Cartoon-to-Photo Facial Translation with Generative Adversarial Networks

Junhong Huang[†] EE_H.JUNHONG@MAIL.SCUT.EDU.CN
School of Electronic and Information Engineering, South China University of Technology
Mingkui Tan[†] MINGKUITAN@SCUT.EDU.CN
School of Software Engineering, South China University of Technology
Yuguang Yan YAN.YUGUANG@MAIL.SCUT.EDU.CN
School of Software Engineering, South China University of Technology
Chunmei Qing[*] QCHM@SCUT.EDU.CN
School of Electronic and Information Engineering, South China University of Technology
Qingyao Wu QYW@SCUT.EDU.CN
School of Software Engineering, South China University of Technology
Zhuliang Yu ZLYU@SCUT.EDU.CN
School of Automation Science and Engineering, South China University of Technology

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Cartoon-to-photo facial translation could be widely used in different applications, such as law enforcement and anime remaking. Nevertheless, current general-purpose image-to-image models usually produce blurry or unrelated results in this task. In this paper, we propose a Cartoon-to-Photo facial translation with Generative Adversarial Networks (CP-GAN) for inverting cartoon faces to generate photo-realistic and related face images. In order to produce convincing faces with intact facial parts, we exploit global and local discriminators to capture global facial features and three local facial regions, respectively. Moreover, we use a specific content network to capture and preserve face characteristic and identity between cartoons and photos. As a result, the proposed approach can generate convincing high-quality faces that satisfy both the characteristic and identity constraints of input cartoon faces. Compared with recent works on unpaired image-to-image translation, our proposed method is able to generate more realistic and correlative images.

Keywords: Generative Adversarial Networks, image-to-image translation, cartoon-to-photo translation

1. Introduction

Cartoon-to-photo facial translation is an interesting yet challenging task, which aims to convert cartoon faces into the photo-realistic ones. This task could be used in many real-world applications. For example, in criminal investigation, a witness can easily produce a cartoon face image of the suspect by selecting the components of hairstyles, noises, eyes,

[†]The first two authors contributed to this work equally.
[*]Correspondence should be addressed to C. Qing.

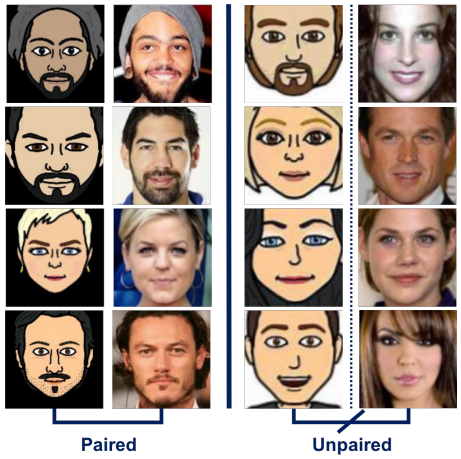

Figure 1: Comparison of paired and unpaired data. The data in cartoon-to-photo facial translation task are often unpaired. Here, the paired samples (the left panel) are created manually by a web interface (bitmoji.com), which can translates a face photo to a corresponding cartoon picture, such that the face and cartoon images have the same identity. In contrast, the unpaired samples (the right panel) consist of a source set $\{x_i\}_{i=1}^{N} \in X$ and a target set $\{y_j\}_{j=1}^{M} \in Y$, where x_i is collected independently regardless of Y. For unpaired data, since there is no extra information to indicate whether x_i corresponds to y_j, the learning task is challenging.

eyebrows, and mouths. It would be much more efficient if the police can generate the corresponding face photo from the cartoon image, rather than drawing a picture based on the witness's oral descriptions. Besides, anime has attracted much more fans than ever before in recent years. It could be very interesting to see how the cartoon characters look like in reality. Moreover, cartoon-to-photo facial translation can also help to remake anime productions into the realistic version.

One potential method for cartoon-to-photo translation is to leverage paired training samples (the left panel in Figure 1), which, however, suffers from expensive cost of data collection. As alternatives, some existing works of image-to-image translation leverage unpaired samples (the right panel in Figure 1) to perform the cartoon-to-photo application (Liu et al., 2017; Zhu et al., 2017; Benaim and Wolf, 2017). Nevertheless, there are still several challenges remaining to be addressed. First, in a generated face photo, some local facial regions are easy to collapse, resulting in unrealistic face components. Next, a generated photo with intact facial regions may look blurry as a whole. Third, input and output faces may be unrelated, which means that two photos involve two different identities.

