Unpaired Photo-to-Caricature Translation on Faces in the Wild

Ziqiang Zheng[a], Chao Wang[a], Zhibin Yu[a], Nan Wang[a], Haiyong Zheng[a,b,*], Bing Zheng[a]

[a]*Department of Electronic Engineering, Ocean University of China, Qingdao 266100, China*
[b]*Department of Mathematics, School of Science and Engineering, University of Dundee, Dundee DD1 4HN, U.K.*

Abstract

Recently, image-to-image translation has been made much progress owing to the success of conditional Generative Adversarial Networks (cGANs). And some unpaired methods based on cycle consistency loss such as DualGAN, CycleGAN and DiscoGAN are really popular. However, it's still very challenging for translation tasks with the requirement of high-level visual information conversion, such as photo-to-caricature translation that requires satire, exaggeration, life-likeness and artistry. We present an approach for learning to translate faces in the wild from the source photo domain to the target caricature domain with different styles, which can also be used for other high-level image-to-image translation tasks. In order to capture global structure with local statistics while translation, we design a dual pathway model with one coarse discriminator and one fine discriminator. For generator, we provide one extra perceptual loss in association with adversarial loss and cycle consistency loss to achieve representation learning for two different domains. Also the style can be learned by the auxiliary noise input. Experiments on photo-to-caricature translation of faces in the wild show considerable performance gain of our proposed method over state-of-the-art translation methods as well as its potential real applications. Our code is available at `https://github.com/zhengziqiang/P2C`.

*Corresponding author
Email address: `zhenghaiyong@ouc.edu.cn` (Haiyong Zheng)

Keywords: Generative Adversarial Network, Image-to-image translation, Photo-to-caricature translation, Dual discriminators, Unpaired image translation

1. Introduction

Image-to-image translation has been made much progress [1, 2, 3, 4, 5, 6, 7] because many tasks [8, 9, 10, 11] in image processing, computer graphics, and computer vision can be posed as translating an input image into a corresponding output image [12, 13, 14, 15, 16, 17, 18]. And its achievements mainly owes to the success of Generative Adversarial Networks (GANs) [19], especially conditional GANs (cGANs) [20, 21, 1]. However, the current studies mainly concern image-to-image translation tasks with low-level visual information conversion, *e.g.*, photo-to-sketch [22].

A caricature is a rendered image showing the features of its subject in an exaggerated way and usually used to describe a politician or movie star for political or entertainment purpose. Creating caricatures can be considered as artistic creation tracing back to the 17th century with the profession of caricaturist. Then some efforts have been made to produce caricatures semi-automatically using computer graphics techniques [23], which intend to provide warping tools specifically designed toward rapidly producing caricatures. But there are very few software programs designed specifically for automatically creating caricatures, and to the best of our knowledge, none can work to be comparable with caricaturist. Nowadays, besides the political and public-figure satire, caricatures are also used as gifts or souvenirs, and more and more museums dedicated to caricature throughout the world were opened. So it would be very useful and meaningful if computers can create caricatures from photos automatically and intelligently.

Photo-to-caricature is a typical high-level image-to-image translation problem but with bigger challenge than other low-level translation problems such as photo-to-label, photo-to-map, or photo-to-sketch [1], because caricatures

- require **satire** and **exaggeration** of photos;
- need **artistry** with different styles;
- must be **lifelike**, especially the expression of a face photo.

Specifically, for a face photo, we want to create the face caricatures with different styles, which exaggerate the face shape or facial sense organs (*i.e.*, ears, mouth, nose, eyes and eyebrow) but keep the vivid expression while producing artistry.

In this paper, we propose a GAN-based method for learning to translate faces in the wild from the source photo domain to the target caricature domain (see Figure 1 for translating examples and Figure 2 for the architecture of our proposed method). Although deep convolutional neural networks with adversarial training [24] can generate images with enough precise facial features [25], these images sometimes still have wrong relationships between facial features or mismatch among facial organs such as nose and eyes, *e.g.*, a face with more than two eyes or crooked nose. Traditional GANs can produce correct facial organs but wrong relationships between them, and also it's very challenging to abstract and exaggerate face and facial organs. We attribute this problem to the deficient capacity of the discriminator of GAN to distinguish real-fake images. Therefore, our motivation is to design an adversarial training with multiple discriminators to improve the ability of GAN's discriminator for feature representation.

Based on the model of CycleGAN [2], we design a dual pathway model of GAN for high-level image-to-image translation tasks, where one pathway of coarse discriminator is in charge of abstracting the global structure information, while another pathway of fine discriminator is responsible for concerning the local statistics information. For generator, besides the adversarial loss, we provide one more extra perceptual similarity loss to constrain consistency for generated output itself and with the unpaired target domain image. By using our proposed method, the photos of faces in the wild can be translated to caricatures with learned general-purpose exaggerated artistic styles while still keeping the original lifelike expression (see Figure 1 and 11 for references). Considering that traditional GANs are not robust and easily attracted by noise, we design

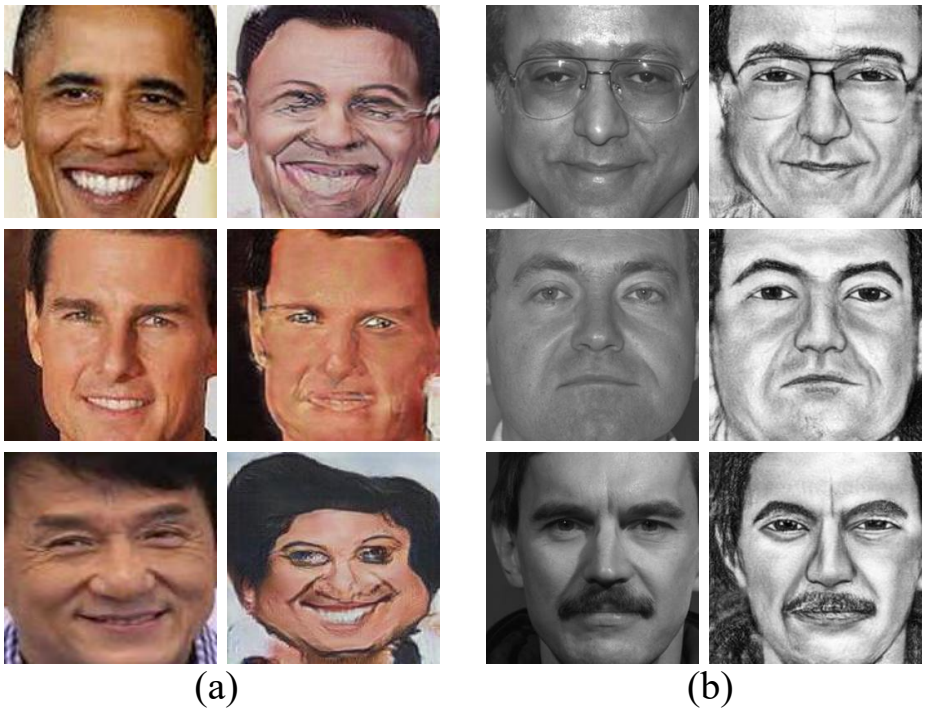

Figure 1: Translating faces in the wild from photo to caricature with different styles by our proposed method. (a) Example results on IIIT-CFW dataset [26]; (b) Example results on PHOTO-SKETCH dataset [27, 28].

a noise-added training procedure to improve the robustness of our model. Inspired by InfoGAN [29], we find that auxiliary noise can help model learning the caricature style information while translating images in our task.

