is enforced on $G(F(d))$ and d. By tolerating the reconstruction information loss between $F(G(p))$ and p, the objective function optimization has enough flexibility to recover all important facial features in APDrawings. A truncation loss is further proposed to enforce the embedded information to be visible, especially around the local area of the selected edges where relaxed cycle-consistency loss works. Furthermore, local drawing discriminators for the nose, eyes and lips are introduced to enforce their existence and ensure quality for these regions in the generated drawings. By using these techniques, our method generates high-quality portrait line drawings with complete facial features.

Our model also deals with multi-style APDrawing generation. The APDrawing data we collected from the Internet contains a variety of styles, of which only some are tagged with author/source information. We select representative styles from the collected data (see Fig. 2), and train a classifier for the collected drawings. Then a learned representation is extracted as a style feature and inserted into the generator to control the generated drawing style. An additional style loss is introduced to optimize for each style.

The four networks in our model are trained in an adversarial manner [5]: the two discriminators $D_{\mathcal{D}}$ and $D_{\mathcal{P}}$ are trained to maximize the probability of assigning correct labels to real and synthesized drawings and photos; and meanwhile the two generators G and F are trained to minimize the probability of the discriminators assigning the correct labels. The loss function $L(G, F, D_{\mathcal{D}}, D_{\mathcal{P}})$ contains five types of loss terms: adversarial loss $L_{adv}(G, D_{\mathcal{D}}) + L_{adv}(F, D_{\mathcal{P}})$, relaxed cycle consistency loss $L_{relaxed-cyc}(G, F)$, strict cycle consistency loss $L_{strict-cyc}(G, F)$, truncation loss $L_{trunc}(G, F)$, and style loss $L_{style}(G, D_{\mathcal{D}})$. Then the function Φ is optimized by solving the minimax problem with loss function $L(G, F, D_{\mathcal{D}}, D_{\mathcal{P}})$:

$$\begin{aligned}&\min_{G,F}\max_{D_{\mathcal{D}},D_{\mathcal{P}}} L(G, F, D_{\mathcal{D}}, D_{\mathcal{P}})\\ &= (L_{adv}(G, D_{\mathcal{D}}) + L_{adv}(F, D_{\mathcal{P}})) + \lambda_1 L_{relaxed-cyc}(G, F)\\ &+ \lambda_2 L_{strict-cyc}(G, F) + \lambda_3 L_{trunc}(G, F) + \lambda_4 L_{style}(G, D_{\mathcal{D}})\end{aligned} \tag{1}$$

In Section 3.2, we introduce the architectures of our model and our different designs for $G, D_{\mathcal{D}}$ and $F, D_{\mathcal{P}}$. In Section 3.3, we introduce our asymmetric cycle-consistency requirements and five loss terms. An overview of our method is illustrated in Fig. 4.

3.2. Architecture

Our GAN model consists of a generator G and a discriminator $D_{\mathcal{D}}$ for face photo to drawing translation, and another generator F and discriminator $D_{\mathcal{P}}$ for the inverse drawing to photo translation. Considering information imbalance between the face photo in $\mathcal{P}$ and the APDrawing in $\mathcal{D}$, we design different architectures for $G, D_{\mathcal{D}}$ and $F, D_{\mathcal{P}}$.

3.2.1 Face photo to drawing generator G

The generator G takes a face photo p and a style feature s as input, and outputs a portrait line drawing $G(p, s)$ whose style is specified by s.

Style feature s. We first train a classifier C (based on VGG19) that classifies portrait line drawings into three styles (Fig. 2), using tagged web drawing data. Then we extract the output of the last fully-connected layer and use a softmax layer to calculate a 3-dimensional vector as the style feature for each drawing (including untagged ones).

Network structure. G is an encoder-decoder with residual blocks [7] in the middle. It starts with a flat convolution and two down convolution blocks to encode face photos and extract useful features. Then the style feature is mapped to a 3-channel feature map and inserted into the network by concatenating it with the feature map of the second down convolution block. An additional flat convolution is used to merge the style feature map with the extracted feature map. Afterwards, nine residual blocks of the same structure are used to construct the content feature and transfer it to the target domain. Then the output drawing is reconstructed by two up convolution blocks and a final convolution layer.