To tackle the above issues, we propose a task-specific method to invert cartoon faces to generate photo-realistic face images based on unpaired samples. We exploit Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to learn a joint distribution of images in two domains. To make the generated faces sharp and recover the collapsed facial parts, we divide an overall face image into several local regions and propose global and local

discriminators to capture the joint distributions of overall faces and local regions, respectively, in cartoon and photo domains. These encourage the generated faces to achieve high face quality with intact facial parts. Moreover, inspired by neural style transfer (Johnson et al., 2016b), we adopt a perceptual loss in the feature space defined by a face recognition network to preserve the characteristic and identity of cartoon faces.

The main contributions of our work are summarized as follows:

- We propose a new task-specific method CP-GAN for cartoon-to-photo facial translation task;
- We combine global discriminator with well-designed local discriminator to encourage the generated faces to achieve high face quality with intact facial parts;
- Experimental results demonstrate that CP-GAN not only highly preserves the characteristic and identity of cartoon faces, but also achieves good face quality on cartoon-to-photo facial translation task.

The remainer of the paper is organized as follows. In Section 2, we briefly review some important related works. Next, we present our proposed CP-GAN method in Section 3. Section 4 shows experimental results on real-world datasets, and Section 5 concludes the whole paper.

2. Related Work

2.1. Generative Adversarial Networks

Generative adversarial networks (GANs) have shown impressive results in image generation (Arjovsky et al., 2017; Gulrajani et al., 2017; Radford et al., 2016; Karras et al., 2018; Cao et al., 2018) and translation (Arjovsky et al., 2017; Gulrajani et al., 2017; Kim et al., 2017). GANs conduct a two-player minimax game, in which a discriminator aims to distinguish between real and generated samples, while a generator attempts to fool the discriminator with realistic generate samples. During the adversarial learning procedure, the discriminator is trained to obtain better distinguishable ability, and the generator is trained to generate higher-quality samples that are indistinguishable from real samples.

In order to improve the stability of the training of GANs, some variants of GANs are proposed in the last two years Nowozin et al. (2016); Chen et al. (2016); Roth et al. (2017). Salimans et al. (2016) developed several techniques according to experimental experience. Motivated by the Wasserstein discrepancy in optimal transport, Arjovsky et al. (2017) proposed Wasserstein GANs with a new objective function, which is derived from the dual form of the Wasserstein discrepancy. After that, Gulrajani et al. (2017) further improved Wasserstein GANs by penalizing the gradient values of parameters. Mao et al. (2017) also designed a new objective function for GANs based on the least square loss. Qi (2017) proposed loss-sensitive GANs to focus on generated samples with poor qualities.

In order to generate images involving specific information, some variants of GANs are proposed for different applications. For example, in (Mirza and Osindero, 2014), label condition is introduced into GANs to generate images with a given label. GAN-CLS generates object pictures according to given text descriptions (Reed et al., 2016). In (Isola et al.,

2017), image-to-image translation is investigated based on paired training data. CAAE applies conditional model to generate faces with a predefined age given a face photo Zhang et al. (2017). DR-GAN finds pose-invariant face features and synthesizes identity-preserving faces given a face photo and a target pose code Tran et al. (2017).

2.2. Unpaired Image-to-Image Translation

In many real-world applications, it is expensive or difficult to collect sufficient paired training data for image translation. Therefore, image-to-image translation based on unpaired training data has been attracting more and more attention. CoGAN exploits a parameter-sharing strategy to learn common representations shared by different image domains (Liu and Tuzel, 2016). UNIT extends CoGAN later by combining GANs and variational auto-encoders (Liu et al., 2017). SimGAN (Shrivastava et al., 2017) applies ℓ_1 loss to measure the pixel difference between a synthetic picture and a generated refined picture. Similarly, DTN (Taigman et al., 2017) measures the difference between an input and output in a feature space. In (Zhu et al., 2017), CycleGAN leverages the cycle consistency scheme to preserve the common information between input and output images. The similar idea is also used in DualGAN (Yi et al., 2017) and DiscoGAN (Kim et al., 2017).

The above-mentioned works are general-purpose approaches for unpaired image translation without considering the specific characteristics in the application of cartoon-to-photo. In this paper, instead, we target the cartoon-to-photo facial translation task and propose a generative adversarial network to produce realistic face photos with preserved identity information. To achieve this, we introduce a specific content network to capture identity information. Moreover, we design a local discriminator to distinguish between real and generated local regions of face photos.

2.3. Neural Style Transfer

Neural style transfer is a special case of unpaired image-to-image translation (Gatys et al., 2016; Johnson et al., 2016b; Luan et al., 2017). Different from the above works that consider multiple image domains, neural style transfer usually combines the content of one picture and the style of another picture. The perceptual loss is used to preserve the content, and the Gram matrix statistics are commonly used to match the style. This task focuses on the content and style of a single image. In this cartoon-to-photo application, we consider two different image domains and leverage the adversarial loss to capture the "style" of the face images.