We have extensively evaluated our method on IIIT-CFW dataset [26], PHOTO-SKETCH dataset [27, 28], Caricature [30], FEI dataset [31], Yale dataset [32], KDEF dataset [33] and CelebA dataset [34]. The experimental results show that our method can create acceptable caricatures from face photos while current state-of-the-art image-to-image translation methods can't. Also the designed experiments indicate the effectiveness of our proposed dual pathway of discriminators, additional noise input and extra perceptual loss, respectively. Besides, we tested our photo-to-caricature translation method for producing caricatures with adding different proportions of noise to show the translating robustness

and style diversity. Furthermore, the proposed method can create caricatures for arbitrary face photos without pre-training on extra face datasets. Another prominent performance of our methods is that our model can capture the expression information and make some abstraction and exaggeration. This might be helpful to fill aforementioned gap of automatic and intelligent caricature creation.

2. Related Work

Image-to-image translation. Owing to the success of GANs [19], especially various conditional GANs [20, 35, 36, 37, 14, 38], image-to-image translation problems have been made much progress recently, which aims to translate an input image from one domain to another domain given input-output images pairs [1]. Earlier image-conditional models for specific applications have achieved impressive results on inpainting [39], de-raining [38], texture synthesis [13], style transfer [14], video prediction [35] and super-resolution [15]. The general-purpose solution for image-to-image translation developed by Isola *et al.* [1] with the released `pix2pix` software has achieved reasonable results on many translation tasks by using paired images for training such as photo-to-label, photo-to-map and photo-to-sketch. Then CycleGAN [2], DualGAN [3], and DiscoGAN [5] were proposed for unpaired or unsupervised image-to-image translation with almost comparable results to paired or supervised methods. However, the translation tasks that these work can tackle usually concern the conversion of low-level visual information such as line (photo→sketch), color (summer→winter), and texture (horse→zebra), but it's still very challenging for some translation tasks with the requirement of high-level visual information conversion, *e.g.*, abstraction and exaggeration (photo→caricature). Recently, Iizuka *et al.* [40] combined two discriminators called global discriminator and local discriminator to improve the adversarial training for image completion task. Experimental results have proven that the global discriminator can distinguish the images based on the global parts while the local discriminator pays

attention to the details of parts. We exploit the design of two discriminators for high-level image-to-image translation tasks in this paper.

Photo-to-cartoon translation. Translating photo to cartoon by computer algorithms has been studied for a long time because of the corresponding interesting and meaningful applications. The earlier work mainly relied on the analysis of facial features [41, 42, 43], which is hard to be applicable for large-scale faces in the wild. Thanks to the invention of GANs [19], automatic photo-to-cartoon translation becomes feasible. But most of the current related work mainly focused on the generation of anime [44] or emoji [16] with specific style[1]. And these works have nothing to do with caricature creation that needs to be exaggerated, lifelike and artistic.

Unlike the prior works for image-to-image translation dealing with low-level visual information conversion, our study mainly focuses on the translation problems of high-level visual information conversion, *e.g.*, photo-to-caricature. For this purpose, our method differs from the past work in network architecture as well as the layers and losses. We design a dual pathway model of GAN with two discriminators named coarse discriminator and fine discriminator to capture global structure and local statistics respectively, and we apply the perceptual similarity loss for generator. For learning the style information and improving the robustness of our model, we provide the input with auxiliary noise. Here we show that our approach is effective on the face photo-to-caricature translation task, which typically requires high-level visual information conversion.

3. Method

Conditional GANs are generative models that learn a mapping from observed input image x, to target image y, $G : \{x\} \rightarrow y$. The objective of our conditional

[1] We consider that anime, emoji, and caricature are three types of cartoon or three subsets of cartoon.

GAN can be expressed as

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) =& \mathbb{E}_{x,y\sim P_{data}(x,y)}\left[\log D(x,y)\right] + \\ & \mathbb{E}_{x\sim P_{data}(x)}\left[\log(1-D(x,G(x)))\right].\end{aligned} \tag{1}$$

The goal is to learn a generator distribution over data y that matches the real data distribution P_{data} by transforming an observed input image $x \sim P_{data}(x)$ into a sample $G(x)$. This generator is trained by playing against an adversarial discriminator D that aims to distinguish between samples from the true data distribution and the generator's distribution. To deal with the high-level visual information conversion for some more challenging image-to-image translation tasks, we exploit one discriminator to dual discriminators and add two more losses (perceptual loss and cycle consistency loss) to adversarial loss, so our method can tackle conversions of both the low-level (line, color, texture, *etc.*) and the high-level (expression, exaggeration, artistry, *etc.*) visual information. And in order to improve the robustness of our model, we provide our model a noise-added training and use auxiliary noise to learn the style information while translating. The overall network architecture and data flow are illustrated in Figure 2.

3.1. Network architecture

We adopt our generator architecture from [2], which use the modules of the form `convolution-InstanceNorm-ReLu` (Conv block), `deconvolution-InstanceNorm-ReLu` (Deconv block) and residual blocks. As for the discriminator, we adopt the PatchGAN from [1] and use modules of the form `convolution-InstanceNorm-LeakyReLu` (Conv block). Here we set leak 0.2 for our discriminators. Different from [1], we do not add a Sigmoid activity function at the last layer. We add a Tanh activity function at the last of the generator to normalize the generated image to $[-1, 1]$. As shown in Figure 2 with the details of our network architecture (the size of feature maps of every layers), we set convolution kernel size 3 and stride 2 for our generator except the first convolution layer and the last deconvolution layer with kernel size 7 and stride 1. We set convolution kernel size 4