3.2.2 Drawing discriminator $D_{\mathcal{D}}$

The drawing discriminator $D_{\mathcal{D}}$ has two tasks: 1) to discriminate generated portrait line drawings from real ones; and 2) to classify a drawing into three selected styles, where a real drawing d is expected to be classified into the correct style label (given by C), and a generated drawing $G(p, s)$ is expected to be classified into the style specified by the 3-dimensional style feature s.

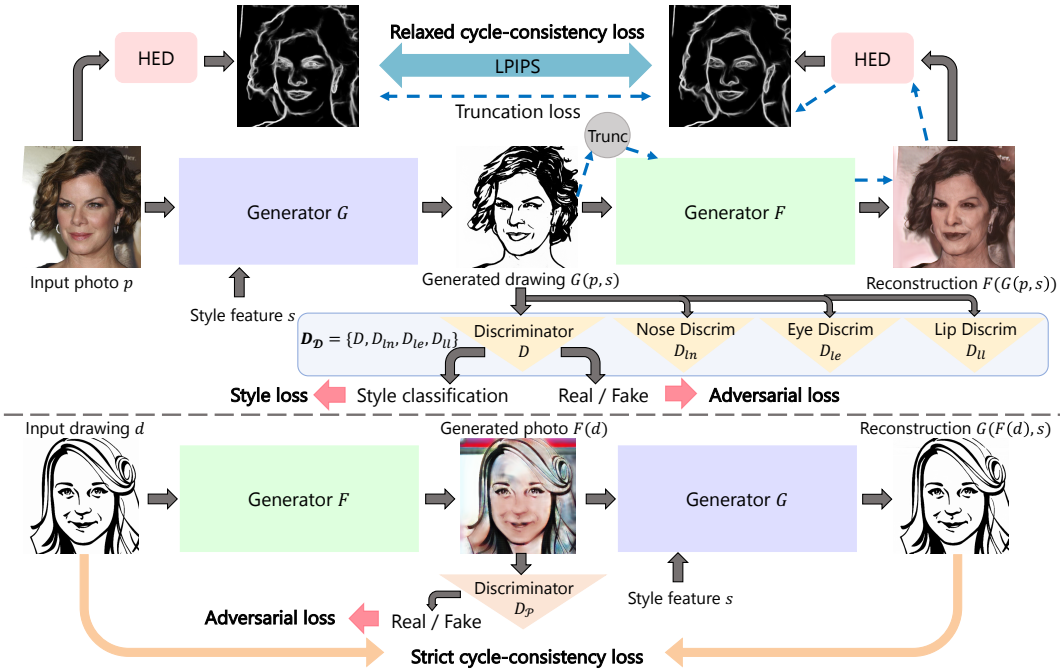


Figure 4. Our model is an asymmetric cycle-structure GAN model that consists of a photo to drawing generator G, a drawing to photo generator F, a drawing discriminator $D_{\mathcal{D}}$ and a photo discriminator $D_{\mathcal{P}}$. We use a relaxed cycle-consistency loss between reconstructed face photo $F(G(p))$ and input photo p, while enforcing a strict cycle-consistency loss between reconstructed drawing $G(F(d))$ and input drawing d. We further introduce local drawing discriminators D_{ln}, D_{le}, D_{ll} for the nose, eyes and lips and a truncation loss. Our model deals with multi-style generation by inserting a style feature into the generator and adding a style loss.

For the first task, to enforce the existence of important facial features in the generated drawing, besides a discriminator D that analyzes the full drawing, we add three local discriminators D_{ln}, D_{le}, D_{ll} to focus on discriminating the nose drawing, eye drawing and lip drawing respectively. The inputs to these local discriminators are masked drawings, where masks are obtained from a face parsing network [6]. $D_{\mathcal{D}}$ consists of $D, D_{ln}, D_{le}, D_{ll}$.

Network structure. The global discriminator D is based on PatchGAN [10] and modified to have two branches. The two branches share three down convolution blocks. Then one branch D_{rf} includes two flat convolution blocks to output a prediction map of real/fake for each patch in the drawing. And the other classification branch D_{cls} includes more down convolution blocks and outputs probability values for the three style labels. Local discriminators D_{ln}, D_{le}, D_{ll} also adopt the PatchGAN structure.