3. Method

Our goal is to learn a joint distribution of faces in cartoon domain X and photo domain Y, given unpaired training samples $\{x_i\}_{i=1}^N \in X$ and $\{y_j\}_{j=1}^M \in Y$. As shown in Figure 2, our network consists of four components: a cartoon-to-photo facial translation network G, a global discriminator D_g, a local discriminator D_l, and a content network C. D_g aims to distinguish between real face photos $\{y\}$ and generated face photos $\{G(x)\}$, and D_l aims to distinguish between real photo patches $\{y_i^P\}$ and generated patches $\{G(x)_i^P\}$, where three patches are extracted from each photo, *i.e.*, $i \in \{1, 2, 3\}$. G tries to generate realistic face

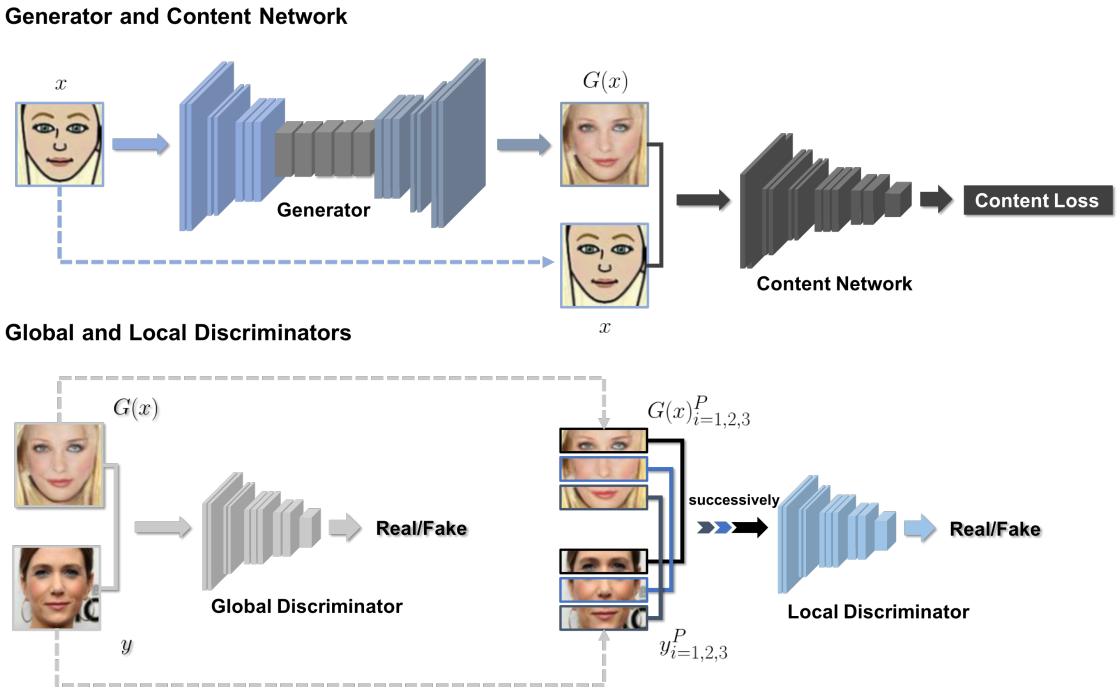

Figure 2: Illustration of the proposed method. After the generator G transfers a cartoon face x into the generated face $G(x)$, x and $G(x)$ are fed successively to the pre-trained content network C. G is updated based on the content loss between two outputs of C, while C is not updated. Next, we extract three patches respectively from $G(x)$ and a real sample y, and then feed them into the local discriminator D_l. D_l forces three patches of $G(x)$ to be plausible and photo-realistic. Moreover, the global discriminator D_g encourages G to translate x into the output $G(x)$ indistinguishable from y. In this figure, the full lines connecting to two images denote that we feed these two images successively into the model.

photos $\{G(x)\}$ to fool the discriminators D_g and D_l, such that the generated pictures will be incorrectly classified as real. Moreover, we introduce the content network C to capture the characteristic and identity within an input cartoon picture and the corresponding generated face photo, and minimize the distance between them.

3.1. Global and Local Adversarial Losses

Both cartoon face and real face consist of several local components, such as eyes, eyebrows, nose, mouth and so on. These components are connected to each other with their positions and shapes. To generate high-quality real faces when given the cartoon faces, we need not only to learn the data distribution of the overall faces but also pay attention to refine the local elements of the face. To this end, we propose a global and local adversarial learning scheme to train our network.

Global Discriminator. Global discriminator is used to encourage the generator to capture the joint distribution of overall faces in cartoon and photo domains. It aims to capture overall facial information and produces real faces with plausible facial shape and skin texture. Through experiments, we observe that there exist two problems when using the global discriminator alone. Firstly, the generated faces achieve high image quality but some facial parts collapse. Secondly, each facial part is intact while the overall faces look blurry. To tackle the above problems and better preserve the local face details, we propose a local discriminator.