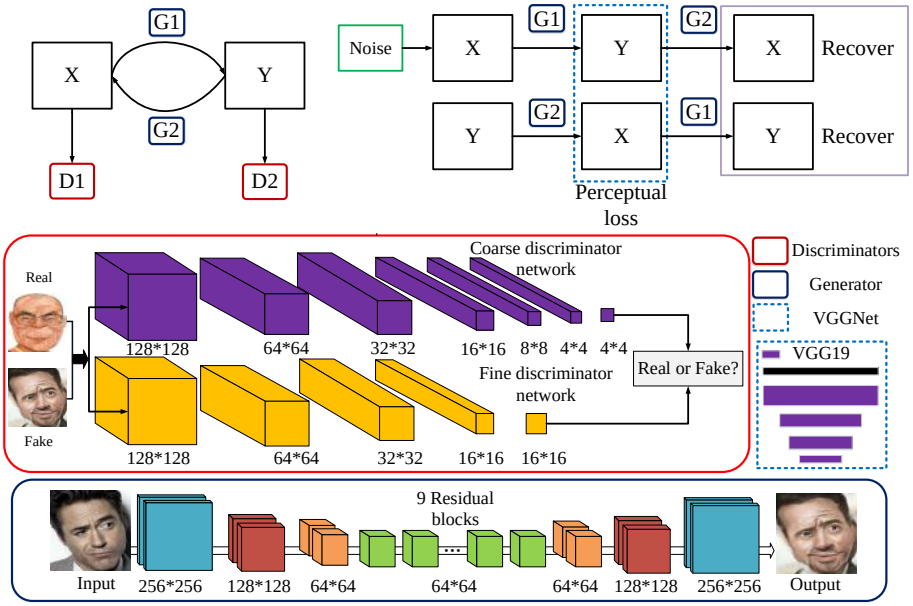

Figure 2: Network architecture and data flow chart of our proposed method for face photo-to-caricature translation.

and stride 2 for both coarse discriminator and fine discriminator. Let `Ck` denote a form `convolution-InstanceNorm-ReLu` with `k` filters, `CDk` denote a form `deconvolution-InstanceNorm-ReLu` with `k` filters, `Rk` denote a residual block with `k` filters, and `Dk` denote a form `convolution-InstanceNorm-LeakyReLu` with `k` filters for discriminators, the network architecture consists of:

Generator:

`C64-C128-C256-R256-R256-R256-R256-R256-R256-R256-R256-R256-CD128-CD64`

Coarse discriminator: `D64-D128-D256-D512-D512-D512`

Fine discriminator: `D64-D128-D256-D512`

We also show the details of network architecture in Tables 1, 2 and 3.

3.2. Cycle consistency loss

Researchers apply adversarial training to learn the mapping function between two different image domains. Here we use two same generators (named

Table 1: The netwrok architecture of generator.

Type	Kernel size	Stride	Output
Conv Block	7	1	64
Conv Block	3	2	128
Conv Block	3	2	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Residual Block	3	1	256
Deconv Block	3	2	128
Deconv Block	3	2	64
Deconv	7	1	3

`G1` and `G2` in Figure 2) for observing the sample distribution of two domains. And X denotes the photo domain while Y denotes the caricature domain in our task. We use our generators to emulate the translation between X and Y. However, only using adversarial loss can't guarantee a plausible generation. So researchers use Cycle Consistency Loss ($\mathcal{L}_{cyc}$) to help establishing mapping function between domains [2]. And $\mathcal{L}_{cyc}$ is expressed as

$$\begin{aligned}\mathcal{L}_{cyc}(G_1, G_2) =& \mathbb{E}_{x\sim P_{data}(x)}\left[||G_2(G_1(x)) - x||_1\right] + \\ & \mathbb{E}_{y\sim P_{data}(y)}\left[||G_1(G_2(y)) - y||_1\right],\end{aligned} \tag{2}$$

where G_1 and G_2 represent the two generators, and x and y are samples from X and Y domain respectively. We use $L1$ loss for the cycle consistency loss following Zhu *et al.* [2].

Table 2: The network architecture of coarse discriminator.

Type	Kernel size	Stride	Output
Conv+Leaky ReLU	4	2	64
Conv Block	4	2	128
Conv Block	4	2	256
Conv Block	4	2	512
Conv Block	4	2	512
Conv Block	4	2	512
Conv	4	1	1

Table 3: The network architecture of fine discriminator.

Type	Kernel size	Stride	Output
Conv+Leaky ReLU	4	2	64
Conv Block	4	2	128
Conv Block	4	2	256
Conv Block	4	2	512
Conv	4	1	1

3.3. Perceptual loss

To further reduce the space gap of possible mapping functions between domains, we apply the perceptual loss $\mathcal{L}_p$ for our model. For a constrained translation problem, finding an appropriate loss function is critical. We adopt the content loss of Gatys *et al.* [45], which is also referred as a perceptual similarity loss or feature matching [46, 47, 48, 49]. We apply the perceptual loss to our model followed by the cycle consistence loss, and compute the perceptual loss between unpaired images from different domains to push generator to capture the feature representations. Let Φ denotes a pre-trained visual perception network (we use pre-trained `VGG19` in our experiments) and n denotes the number of feature maps. Different layers in the networks represent low-to-high level information: from edges and color to object and semantic representation. Matching both low and high layers in the perception network can help achieving

fantastic translation. And the perceptual loss $\mathcal{L}_p$ can be expressed as

$$\mathcal{L}_p = \sum_{n}^{N} \lambda_n ||\Phi_n(y) - \Phi_n(G(x);\theta))||_1, \tag{3}$$

where y denotes the image from caricature domain in our task, $G(x)$ denotes the synthesized caricature image using x, θ means the corresponding parameters. The hyper parameters λ_n are used for balance and we set them following [4]. Note that these two images are unpaired.

3.4. Auxiliary noise input

Previous researches have fully proved that we can get plausible image results from noise input [21, 29, 50, 25]. In order to improve the robustness and enrich the diversity of image translation between domains, we design a noise-added training procedure before the translation shown in Figure 2. First we obtain a random noise input from random uniform distribution (range from 0 to 255), then we merge the noise input and the raw image input to acquire the final input using approximate weights. Here we define α to denote the proportion of the raw image accounting for the final image. This can be expressed as

$$x = x_i * \alpha + (1 - \alpha) * \tau, \quad 0 \leq \alpha \leq 1, \tag{4}$$

where x_i denotes the raw input from photo domain and τ denotes the noise that has a uniform distribution P_{noise}. Figure 3 shows one sample with adding noise. With auxiliary noise input, the GAN object is expressed as:

$$\begin{aligned} \mathcal{L}(G, D) =& \mathbb{E}_{x,y\sim P_{data}(x,y)} \left[\log D\left(x, y\right)\right] + \\ & \mathbb{E}_{x\sim P_{data}(x),\tau\sim P_{noise}(\tau)} \left[\log(1 - D(x, G(x, \tau))\right]. \end{aligned} \tag{5}$$

We add the auxiliary noise to improve the robustness of our model. And we can get different styles of output through adjusting the noise input (see Section 4.5). Besides, we add the noise from a uniform distribution for making the style information more representable and matching the style space while translating. Furthermore, we also hope to control the style by adding different noise and make the translating conditional, which will be our future work.