3.2.3 Drawing to face photo generator F and Photo discriminator $D_{\mathcal{P}}$

The generator F in the inverse direction takes a portrait line drawing d as input and outputs a face photo $F(d)$. It adopts an encoder-decoder architecture with nine residual blocks in the middle. Photo discriminator $D_{\mathcal{P}}$ discriminates generated face photos from real ones, and also adopts the PatchGAN structure.

3.3. Loss Functions

There are five types of losses in our loss function (Eq. (1)). We explain them in detail as follows:

Adversarial loss. The adversarial loss judges discriminator $D_{\mathcal{D}}$'s ability to assign correct labels to real and synthesized drawings. It is formulated as:

$$L_{adv}(G, D_{\mathcal{D}}) = \sum_{D \in D_{\mathcal{D}}} \mathbb{E}_{d \in S(d)}[\log D(d)] + \sum_{D \in D_{\mathcal{D}}} \mathbb{E}_{p \in S(p)}[\log(1 - D(G(p, s))] \quad (2)$$

where s is randomly selected from the style features of drawings in $S(d)$ for each p. As $D_{\mathcal{D}}$ maximizes this loss and G minimizes it, this loss drives the generated drawings to become closer to real drawings.

We also adopt an adversarial loss for the photo discriminator $D_{\mathcal{P}}$ and the inverse mapping F:

$$L_{adv}(F, D_{\mathcal{P}}) = \mathbb{E}_{p \in S(p)}[\log D_{\mathcal{P}}(p)] + \mathbb{E}_{d \in S(d)}[\log(1 - D_{\mathcal{P}}(F(d))] \quad (3)$$

Relaxed forward cycle-consistency loss. As aforementioned, we observe that there is much less information in domain $\mathcal{D}$ than information in domain $\mathcal{P}$. We do not expect $p \rightarrow G(p, s) \rightarrow F(G(p, s))$ to be pixel-wise similar to p. Instead, we only expect the edge information in

p and $F(G(p,s))$ to be similar. We extract edges from p and $F(G(p,s))$ using HED [19], and evaluate the similarity of edges by the LPIPS perceptual metric proposed in [24]. Denote HED by H and the perceptual metric by L_{lpips}, the relaxed cycle-consistency loss is formulated as:

$$L_{relaxed-cyc}(G,F) = \mathbb{E}_{p\in S(p)}[L_{lpips}(H(p), H(F(G(p,s))))] \tag{4}$$

Strict backward cycle-consistency loss. On the other hand, the information in the generated face photo is adequate to reconstruct the drawing. Therefore, we expect $d \rightarrow F(d) \rightarrow G(F(d), s(d))$ to be pixel-wise similar to d, here the style feature $s(d)$ is the style feature of d. The strict cycle-consistency loss in the backward cycle is then formulated as:

$$L_{strict-cyc}(G,F) = \mathbb{E}_{d\in S(d)}[||d - G(F(d), s(d))||_1] \tag{5}$$

Truncation loss. The truncation loss is designed to prevent the generated drawing from hiding information in small values. It is in the same format as the relaxed cycle-consistency loss, except that the generated drawing $G(p,s)$ is first truncated to 6 bits (a general digital image stores intensity in 8 bits) to ensure encoded information is clearly visible, and then fed into F to reconstruct the photo. Denote the truncation operation as $T[\cdot]$, the truncation loss is formulated as:

$$L_{trunc}(G,F) = \mathbb{E}_{p\in S(p)}[L_{lpips}(H(p), H(F(T[G(p,s)])))] \tag{6}$$

In the first period of training, the weight for the truncation loss is kept low, otherwise it would be too hard for the model to optimize. The weight gradually increases as the training progresses.