Local Discriminator. The local discriminator D_l is used to recover the collapsed parts and make faces sharper. Ideally, local discriminator aims to learn the joint distribution of each facial component in cartoon and photo domains. However, we cannot extract facial components precisely because of the limitation of labels. In practice, three equal-sized patches are extracted from a real image y based on its 5 landmarks (left eye, right eye, nose, left mouth, left and right mouth corners), where each patch is padded as the size of y before fed into D_l, as shown in Figure 3. The first patch contains eyebrows and eyes, while the other two patches contain nose and mouth, respectively. We do the same operation for the generated face $\{G(x)\}$. Moreover, we preserve the original position of each patch in the padded images as additional conditions for D_l, which helps D_l capture the distribution of each part independently.

The objective of the global and local facial adversarial learning scheme can be expressed as (Goodfellow et al., 2014):

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_g) =& \mathbb{E}_{y\sim p_{data}(y)}[\log D_g(y)] \\ &+ \mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D_g(G(x)))],\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_l) =& \mathbb{E}_{y\sim p_{data}(y)}\mathbb{E}_{i=1,2,3}[\log D_l(y_i^P)] \\ &+ \mathbb{E}_{x\sim p_{data}(x)}\mathbb{E}_{i=1,2,3}[\log(1 - D_l(G(x)_i^P))].\end{aligned} \tag{2}$$

Here, we hope to find G to generate the global face $G(x)$ and local specified faces $G(x)_{i=1,2,3}^p$, so that they can achieve similar appearance to real ones by minimizing $\mathcal{L}_{GAN}(G, D_g)$ and $\mathcal{L}_{GAN}(G, D_l)$. Specifically, we seek to solve the following minimax problem to obtain an optimal solution (G^*, D_g^*, D_l^*):

$$(G^*, D_g^*, D_l^*) = \arg\min_G \max_{D_g, D_l} \mathcal{L}_{GAN}(G, D_g) + \lambda\mathcal{L}_{GAN}(G, D_l), \tag{3}$$

where λ is a trade-off parameter. In our design, the global discriminator is used to encourage the generator to learn overall face information including facial shape and skin texture, while local discriminator is used to ensure that each local facial part is intact.

3.2. Content Loss

Besides the high quality in image generation, we also require that the generated face photos look similar to the given cartoon pictures. To achieve this, we exploit the content network C to capture the characteristic and identity information of an input picture, thus obtaining the content representation of the input. By minimizing the difference between the content representations of the input cartoon picture x and the corresponding generated face photo

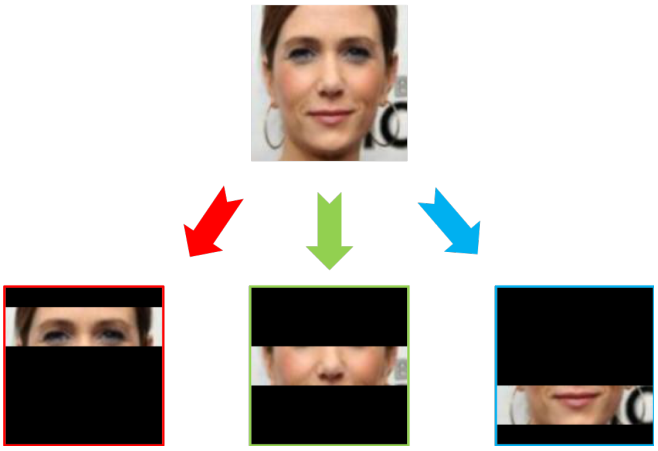

Figure 3: The cropping and padding scheme for local adversarial learning. Three equal-sized regions are extracted from the input and padded as the size of the input. These specified regions contain eyebrows and eyes, nose, and mouth.

$G(x)$, these two pictures are encouraged to have similar representations. As a result, the content information of the input picture can be preserved. The objective to minimize is given as follows:

$$\mathcal{L}_{content}(G, C) = \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}||C(G(x)) - C(x)||^2, \tag{4}$$

where $C(x)$ and $C(G(x))$ are the feature maps.

In this task, the face identity is the common information of cartoon and photo faces. Here, we use the face identity feature extracted by a pretrained model with the high accuracy of face recognition, instead of the one learned from scratch without huge number of training samples and face identity labels.

3.3. Full Objective Function

By considering the content loss, we reformulate our final objective as follows:

$$\mathcal{L}(G, D, C) = \mathcal{L}_{GAN}(G, D_g) + \lambda\mathcal{L}_{GAN}(G, D_l) + \gamma\mathcal{L}_{content}(G, C), \tag{5}$$

where λ and γ are trade-off parameters.