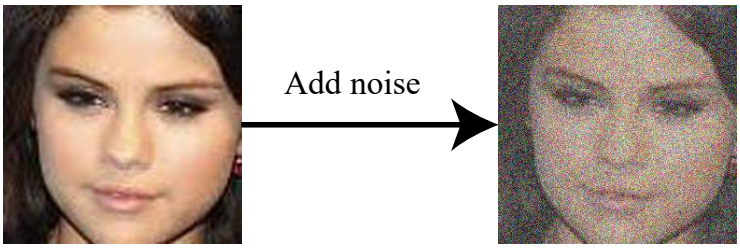

Figure 3: Sample for adding noise, here we use $\alpha = 0.5$.

3.5. Dual discriminators

Traditional GANs usually have one generator and one discriminator to leverage adversarial training. Different from them, we design two different discriminators to capture different level information.

In our method, we propose two different discriminators called coarse discriminator and fine discriminator respectively. The coarse discriminator aims to encourage generator to synthesize images based on global style and structure information for domain translation. In our high level image-to-image task, the coarse discriminator is capable for capturing the structure information and abstracting the representative information of face photo such as emotion and style. While the fine discriminator aims to achieve the feature matching and help generating more plausible and precise images, and the fine discriminator builds the adversarial training for the face details with generator such as lip and eyes. Different from Satoshi *et al.* [40], we are not using the image patches as input of local discriminator, we provide both the two discriminators with the whole image as input while the outputs of the two discriminators are different (see Figure 2). The output of coarse discriminator is a 4×4 patch matrix after the sigmoid activity function while the output of fine discriminator is 16×16. Note that both the two discriminators are using sigmoid function at the last layer. We have tried using different output size combinations to get different results, and the experiments show that the combination of 16×16 for fine discriminator and 4×4 for coarse discriminator can obtain the best result for translation. The coarse discriminator has a smaller feature map with more ab-

stractive representation compared to the fine discriminator. The `D1` and `D2` in Figure 2 represent the dual discriminators for translating two different domains X and Y.

3.6. Generator

Previous studies have found it beneficial to mix the GAN objective with a more traditional norm, such as $L2$ [39] and $L1$ [1] distance. We explore this opinion to cycle consistency loss by applying $L1$ distance to compute L_{cyc}. Our final objective is

$$G^* = \arg\min_G \max_D \mathcal{L}(G, D) + \gamma \mathcal{L}_{cyc} + \sigma \mathcal{L}_p, \tag{6}$$

$$\text{w.r.t} \quad \mathcal{L}(G, D) = \begin{cases} \mathcal{L}_c(G, D_c) & \text{for } D_c, \\ \mathcal{L}_f(G, D_f) & \text{for } D_f, \end{cases} \tag{7}$$

where $\mathcal{L}_{cyc}$ means cycle consistency loss, $\mathcal{L}_p$ indicates perceptual loss, D_c and D_f denote coarse and fine discriminators while $\mathcal{L}_c$ and $\mathcal{L}_f$ represent the corresponding objective functions respectively, γ and σ are hyper parameters to balance the contribution of each loss to the objective. As to $\mathcal{L}_{cyc}$, γ controls the content consistency for reconstruction so that it cannot be very small. As to $\mathcal{L}_p$, σ controls the style exaggeration for translation and it cannot be very large for keeping reasonable transformation. We use greedy search to optimize the hyper parameters with $\gamma = 10$ and $\sigma = 2.0$ for all the experiments in this paper.

As shown in Figure 2, we use Conv-Residual blocks-Deconv [51] as the generator to share the low-level and high-level information between the input and output directly across the net.

4. Experiments

As a typical image-to-image translation task, photo-to-caricature requires high-level visual information conversion, which is very challenging for the state-of-the-art general-purpose solutions. To explore the effect of our proposed

model, we tested the method on a variety of datasets for translating faces in the wild from photo to caricature with different styles, and the qualitative results are shown in Figures 4 and 11.

4.1. Dataset and training

Our proposed model is trained in a supervised unpaired fashion on a paired face photo-caricature dataset, named IIIT-CFW-P2C dataset, which was rebuilt based on IIIT-CFW [26]. The IIIT-CFW is a dataset for the cartoon faces in the wild and contains 8928 annotated cartoon faces of famous personalities in the world with varying profession. Also it provides 1000 real faces of the public figure for cross modal retrieval tasks. However, it's not suitable for the training of photo-to-caricature translation task using some paired methods (such as pix2pix) because the face photos and face cartoons[2] are not paired, *e.g.*, the facial orientation and expression of the photo and caricature for the same person are varying a lot. So we rebuild a photo-caricature dataset with 1171 paired images by searching the IIIT-CFW dataset and Internet as the training set for compared experiments. Here we use 800 for training and the left for testing. At inference time, we run the generator in exactly the same manner as during the training phase. Besides, we also extensively evaluated our method on a variety of datasets with faces in the wild, including Caricature [30], FEI [31], IIIT-CFW [26], Yale [32], KDEF [33] and CelebA [34].

Besides, we also consider photo-to-sketch as a photo-to-caricature task for experiments using PHOTO-SKETCH dataset [27, 28], which has 1194 paired images and hence can be directly used for supervised training of some compared paired methods. And following DualGAN, we use 995 unpaired images for training and 199 for testing. Note that we train CycleGAN, DualGAN and our model using unpaired images and train pix2pix using paired images of the two datasets.

For training and optimization, we use Adam optimizer, and set the learning

[2]We consider that caricature is a type of a cartoon or a subset of a cartoon.

rate 0.0002 and beta 0.5. Follow the CycleGAN, we use 9 residual blocks for the bottle neck of the generator. First we feed the input images to the two generators for generating fake images, then the coarse discriminator and fine discriminator try to distinguish the fake images and the real images. So we first optimize the generators and then optimize the coarse and fine discriminators. For the details of our generators and discriminators, please refer to Figure 2.

4.2. Comparison with state-of-the-arts

Using IIIT-CFW-P2C dataset, we first compare our proposed method with pix2pix [1], DualGAN [3], DiscoGAN [5], CycleGAN [2], and Fast Style Transfer [48] on photo-to-caricature translation task. Pix2pix is designed for paired image-to-image translation, while the others are developed for unpaired image-to-image translation. Only Fast Style Transfer is non-GAN method and the others are GAN-based methods. The idea behind the DualGAN, DiscoGAN, and CycleGAN is nearly the same, which uses two generators and two discriminators as the network structure, where one generator transforms images of domain X to domain Y, the other one do the opposite. Two discriminators D_X and D_Y try to distinguish real and fake images in each domain (fake means transformed from another domain). All the five methods were trained on the same training dataset and tested on novel data from IIIT-CFW-P2C dataset that does not overlap those for training. For Fast Style Transfer method, we randomly selected one caricature image as the style image following the instructions from the authors.