Style loss. The style loss is introduced to help G generate multiple styles with different style features. Denote the classification branch in $D_{\mathcal{D}}$ as D_{cls}, the style loss is formulated as

$$\begin{aligned} L_{cls}(G, D_{\mathcal{D}}) = \mathbb{E}_{d\in S(d)}[-\sum_c p(c)\log D_{cls}(c|d)] \\ + \mathbb{E}_{p\in S(p)}[-\sum_c p'(c)\log D_{cls}(c|G(p,s))] \end{aligned} \tag{7}$$

For real drawing d, $p(c)$ is the probability over style label c given by classifier C, $D_{cls}(c|d)$ is the predicted softmax probability by D_{cls} over c. We multiply by the probability $p(c)$ in order to take into account those real drawings that may not belong to a single style but lie between two styles, e.g. softmax probability $[0.58, 0.40, 0.02]$. For generated drawing $G(p,s)$, $p'(c)$ denotes the probability over style label c and is specified by style feature s, $D_{cls}(c|G(p,s))$ is the predicted softmax probability over c. This classification loss drives D_{cls} to classify a drawing into the correct style and drives G to generate a drawing close to a given style feature.

Figure 5. Comparison with two state-of-the-art neural style transfer methods, i.e., Gatys [4] and LinearStyleTransfer [14].

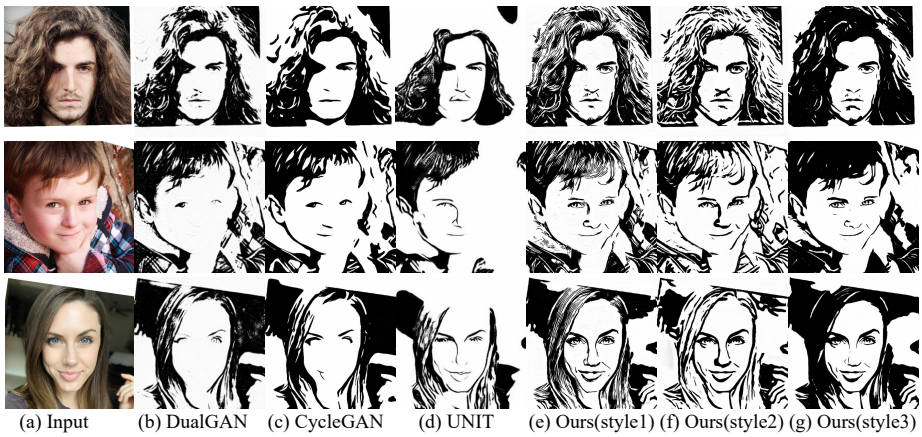
4. Experiments

We implemented our method in PyTorch. All experiments are performed on a computer with a Titan Xp GPU. The parameters in Eq. 1 are $\lambda_1 = 5 - \frac{4.5i}{n}$, $\lambda_2 = 5$, $\lambda_3 = \frac{4.5i}{n}$, $\lambda_4 = 1$, where i is the current epoch number, and n is the total epoch number.

4.1. Experiment Setup

Data. We collect face photos and APDrawings from the Internet and construct a training corpus of 798 face photos and 625 delicate portrait line drawings, and a test set of 154 face photos. Among the collected drawings, 84 are labeled with artist Charles Burns, 48 are labeled with artist Yann Legendre, 88 are labeled with artist Kathryn Rathke, 212 are from website vectorportral.com, while others have no tagged author/source information. We observed that both Charles Burns and Yann Legendre use similar parallel lines to draw shadows, and so we merged drawings of these two artists into style1. We select the drawings of Kathryn Rathke as style2 and the drawings of vectorportral as style3. Both of them have distinctive features: Kathryn Rathke uses flowing lines but few dark regions and vectorportral uses thick lines and large dark regions. All the training images are resized and cropped to 512×512 pixels.

Training process. 1) Training classifier C. We first train a style classifier C (Section 3.2.1) with the tagged drawings and data augmentation (including random rotation, translation and scaling). To balance the number of drawings in each style, we take all drawings from the first and second styles but only part of the third style in training stage of C, to achieve more balanced training for different styles. 2) Training our model. Then we use the trained classifier to obtain style features for all 625 drawings. We further augment training data using synthesized drawings. Training our network with the mixed data of real drawings and synthesized drawings results in high-quality generation for all three styles (Figs. 5-7, where our results



(a) Input (b) DualGAN (c) CycleGAN (d) UNIT (e) Ours(style1) (f) Ours(style2) (g) Ours(style3)

Figure 6. Comparison with three single-modal unpaired image-to-image translation methods: DualGAN [21], CycleGAN [25], UNIT [15].