In Eq. (5), we exploit both global and local adversarial losses to make the generated faces plausible, and apply the content loss to capture the common information between cartoon faces and generated faces, *i.e.*, the characteristic and identity. The content network and the local discriminator work cooperatively, which can achieve more precise translation. In this paper, we employ the alternative optimization method to address Problem (5) (Goodfellow et al., 2014; Mirza and Osindero, 2014).

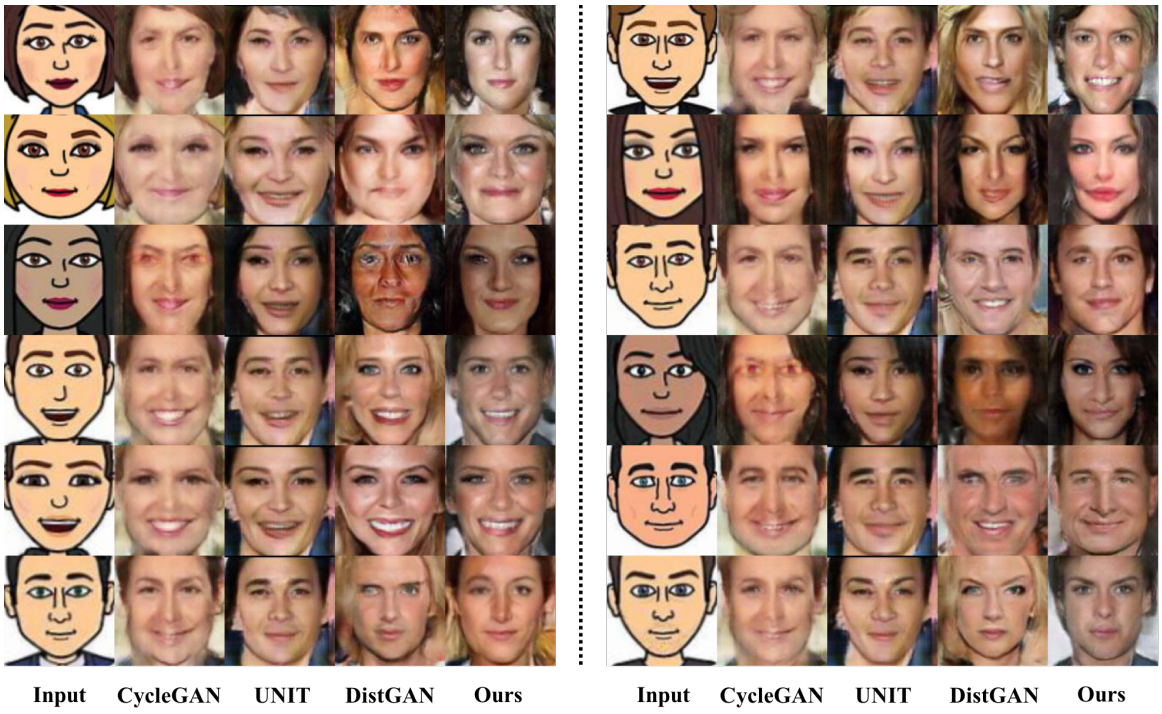

Figure 4: Comparison with state-of-the-art. (a) Input, (b) CycleGAN (Zhu et al., 2017), (c) UNIT (Liu et al., 2017), (d) Distance GAN (Benaim and Wolf, 2017), (e) Ours.

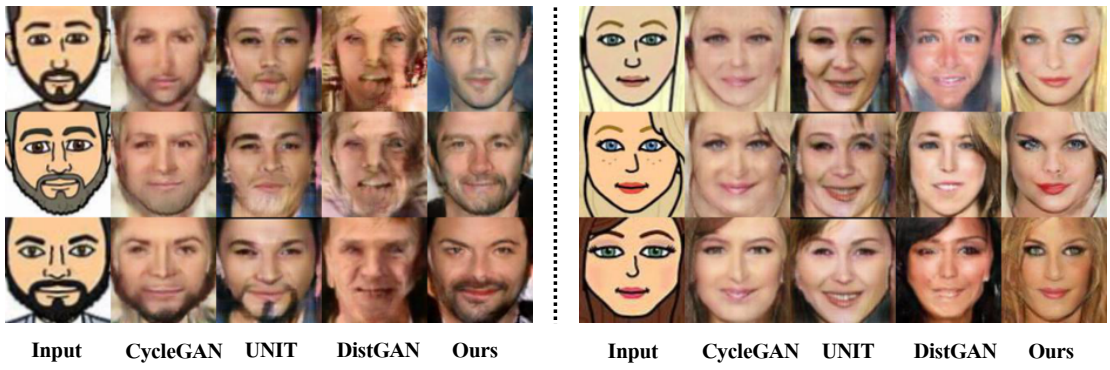

Figure 5: Challenging characteristics: the facial attributions, e.g. beard, lipstick, eyelashes are preserved by CP-GAN.