Qualitative evaluation. Figure 4 shows the experimental results of the comparison, it can be seen that, DualGAN only learned the color and edge translation rather than structure information, pix2pix makes structure error in almost all cases, while CycleGAN can keep the structure information of the input image but without enough conversion for caricature creation task. For DiscoGAN, it's really hard to generate plausible and meaningful results due to the lack data of training (800 pairs in our experiments vs. tens of thousands of pairs

in DiscoGAN's experiments) and the big challenge of task (photo-to-caricature vs. edge-to-photo). Fast Style Transfer method could not achieve exaggeration and abstraction between photo and caricature, and it only changes the color and texture information compared with the inputs. Although it's still not good enough for the results of our method compared to human caricaturists, the experiments on photo-to-caricature translation of faces show considerable performance gain of our proposed method over state-of-the-art image-to-image translation methods, especially the encouraging ability of exaggeration and abstraction. However, due to the very challenging task with less training data but on various styles, our method also messes some details while translation (see the mouth of the first row and the eyes of the fourth row on our results in Figure 4).

This experiment also expresses that high-level image-to-image translation tasks like photo-to-caricature are generally more difficult than those low-level translation tasks such as photo-to-sketch, because it not only needs to abstract the facial features, but also requires to exaggerate the emotional expression. So that the pixel-level methods (like pix2pix) might fail as they force the generator to concentrate on local information rather than whole structure.

Besides, we also evaluate our methods on PHOTO-SKETCH dataset. Figure 5 shows the compared results. Comparing with pix2pix, our method can reduce the effect of being blurry and artifact. Although DualGAN and CycleGAN can also reserve the structure information of input faces, they are not good at achieving abstraction and artistry, and this also happens to the Fast Style Transfer method. Similarly, DiscoGAN collapses when facing insufficient training data and big challenging task.

Quantitative evaluation. Beyond the visually qualitative evaluation, we also evaluated the translated cartoon results of different methods quantitatively on IIIT-CFW-P2C dataset and PHOTO-SKETCH dataset in terms of both human judge and machine grade, and the average results are shown in Table 4 and Table 5 respectively.

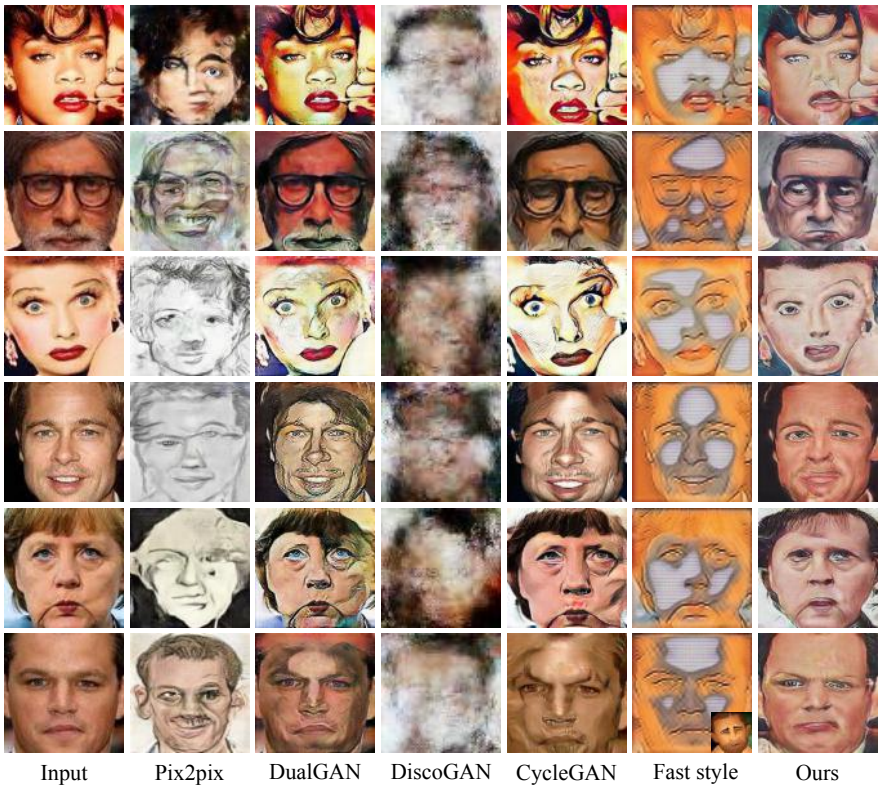

Figure 4: Comparison of state-of-the-art image-to-image translation methods with our proposed method for face photo-to-caricature translation on IIIT-CFW-P2C dataset.

For human score, we invited 40 volunteers to evaluate the generated image quality of different methods in terms of satire, exaggeration, lifelikeness and artistry compared with the given original photo, by grading from 1 to 5 (1 represents the worst and 5 represents the best). For inception score, we used a pre-trained classifier network and sampled images for evaluation followed [52]. The results show that our proposed method outperforms state-of-the-art image-to-image translation methods with the highest human and inception scores.

4.3. Dual discriminators

In this experiment, we verify the effectiveness of our proposed dual pathway of discriminators. We first use only one coarse discriminator (coarse D) and

Table 4: The generated image equality evaluation results of different methods on IIIT-CFW-P2C dataset.

Method	Human score	Inception score
Pix2pix	2.0106	1.5069±0.1090
DualGAN	1.9946	1.4843±0.1049
DiscoGAN	1.5014	1.3366±0.0714
CycleGAN	3.5001	1.5684±0.1331
Fast Style Transfer	-	1.0624±0.0325
Ours	4.0120	1.6043±0.0918

Table 5: The generated image equality evaluation results of different methods on PHOTO-SKETCH dataset.

Method	AMT	Inception score
Pix2pix	2.0971	1.3625±0.0706
DualGAN	2.3810	1.4063±0.0822
DiscoGAN	1.0858	1.3142±0.0206
CycleGAN	3.2272	1.3980±0.1130
Fast Style Transfer	-	1.2985±0.0421
Ours	4.0750	1.4298±0.0818

Figure 5: Comparison of state-of-the-art image-to-image translation methods with our proposed method for face photo-to-caricature translation on PHOTO-SKETCH dataset.

one fine discriminator (fine D) separately, and then use the two discriminators with different weights, while keeping all other architectures and settings fixed for training and testing. Some example results are shown in Figure 6, it can be seen that, one fine D model almost misses the key structure information on faces; if the coarse D weights more, the results are more blur, while if the fine D weights more, the results are more detailed; and our dual coarse D + fine D (`C1F1`) model can render the structure of facial features well with better details. It further proves that the fine D model only concerns the local statistics for tackling low-level image-to-image translation tasks (*e.g.*, photo-to-sketch).