(a) Input (b) MUNIT (c) ComboGAN (style1, 2, 3) (d) Ours (style1, 2, 3)

Figure 7. Comparison with two unpaired image-to-image translation methods that can deal with multi-modal or multi-domain translation: MUNIT [9], ComboGAN [1].

of styles 1, 2, 3 are generated by feeding in a style feature of $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, 1]$ respectively).

4.2. Comparisons

We compare our method with two state-of-the-art neural style transfer methods: Gatys [4], LinearStyleTransfer [14], and five unpaired image-to-image translation methods: DualGAN [21], CycleGAN [25], UNIT [15], MUNIT [9] and ComboGAN [1].

Comparisons with neural style transfer methods are shown in Fig. 5. Gatys' method fails to capture portrait line drawing styles because it uses the Gram matrix to model style as texture and APDrawings have little texture. LinearStyleTransfer produces visually better results but still not the desired line drawing: the generated drawings have many thick lines but they are produced in a rough manner. Compared to these example-guided style transfer methods, our method learns from a set of APDrawings and generates delicate results for all three styles.

Comparisons with single-modal unpaired image-to-image translation methods are shown in Fig. 6. DualGAN and CycleGAN are both based on strict cycle-consistency loss. This causes a dilemma in photo to line drawing translation: either a generated drawing looks like a real drawing (i.e. close to binary, containing large uniform regions) but cannot properly reconstruct the original photo; or a generated drawing has grayscale changes and good reconstruction but does not look like a real drawing. Also, compared to CycleGAN, DualGAN is more grayscale-like, less abstract and worse in line drawing style. UNIT adopts feature-level cycle-consistency loss, which makes the results less constrained at the image level, making the face appear deformed. In comparison, our results both preserve face structure and have good image and line quality.

Comparisons with unpaired image-to-image translation methods that can deal with multi-modal or multi-domain translation are shown in Fig. 7. Results show that MUNIT does not capture the line drawing style in our task and the results are more similar to a pencil drawing with shading and many gray regions. ComboGAN fails to capture all three representative styles and performs better on styles 2 and 3 than style 1. Our architecture inserts style information earlier in the generator, which gives more space for multi-style generation. As a result, our method generates distinctive re-

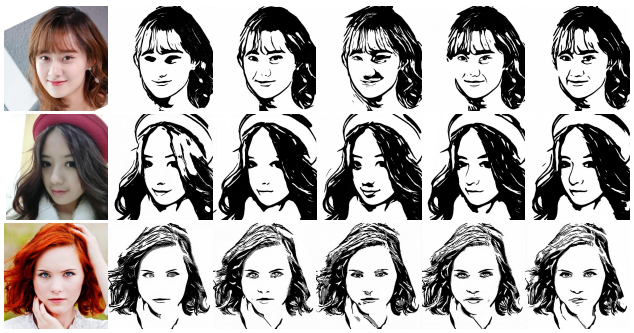


Figure 8. Ablation study: (a) input photos, (b) results of removing relaxed cycle-consistency loss (i.e. using L_1 loss) and removing local discriminators, (c) results of removing relaxed cycle-consistency loss, (d) results of removing local discriminators, (e) results of removing HED in calculating relaxed cycle-consistency loss, (f) our results.

Table 1. User study results. The i-th row shows the percentages of different methods (LinearStyleTransfer (LST) [14], CycleGAN [25], ComboGAN [1] and Ours) being ranked as the i-th among four methods.

	LST	ComboGAN	CycleGAN	Ours
Rank1	1.3%	14.9%	15.2%	**68.5%**
Rank2	7.4%	31.8%	38.8%	22.0%
Rank3	31.2%	31.4%	30.0%	7.4%
Rank4	60.1%	21.9%	15.9%	2.1%

sults for three styles and reproduces the characteristics for each style well.

