2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Unsupervised Contrastive Photo-to-Caricature Translation based on Auto-distortion

Yuhe Ding*†, Xin Ma†‡, Mandi Luo†‡, Aihua Zheng* and Ran He†‡
*School of Computer Science and Technology, Anhui University
†NLPR & CEBSIT, CASIA
‡School of Artificial Intelligence, University of Chinese Academy of Sciences
Email: madao3c@foxmail.com, xin.ma@cripac.ia.ac.cn, mandi.luo@cripac.ia.ac.cn, ahzheng@foxmail.com, rhe@nlpr.ia.ac.cn

***Abstract*—Photo-to-caricature translation aims to synthesize the caricature as a rendered image exaggerating the features through sketching, pencil strokes, or other artistic drawings. Style rendering and geometry deformation are the most important aspects in photo-to-caricature translation task. To take both into consideration, we propose an unsupervised contrastive photo-to-caricature translation architecture. Considering the intuitive artifacts in the existing methods, we propose a contrastive style loss for style rendering to enforce the similarity between the style of rendered photo and the caricature, and simultaneously enhance its discrepancy to the photos. To obtain an exaggerating deformation in an unpaired/unsupervised fashion, we propose a Distortion Prediction Module (DPM) to predict a set of displacements vectors for each input image while fixing some controlling points, followed by the thin plate spline interpolation for warping. The model is trained on unpaired photo and caricature while can offer bidirectional synthesizing via inputting either a photo or a caricature. Extensive experiments demonstrate that the proposed model is effective to generate hand-drawn like caricatures compared with existing competitors.**

I. Introduction

As a special image-to-image translation task, caricature generation requires exaggerating on face features, and re-rendering the facial texture to form a portrait. Existing methods mainly fall into three classes, deformation-based, texture-based, and methods taking both aspects into consideration.

Deformation-based methods focus on the geometric distortion by using a certain guidance, such as 2D landmarks or 3D mesh [1]. However, it is challenging to guarantee the precise guidances. Furthermore, without consideration of texture render, they tend to generate less hand-drawn painting like results.

Texture-based methods devote to obtain caricature's style via the prevalent GANs [2]. Zheng *et al*. [3] use the cycle generators to preserve the texture consistency in caricature generation. Benefit from facial masks, Li *et al*. [4] transfer the texture of the input image through weakly paired adversarial learning. However, they only consider the geometric deformation in representation space thus result in limited deformation.

More recently, to exaggerate deformation and obtain plausible texture simultaneously, Cao *et al*. [5] propose to first employ two CycleGANs [6] to transfer photos and landmarks, respectively, then warp the rendered image with the help of landmarks. Shi *et al*. [7] propose a warp controller to predict a set of controlling points and its displacements, followed by an AdaIN-based [8] rendering network to transfer the texture. Despite the great progress for caricature translation by taking both deformation and texture into consideration in these two methods, there are two intuitive defects remaining improvement.

On the one hand, their style rendering effect appears either not plausible enough or with prominent artifacts. Cao *et al*. [5] retain too much photo information since the CycleGAN-based architecture emphasizes single style translation learning, which is hard to learn the diverse texture styles existing in caricatures. Shi *et al*. [7] tend to generate more artifacts since the AdaIN-based structure assumes that feature maps in different channels are uncorrelated, which ignores the global information in style rendering. Therefore, we propose a contrastive photo-caricature translation method in this paper. First, we design our transferring architecture based on weight sharing strategy [9] to maintain the global information without the uncorrelated assumption between different channels. Second, Hadsell *et al*. [10] evidences that contrastive loss can pull closer similar pairs and push away dissimilar pairs, However, conventional Euclidean distance based contrastive loss is simply defined by subtraction between two images, which is not suitable to measure the similarity of image styles. Therefore, we propose a style distance by gram matrix, which can enhance the texture details by calculating the dot product of feature maps for any two channels, and introduce the proposed contrastive style loss to enforce the texture similarity of the rendered photo to caricatures and its discrepancy to the photos.

On the other hand, although Cao *et al*. [5] support unsupervised translation due to their cycle architecture, the precise landmarks required as guidance during the translation are hard to be guaranteed. Shi *et al*. [7] propose a warp controller to predict a set of points for warping, which avoids extra guidance information such as landmarks. However, as the weakly supervision information, the identity labels are hard to obtain in the wild, which is the crucial information in Shi *et al*. [7] to generate characteristic exaggeration. To obtain exaggerating deformation in the unsupervised fashion without the guidance condition, we propose the Distortion Prediction

[1]Yuhe Ding and Xin Ma contribute equally to this work.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412569

Module (DPM), to automatically predict a set of displacements for the predefined controlling points in this paper. Then we use the controlling points and their displacements for warping via a classical interpolation method: thin plate spline [11]. Note that since we only predict the displacements rather than simultaneously predict the controlling points, we can decrease the unexpected deformation as well as simplify our network complexity. Particularly, in the test stage, random-perturbed photos are input to DPM to achieve diverse deformation in caricatures. Note that our goal is the unsupervised/unpaired translation which means identity label is not available/required. Furthermore, it is available for bidirectional synthesizing with either a photo or a caricature as the input. More comparisons to the state-of-the-art caricature generation methods are shown in Table I.