Actually, using only one discriminator is hard to generate normal faces with good shape, and the generation results could be getting worse on distortion when the training was going on. However, coarse D and fine D, taking care

of coarse structure and fine details respectively, can constraint each other to generate better results with good shape, because these two discriminators build an adversarial training with generator and they can communicate with each other through the adversarial loss.

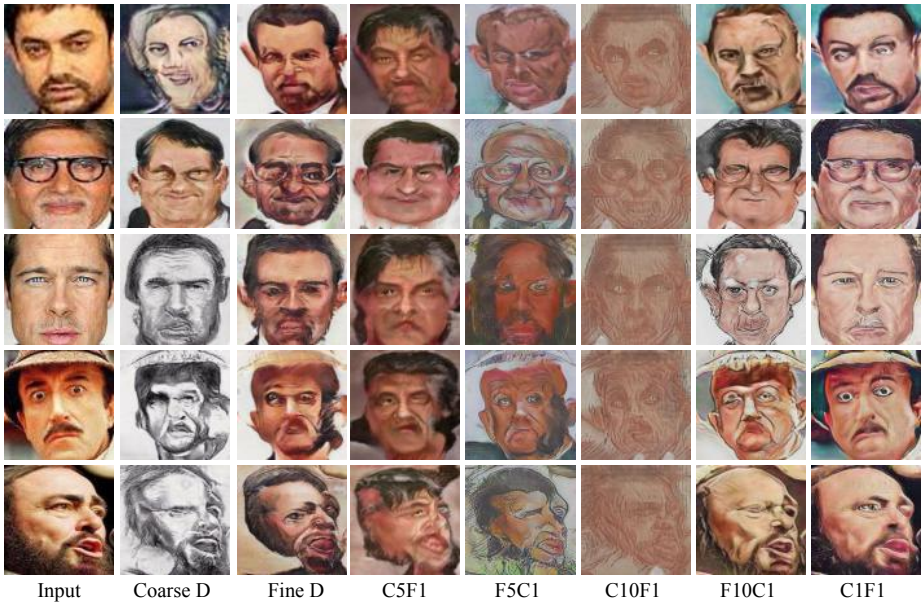

Figure 6: Comparison of using different weights of coarse discriminator (coarse D) and fine discriminator (fine D), where $CmFn$ and $FnCm$ represent the weight of coarse D and fine D is $m : n$.

And we also took some experiments to greedy search the best combination size of output patches for coarse discriminator and fine discriminator. Figure 7 shows results of different combinations, which indicates that large Fine D patches (*e.g.*, F32) fails to abstract and exaggerate faces while small Fine D patches (*e.g.*, F8) abstracts and exaggerates faces too much.

4.4. Loss selection

We first consider to check if the cycle consistency loss $\mathcal{L}_{cyc}$ should be provided to the GAN objective (Equation 6), and the second column in Figure 8 shows the results without cycle consistency loss. It's easy to see that without

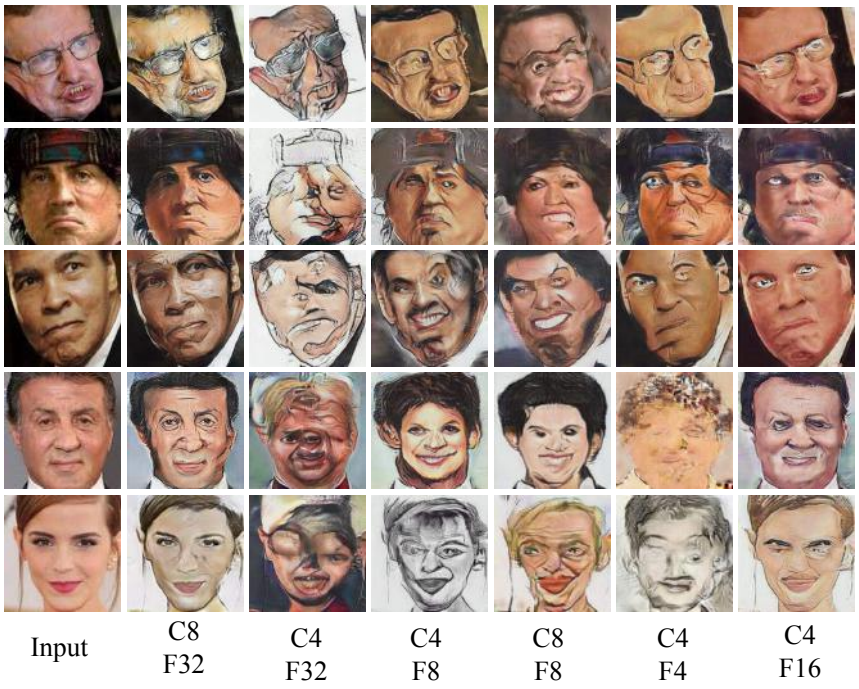

Figure 7: Examples of different combination sizes of output patches for coarse discriminator and fine discriminator. Ck denotes that the output of coarse discriminator is a $k \times k$ patch, and Fk denotes that the output of fine discriminator is a $k \times k$ patch. It can be seen that large Fk such as F32 for fine discriminator fails to abstract faces and achieve exaggerations, while small Fk such as F8 abstracts and exaggerates faces too much.

cycle consistency loss, although the adversarial system with adversarial loss can capture the facial features, it's hard to generate caricature images with plausible objects and meaningful relationship between facial organs. The third column in Figure 8 shows the results by adding cycle consistency loss $\mathcal{L}_{cyc}$. Therefore, the normal adversarial training can lead to some kind of caricature style, but it fails to be lifelike without meaningful components.