4.3. User Study

We conduct a user study to compare our method with LinearStyleTransfer (LST), CycleGAN and ComboGAN (Gatys, DualGAN and UNIT are not included because of lower visual quality and MUNIT is not included because it does not capture the line drawing style). We randomly sample 60 face photos from the test set and translate 20 of them to each style. The style reference needed for LST is randomly chosen from real drawings. Participants are shown a photo, a real drawing (style reference) and four generated drawings at a time, and are asked to drag and sort four results based on style similarity, face structure preservation and image quality. 34 participants attended the user study and 2,040 votes were collected in total. Results of the percentages of each method ranked as 1,2,3,4 are summarized in Table. 1. Our method ranks the best in 68.5% of votes, while LST, ComboGAN and CycleGAN rank the best in 1.3%, 14.9% and 15.2% instances. The average rank of our method is 1.43, compared to CycleGAN's 2.47, ComboGAN's 2.60 and LST's 3.50. These results demonstrate that our method outperforms other methods. All generated drawings evaluated in user study are presented in the supplementary material. And we provide another quantitative evaluation (FID evaluation) in the supplementary material.

4.4. Ablation Study

We perform an ablation study on the key factors in our method: (1) relaxed cycle consistency loss, (2) local discriminators and (3) HED edge extraction. Results show that they are all essential to our method.

As shown in Fig. 8b, without relaxed cycle consistency loss and local discriminators, facial features are often missing (e.g. the nose is missing in the first and second rows, nose and eye details are missing in the third row). Removing only relaxed cycle consistency loss (Fig. 8c) preserves more facial feature regions (e.g., the nose in the first row) than (Fig. 8b) but still some parts are missing. Removing only local discriminators (Fig. 8d) produces few missing parts (much better than (Fig. 8b) in facial structure preservation), but some facial features are not drawn in the desired manner: some black regions or shadows that are usually drawn near facial boundaries or hair appear near the nose. When both relaxed cycle consistency loss and local discriminators are used, results (Fig. 8f) preserve all facial feature regions and no undesired black regions or shadows appear in faces. These results show that both relaxed cycle consistency loss and local discriminators help to preserve facial feature regions and are complementary to each other, and local discriminators also help to avoid undesired elements in facial features.

As shown in Fig. 8e, without HED edge extraction in the relaxed cycle consistency loss calculation, the lines are often discontinuous or blurred (see the nose in the first and second rows, and eyes and lips in the third row). In comparison, our results have clear, sharp and continuous lines. This result demonstrates that using HED edge extraction helps the model to generate clearer and more complete lines.

5. Conclusion

In this paper, we propose a method for unpaired portrait line drawing generation using asymmetric cycle mapping. Our method can learn multi-style portrait drawing generation from mixed web data using an additional style feature input and a soft classification loss. Experiments and a user study demonstrate that our method can generate high-quality distinctive results for three representative styles and outperform state-of-the-art methods.

Acknowledgement

This work was partially supported by NSFC (61725204, 61521002), BNRist, MOE-Key Laboratory of Pervasive Computing and Royal Society-Newton Advanced Fellowship (NA150431).

References

[1] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. ComboGAN: unrestrained scalability for image domain translation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 783–790, 2018. 2, 3, 7, 8

[2] Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a master of steganography. *CoRR*, abs/1712.02950, 2017. 3

[3] Hao Dou, Chen Chen, Xiyuan Hu, and Silong Peng. Asymmetric CycleGan for unpaired NIR-to-RGB face image translation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 1757–1761. IEEE, 2019. 3

[4] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 1, 2, 6, 7

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 2, 4

[6] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3436–3445, 2019. 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[8] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 2

[9] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *15th European Conference (ECCV)*, pages 179–196, 2018. 2, 3, 7

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 1, 2, 5

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *14th European Conference (ECCV)*, pages 694–711, 2016. 2

[12] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 2

[13] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2479–2486, 2016. 2

[14] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3809–3817, 2019. 2, 6, 7, 8

[15] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 700–708, 2017. 2, 3, 7

[16] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1060–1069, 2016. 2

[17] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1349–1357, 2016. 2

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 3

[19] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1403, 2015. 6

[20] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. ApdrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10743–10752, 2019. 1, 2

[21] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876, 2017. 1, 2, 3, 7

[22] Han Zhang, Tao Xu, and Hongsheng Li. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916, 2017. 2

[23] Jichao Zhang, Yezhi Shu, Songhua Xu, Gongze Cao, Fan Zhong, Meng Liu, and Xueying Qin. Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. In *ACM Multimedia Conference on Multimedia Conference (MM)*, pages 392–401, 2018. 2

[24] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 6

[25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. 1, 2, 3, 7, 8