Based on the above discussion, we propose an unsupervised Contrastive Translation for auto-distortion Photo-to-Caricature translation method in this paper. The main contributions include:

- To reduce artifacts in rendered photos, we propose a novel contrastive loss by define a style to enforce the similarity between the rendered photo's texture and caricatures, and enforce its discrepancy to the photos, which can generate plausible textures in the rendered photos with more details.
- To obtain exaggerating deformation in the unpaired setting, we propose a new symmetrical architecture with Distortion Prediction Modules (DPM), which predicts a set of displacements vectors without any guidance to warp the images in an unsupervised fashion.
- Experiments on the benchmark caricature generation dataset WebCaricature [12] compared to the state-of-the-art methods demonstrate our methods can synthesize caricatures with more hand-drawn like texture with diverse deformation without guidance in the unsupervised fashion. In addition, our method supports bidirectional translation due to the symmetrical architecture.

II. Related works

We briefly review the related works on the following three aspects.

A. Caricature Generation

To generate exaggerating caricatures, traditional works define a shape representation such as 2D landmarks and 3D mesh [1], [13], then calculate a mean face to exaggerate the representative feature where have the largest deviation from mean face. However, their capability of geometry distortion are generally limited due to the need for guidance. And their results suffer from poor visual quality because these networks are not suitable for problems with large spatial variation. With the blossom of deep learning, generative adversarial networks (GAN) [2] have a widespread application in computer vision [14], [15], especially in caricature generation. Cao *et al*. [5] recently propose to decouple texture rendering and geometric deformation with two CycleGANs trained on image and landmark space, respectively. But with their face shape modeled in the PCA subspace of landmarks, they suffer from the same problem of the traditional deformation-based methods Shi *et al*. [7] propose a weakly-supervised generation model, and obtain commendable results in the aspect of geometry distortion. A warp controller and a rendering network are used to process geometry and style, respectively. The style rendering network is based on AdaIN [8], which assumes the feature maps in different channels are uncorrelated, hence ignoring the global and identical information. Therefore, it tends to produce many artifacts.

B. Style Rendering

Many style translation tasks have been proposed to render images from the source texture to the target one. These methods fall into two main categories: supervised and unsupervised models. The major difference of the two categories is whether the training data are paired or not. Pix2pix [16] is one of the prevalent supervised frameworks, which learns a mapping function from the source to the target domain. Wang *et al*. further propose Pix2pixHD [17] for high-resolution photo-realism translation. Representative supervised models [16]–[19] are wildly researched and applied in the past decade. However, paired image data restrict its application in real-world applications. As pioneer unsupervised translation, CycleGAN [6], DualGAN [20] and DiscoGAN [21] translate images using cycle consistency. Subsequently, various GAN-based unsupervised translation models [22]–[26] emerge for one-to-many translation, many-to-many translation and so on. However, they mainly focus on style rendering and less consider the spatial distortion that is important for photo-to-caricature translation.

Method	Component				
	Texture	Exaggeration	Diversity	Bidirectional	Unsupervised
CycleGAN [6]	✓	×	×	✓	✓
UNIT [9]	✓	×	×	✓	✓
MUNIT [27]	✓	×	✓	✓	✓
StarGAN [22]	✓	×	×	✓	✓
CariGAN [4]	✓	✓	×	×	×
CariGANs [5]	✓	✓	×	✓	✓
WarpGAN [7]	✓	✓	✓	×	×
Ours	✓	✓	✓	✓	✓

TABLE I: Comparison of different caricature generation methods.

C. Spatial Distortion

To process geometry distortion, parameter-based methods [28], [29] predict a set of deformation parameters, but they can not process with fine-grained distortion because the number of parameters is limited. Some methods [30], [31] use a dense motion field to warp the images, all vertices in a deformation grid are predicted, yet most of them are useless. Besides, the disentanglement-based method [32] performs well at some simple datasets. However, because of the hundreds of styles, caricatures are too complex to disentangle. To