Based on the cycle consistency loss, we provide the perceptual loss $\mathcal{L}_p$ for generator in our system. Figure 9 shows the compared results of without and with $\mathcal{L}_p$. It can be seen that perceptual loss can produce images with the exaggerated facial features such as eyes, nose, and mouth. The perceptual loss,

Figure 8: Comparison of extra loss for final objective of generator: without (w/o) $\mathcal{L}_{cyc}$ and with ($w/$) $\mathcal{L}_{cyc}$.

which expresses some perceptual errors on facial features such as head, eyes, and mouth, could improve the artistic expression of image generation and show better abstraction ability. And it can also reduce the effect of being blurry. The second column of Figure 9 without using perceptual loss illustrates the indistinguishable facial expressions with distorted facial organs while translation, and the third column with adding perceptual loss improves the performance on facial expression and organ translation with caricature effect, *e.g.*, the smile woman of last row. Besides, the reason for transferring a feminine face into a masculine face in line 4 of Figure 9 might be the imbalanced training dataset of feminine faces (192 images) and masculine faces (608 images), so that the adversarial loss has more strong effect than cycle loss to make the translation more masculine. And the perceptual loss aims to achieve feature matching between two images,

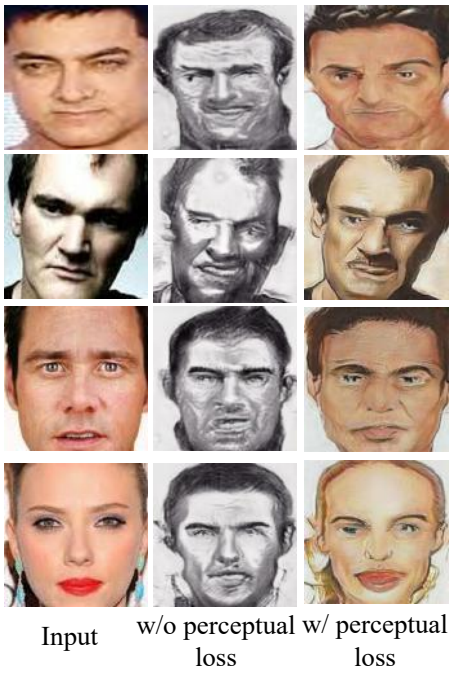

Figure 9: Comparison of extra loss for final objective of generator: without (w/o) $\mathcal{L}_p$ and with ($w/$) $\mathcal{L}_p$.

which can help the generator to focus on the high-level semantic features, thus it can weaken the influence of adversarial loss.

4.5. Auxiliary noise input

By adding auxiliary noise to our photo-to-caricature system, we can improve the robustness and diversity of synthesized facial caricatures. Figure 10 shows the example results of adding auxiliary noise for photo-to-caricature translation, and the output results for adding different proportions (α) of noise indicate that our system can still synthesize meaningful facial caricatures with even more than a half noise ($1 - \alpha$, see Equation 4 for reference) as inputs, besides, the added different proportions of noise also lead to different styles of output results, which indicates that it might be used as a factor for tuning different synthesized styles.

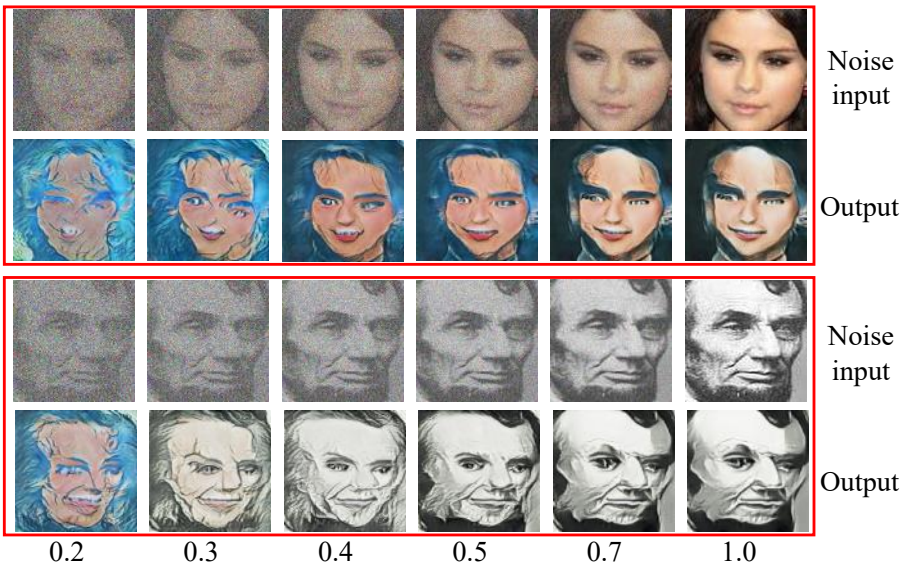

Figure 10: Examples of adding auxiliary noise for robustness and diversity of our photo-to-caricature translation. Different proportions (α in Equation 3) of noise inputs can also lead to meaningful different styles of output caricatures.

4.6. Freestyle face caricature creation

To evaluate the ability for real applications in the daily life, we tested our method on a variety of face datasets, including Caricature [30], FEI [31], IIIT-CFW [26], Yale [32], KDEF [33] and CelebA [34], to illustrate the photo-to-caricature translation on faces in the wild, and the results are shown in Figure 11. These freestyle face caricature creation results validate that our model works not bad on arbitrary faces and show the potential value for the related applications. And we can see that the translated results in KDEF, FEI and Yale datasets also have different facial expression corresponding the input faces. Our methods successfully reserve the emotion information and emulate the facial organs with caricature style. So we can conclude that the more abstracted information such as facial emotion and expression with global structure information are reserved. Besides, our model can also enlarge or narrow the facial organs such as chin, lips, eyes and so on, which is required for high-level image-to-image translation

tasks.

Figure 11: Example translated caricatures of facial photos from several datasets (Caricature [30], FEI [31], IIIT-CFW [26], Yale [32], KDEF [33] and CelebA [34]) using our trained model on IIIT-CFW-P2C.

5. Conclusion and Future Work

We present a novel GAN-based method to deal with high-level image-to-image translation task, *i.e.*, photo-to-caricature translation. The proposed method uses dual discriminators for capturing global structure and local statistics information with abstraction ability, also provides extra perceptual loss on GAN objective to constrain the consistency under exaggeration. Besides, the style information can be learned and representative by adding auxiliary noise input. And the robustness can be improved by the noise-added training. Experi-

mental results show that our method not only outperforms other state-of-the-art image-to-image translation methods, but also works well on a variety of datasets for photo-to-caricature translation of faces in the wild.

Limitations. Translating photo to caricature is a very challenging high-level image-to-image translation task. Thus, our model also fails in some cases, *e.g.*, some generated images of Yale and CelebA dataset in Figure 11. Although our method can keep the structure information of faces, it is still hard to render the details for providing high-quality caricatures and some tiny organs (such as eyes) are lack of details in Figure 4. Besides, it's also sensitive to side face with complex background, *e.g.*, some cases on CelebA and KDEF datasets in Figure 11.

Future work. With regard to future work, first, it would be interesting to investigate our method on other tasks of high-level image-to-image translation (*e.g.*, human-to-cartoon translation for cartoon movies); second, for the proposed method, the model still needs to be improved to provide high-quality rendered translated results; third, we intend to apply our method on the real applications of automatic and intelligent photo-to-caricature translation; fourth, we hope that we can control caricature style while translating images between domains by tuning input noise.