4. Experiments

We organize the experiments as follows. First, we provide the implementation details. Second, we introduce data collection and baselines in Section 4.2 and 4.3, respectively. Third, in Section 4.4, we compare the performance of proposed method with several related baselines. Fourth, we conduct an ablation study on the effect of global and local discriminators in Section 4.5. Last, we investigate the effect of the content loss in Section 4.6.

4.1. Implementation Details

Our approach is based upon the promising fast neural style transfer architecture (Johnson et al., 2016a). We adopt the Light CNN-9 model (Wu et al., 2015) as the specific content network for preserving the face identity, which is trained on MS-Celeb-1M (Guo et al., 2016). We choose the transformer network in (Johnson et al., 2016a) as our generator, which includes two convolutions with stride two, five residual blocks (He et al., 2016) and two specific up-sampling layers. Global and local discriminators have the same network that mainly consists of four blocks, each of which contains a stride-2 convolution, Instance Normalization and a leaky ReLU with the parameter of 0.2. A batch size of 1 and a learning rate of 10^{-4} are adopted for all our experiments. Empirically, we set $\lambda = 10^{-1}$ and $\gamma = 10^{-2}$. Our method is implemented in PyTorch.

4.2. Data Collection and Evaluation Metrics

For cartoon-to-photo facial translation task, we collect two datasets containing cartoon faces and real faces. For cartoon images, we select 10,000 facial cartoon avatars created by an online service(bitmoji.com) and label their 5 landmarks (left eye, right eye, nose, left mouth, left and right mouth corners). In order to obtain real face images, we extract $66,382$ frontal faces from CelebA dataset (Liu et al., 2015). Additionally, to alleviate the artifacts caused by various backgrounds or other factors, each face in these two datasets is aligned and cropped as $128 \times 128 \times 3$ size based on the 5 landmarks.

For quantitative evaluation, we divide the cartoon dataset into a training set (containing 8500 images) and a testing set (containing 1500 images), and adopt Fréchet Inception Distance (FID) (Heusel et al., 2017) as well as MS-SSIM (Odena et al., 2017) to evaluate the qualities of samples transfered by the testing images. Specifically, FID is able to capture the similarity of generated images to real ones. In general, a smaller value of FID means better performance. MS-SSIM measures the diversity of generated samples and the value ranges from 0.0 to 1.0. Higher MS-SSIM values correspond to perceptually more similar images.

We further conduct a human study via Amazon Mechanical Turk (AMT). For each method, we randomly choose 700 cartoon images and the corresponding generated photos, and ask annotators on AMT to pick up the best one that preserves the identity information. Each test case is scored 5 times, resulting in 3,500 human judgments for each method. Finally, we record the percentage of the human judgments where the result is preferred over the other methods.

Table 1: Quantitative results of different methods.

	Cartoon	UNIT	CycleGAN	Distance GAN	CP-GAN
FID	313.73	58.05	81.55	75.62	**31.60**
MS-SSIM	0.3238	0.5420	0.4357	**0.2972**	0.3169

Table 2: Human evaluation on identity perserving.

	UNIT	CycleGAN	Distance GAN	CP-GAN
Similarity	28%	17%	8%	**31%**

4.3. Baselines

To evaluate the performance of the proposed method, three state-of-the-art unpaired image-to-image models are adopted as baselines, including UNIT (Liu et al., 2017), CycleGAN (Zhu et al., 2017), and Distance GAN (Benaim and Wolf, 2017).

- **UNIT (Liu et al., 2017)** This method combines generative adversarial networks and variational auto-encoders based on Coupled GANs (Liu and Tuzel, 2016) to learn a shared latent representation using a weight sharing constraint.
- **CycleGAN (Zhu et al., 2017)** To learn common information between X and Y domains, this approach uses two translators $G : X \to Y$ and $F : Y \to X$ trained simultaneously and adds a cycle constraint that encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$.
- **Distance GAN (Benaim and Wolf, 2017)** DistanceGAN is a method of learning $G : X \to Y$ without learning $F : Y \to X$ by maintaining the distance between different parts of the same sample before and after translation.

4.4. Comparison with State of the Art

In this section, we compare CP-GAN with other state-of-the-art methods. Both visual quality and quantitative comparisons are conducted to evaluate the performance of the proposed method.