Acknowledgement

We thanks the volunteers for grading human scores of translation results from different methods. This work was supported in part by the National Natural Science Foundation of China (61771440, 41776113), in part by the China Scholarship Council (201806335022), and in part by the Qingdao Municipal Science and Technology Program (17-1-1-5-jch).

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: CVPR, 2017, pp. 1125–1134.

[2] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, 2017, pp. 2223–2232.

[3] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: Unsupervised dual learning for image-to-image translation, in: ICCV, 2017, pp. 2849–2857.

[4] Q. Chen, V. Koltun, Photographic image synthesis with cascaded refinement networks, in: ICCV, 2017, pp. 1511–1520.

[5] T. Kim, M. Cha, H. Kim, J. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, arXiv preprint arXiv:1703.05192.

[6] X. Mao, S. Wang, L. Zheng, Q. Huang, Semantic invariant cross-domain image generation with generative adversarial networks, Neurocomputing 293 (2018) 55–63.

[7] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, B. Zheng, Discriminative region proposal adversarial networks for high-quality image-to-image translation, in: ECCV, 2018, pp. 770–785.

[8] Y. Zhou, X. Bai, W. Liu, L. J. Latecki, Similarity fusion for visual tracking, International Journal of Computer Vision 118 (3) (2016) 337–363.

[9] Q. Wang, S. Li, H. Qin, A. Hao, Super-resolution of multi-observed RGB-D images based on nonlocal regression and total variation, IEEE Transactions on Image Processing 25 (3) (2016) 1425–1440.

[10] J. Ma, S. Li, H. Qin, A. Hao, Unsupervised multi-class co-segmentation via joint-cut over l_1-manifold hyper-graph of discriminative image regions, IEEE Transactions on Image Processing 26 (3) (2017) 1216–1230.

[11] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,

IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (11) (2017) 2298–2304.

[12] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, A. A. Efros, Generative visual manipulation on the natural image manifold, in: ECCV, 2016, pp. 597–613.

[13] C. Li, M. Wand, Precomputed real-time texture synthesis with Markovian generative adversarial networks, in: ECCV, 2016, pp. 702–716.

[14] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, in: ECCV, 2016, pp. 318–335.

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: CVPR, 2017, pp. 4681–4690.

[16] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation, in: ICLR, 2017.

[17] M. Sela, E. Richardson, R. Kimmel, Unrestricted facial geometry reconstruction using image-to-image translation, in: ICCV, 2017, pp. 1576–1585.

[18] H.-Y. Fish Tung, A. W. Harley, W. Seto, K. Fragkiadaki, Adversarial inverse graphics networks: Learning 2D-to-3D lifting and image-to-image translation from unpaired supervision, in: ICCV, 2017, pp. 4364–4372.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.

[20] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.

[21] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: ICLR, 2016.

[22] Y. Liu, Z. Qin, T. Wan, Z. Luo, Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks, Neurocomputing 311 (2018) 78–87.

[23] E. Akleman, J. Palmer, R. Logan, Making extreme caricatures with a new interactive 2D deformation technique with simplicial complexes, in: Proceedings of Visual, 2000, pp. 100–105.

[24] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434.

[25] D. Berthelot, T. Schumm, L. Metz, BEGAN: Boundary Equilibrium Generative Adversarial Networks, arXiv preprint arXiv:1703.10717.

[26] A. Mishra, S. N. Rai, A. Mishra, C. Jawahar, IIIT-CFW: A benchmark database of cartoon faces in the wild, in: ECCVW, 2016, pp. 35–47.

[27] W. Zhang, X. Wang, X. Tang, Coupled information-theoretic encoding for face photo-sketch recognition, in: CVPR, IEEE, 2011, pp. 513–520.

[28] X. Wang, X. Tang, Face photo-sketch synthesis and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (11) (2009) 1955–1967.

[29] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets, in: NIPS, 2016, pp. 2172–2180.

[30] B. Abaci, T. Akgul, Matching caricatures to photographs, Signal, Image and Video Processing 9 (1) (2015) 295–303.

[31] C. E. Thomaz, G. A. Giraldi, A new ranking method for principal components analysis and its application to face image analysis, Image and Vision Computing 28 (6) (2010) 902–913.

[32] A. Georghiades, P. Belhumeur, D. Kriegman, Yale face database, Center for computational Vision and Control at Yale University, http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html 2.

[33] D. Lundqvist, A. Flykt, A. Öhman, The karolinska directed emotional faces (KDEF), CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet (1998) 1094–1118.

[34] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: ICCV, 2015, pp. 3730–3738.

[35] M. Mathieu, C. Couprie, Y. LeCun, Deep multi-scale video prediction beyond mean square error, arXiv preprint arXiv:1511.05440.

[36] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: ICML, 2016, pp. 1060–1069.

[37] X. Yan, J. Yang, K. Sohn, H. Lee, Attribute2image: Conditional image generation from visual attributes, in: ECCV, 2016, pp. 776–791.

[38] H. Zhang, V. Sindagi, V. M. Patel, Image de-raining using a conditional generative adversarial network, arXiv preprint arXiv:1701.05957.

[39] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: CVPR, 2016, pp. 2536–2544.

[40] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Transactions on Graphics 36 (4) (2017) 1–14.

[41] W. C. Luo, P. C. Liu, M. Ouhyoung, Exaggeration of facial features in caricaturing, in: Proceedings of International Computer Symposium, 2002.

[42] Y.-L. Chen, W.-H. Tsai, et al., Automatic generation of talking cartoon faces from image sequences, in: Proceedings of Conferences on Computer Vision, Graphics & Image Processing, 2004.

[43] P.-Y. C. W.-H. Liao, T.-Y. Li, Automatic caricature generation by analyzing facial features, in: ACCV, 2004.

[44] L. Zhang, Y. Ji, X. Lin, Style transfer for anime sketches with enhanced residual U-net and auxiliary classifier GAN, arXiv preprint arXiv:1706.03319.

[45] L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: CVPR, IEEE, 2016, pp. 2414–2423.

[46] J. Bruna, P. Sprechmann, Y. LeCun, Super-resolution with deep convolutional sufficient statistics, arXiv preprint arXiv:1511.05666.

[47] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: NIPS, 2016, pp. 658–666.

[48] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: ECCV, Springer, 2016, pp. 694–711.

[49] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, arXiv preprint arXiv:1609.04802.

[50] J. Zhao, M. Mathieu, Y. LeCun, Energy-based generative adversarial network, arXiv preprint arXiv:1609.03126.

[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[52] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: NIPS, 2016, pp. 2234–2242.