We first compare the visual quality of the synthetic images generated by CP-GAN with three baseline methods. As shown in Figure 4, all methods can produce meaningful images and preserve rough facial structures. However, the samples generated by CycleGAN lack photo-realistic details and look blurring. UNIT tends to produce similar faces and Distance GAN fails to preserve identity. Compared to the state-of-the-art methods, CP-GAN is able to produce plausible faces with more and better facial details. Moreover, the identity of the given cartoon face can be preserved on the generated face image.

The beard is one of the obvious male facial characteristics. Except for preserving identity, we still need to capture and translate beard on male cartoon face. As shown in Figure

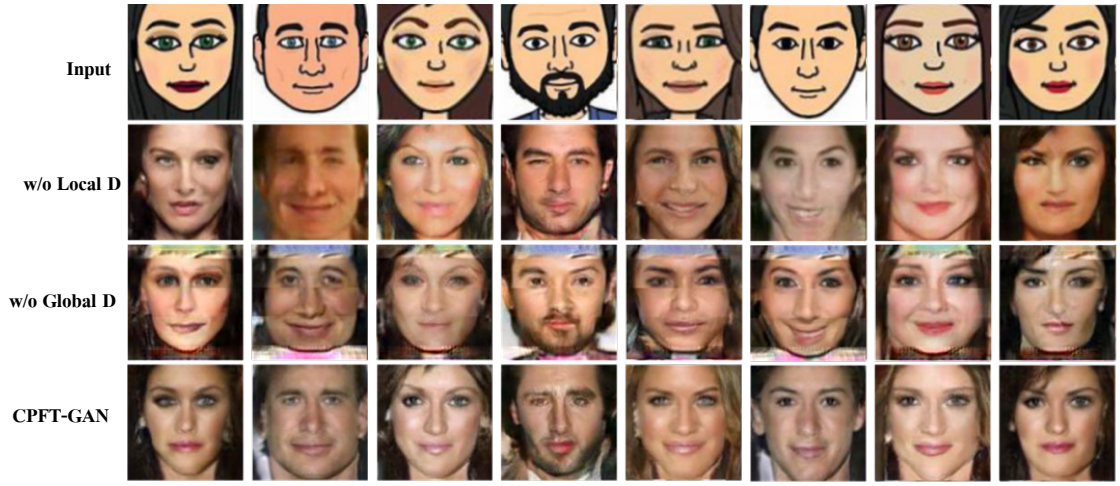

Figure 6: Ablating comparison: the generated results of CP-GAN and its variants.

Table 3: FID and MS-SSIM scores of ablation study.

	w/o Global D	w/o Local D	CP-GAN
FID	113	54.87	**31.60**
MS-SSIM	0.33	0.43	**0.31**

5 (left), compared to the baselines, our method can successfully preserve the local feature of beard without loss of quality and identity. For the women, the makeup can be considered as a necessity. Therefore, we aim to generate female faces with good makeup effects. In Figure 5 (right), benefits from local D, the makeups of mouth and eyes (lipstick, the color of pupils, eyelashes) on female cartoon face can be preserved and transferred to the relevant ones in the realistic world. These results demonstrate the effectiveness of CP-GAN when transforming challenging samples.

We further evaluate the performance of CP-GAN using FID (Heusel et al., 2017) and MS-SSIM (Odena et al., 2017). The scores of different methods are shown in Table 1, where the method "cartoon" means that the results are evaluated on the cartoon images. From the results in Table 1, CP-GAN achieves the best performance in terms of FID, which indicates that the samples generated by CP-GAN are more similar to the real faces. Distance GAN has the lowest value of MS-SSIM, but its FID score is relatively large. It means that although Distance GAN keeps the largest diversity, the qualities of the generated faces are limited. Differently, CP-GAN produces convincing faces while keeping a comparable large diversity.

Moreover, we conduct a perceptual experiment using Amazon Mechanical Turk platform, comparing the proposed method to other baselines. In Table 2, each cell lists the percentage of the human judgments where the result is preferred over the other methods. This human evaluation shows that our method outperforms other methods in term of identity preserving.

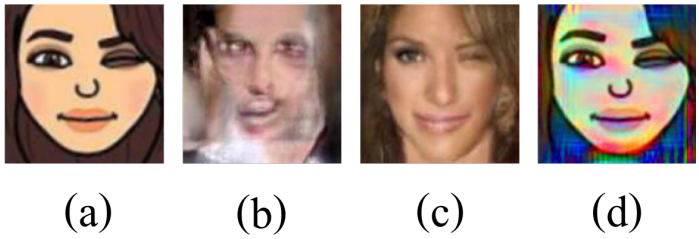

Figure 7: Discussion of the content loss: (a) input, (b) the output when $\gamma = 0.001$, (c) the output when $\gamma = 0.01$, (d) the output when $\gamma = 0.1$.

4.5. Analysis of the Global and Local D

In this section, we perform experiments to evaluate the effectiveness of the proposed global and local discriminators in cartoon-to-photo facial translation. Firstly, we train the generator G just using global D. Next, we replace global D with local D to update G. Finally, the generator G is updated by both global and local D.

As shown in Figure 6, when only using global D, some parts of produced faces collapse and some faces tend to be blurry. With regards to the results generated just using local D, each part of the generated faces is intact and clear, but the skin texture of these faces are fractured between two patches. Especially, the first and last patches are hardly updated by local D and still keep the textures of given cartoon faces. Because our clipping and padding scheme ignores these two patches and just focuses on the main three patches. The results of both using global and local D illustrate that global and local discriminators encourage G to generate the results which achieve high face quality with intact facial parts.

In Table 3, the model without global D has the highest the FID score since its generated faces consist of fractured patches. CP-GAN has both the lowest FID and MS-SSIM scores, which demonstrates that local GAN can help global GAN produce more convincing results.

4.6. Analysis of the Content Loss

The comparison results using different γ are shown in Figure 7. Furthermore, in Figure 8, we present the tendencies of the content loss using different γ during training. In all three experiments, we set $\lambda = 0.1$. In Figure 8, when $\gamma = 0.01$, the content loss is remained constant at 1.00. The main trends of the content loss using $\gamma = 0.001$ and $\gamma = 0.1$ are opposite with divergence and convergence, respectively. According to the results shown in Figure 7, we observe that the produced face using $\gamma = 0.001$ loses the content information since the tendency presented on Figure 8 is divergent. On the contrary, the generated face using $\gamma = 0.1$ is highly similar to the input, because the content loss is convergent. These illustrate that if γ is too small, the results would lose the identity of the cartoon faces, while if γ is too large, the effect of adversarial component could be ignored. Empirically, only keeping the content loss roughly constant can make the content network and adversarial component work cooperatively, which can produce convincing results.

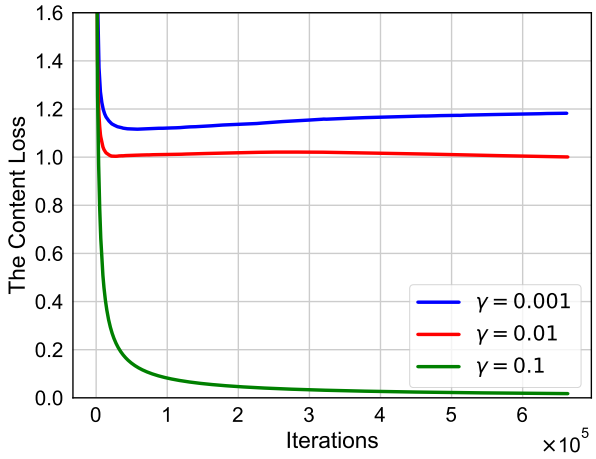

Figure 8: Evolution of content loss with different γ.

5. Conclusions

In this paper, we propose CP-GAN to tackle the problem of cartoon-to-photo facial translation. The proposed global discriminator encourages the generator to produce faces with high-quality global features, and the proposed local discriminator recovers the collapsed facial regions and makes the generated faces sharper. In order to preserve the characteristics and identities of cartoon faces, we exploit a face recognition model as the content network, instead of learning the preserved common information from scratch. As a result, our network can generate high-quality faces with intact facial regions and preserve the identity information in the given cartoon faces. Experimental results demonstrate that our proposed approach is able to generate convincing images and outperform the state-of-the-art methods.

6. Acknowledgements

This work was partially supported by National Natural Science Foundation of China (NSFC) 61602185, 61401163, 61876208, 61502177 and 61602185, Recruitment Program for Young Professionals, Fundamental Research Funds for the Central Universities D2172480, and Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183 and Guangdong Provincial Scientific and Technological funds 2017B090901008, 2017A010101011, 2017B090910005, and Pearl River S&T Nova Program of Guangzhou 201806010081 and CCF-Tencent Open Research Fund RAGR20170105.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Jiezhang Cao, Yong Guo, Qingyao Wu, Chunhua Shen, Junzhou Huang, and Mingkui Tan. Adversarial learning with local coordinate coding. In *International Conference on Machine Learning (ICML)*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, 2016.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

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 *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*. Springer, 2016a.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*. Springer, 2016b.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations (ICLR)*, 2018.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *International Conference on Machine Leraning (ICML)*, 2017.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. *CoRR, abs/1703.07511*, 2017.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *International Conference on Machine Leraning (ICML)*, 2017.

Guo-Jun Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2016.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *International Conference on Machine Learning (ICML)*, 2016.

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *International Conference on Learning Representations (ICLR)*, 2017.

Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xiang Wu, Ran He, and Zhenan Sun. A lightened cnn for deep face representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.