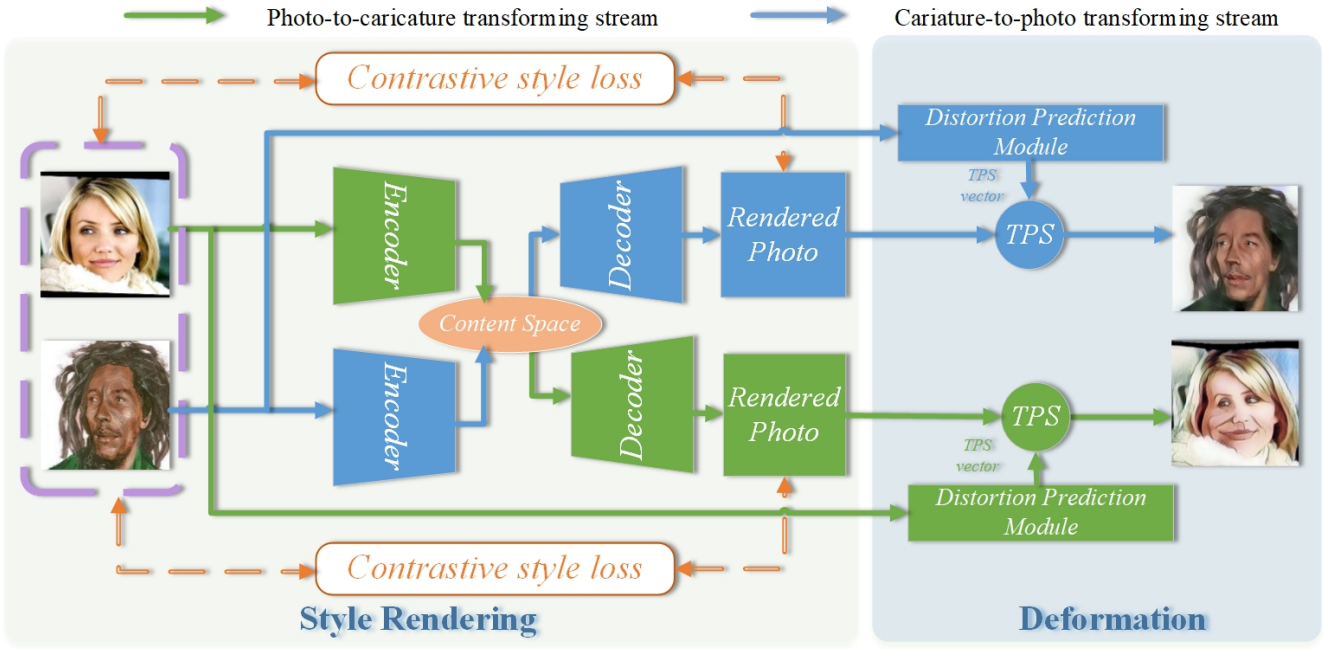

Fig. 1: Overview of our symmetric architecture. The green and blue lines represent photo-to-caricature and caricature-to-photo transforming streams respectively. *Content Space* represents the common content space. Our translation fall into two stages: style rendering and deformation. We propose contrastive style loss and distortion prediction module in these two stages, respectively.

caricature's deformation, there are some specific approaches. Cao *et al*. [5] use PCA landmarks in a way of CycleGAN [6], yet the manual annotations are needed. Shi *et al*. [7] propose a warp controller to predict a set of control points and its displacements automatically. This method works well under a weakly supervised setting so that, it is not suitable for the unsupervised setting.

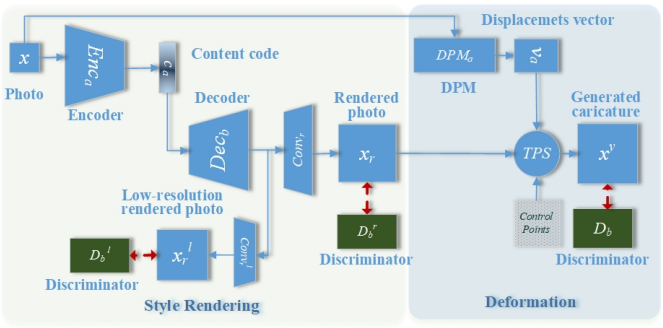

Fig. 2: The pipeline of photo-to-caricature transforming. The caricature-to-photo transforming can be designed in the same manner.

III. Method

In this paper, we propose a symmetrical encoder-decoder architecture with contrastive style loss and Distortion Prediction Module (DPM) for photo to caricature transforming, to decrease the rendered photo's artifacts, and exaggerating geometric deformation under unpaired setting. Fig. 1 illustrates the whole architecture of the proposed model. It is a symmetrical structure that supports the bidirectional transforming. Each direction of transforming stream consists of an encoder, a decoder, a distortion prediction module (DPM) and three discriminators.

A. Transforming Stream

Let $x \in X$ be the images from human face photos, and $y \in Y$ be the images from hand-drawn caricatures. Given an input photo x, our goal is to generate a corresponding caricature $x^y \in Y$. As shown in Fig. 2, our transforming stream falls into two stages: style rendering and geometry deformation.

First, for an input face photo x, the encoder Enc_a first maps x to a code in content space, and then decodes the content code c_a to transform the photo into rendered photo x^r via decoder Dec_b. Note that an auxiliary rendered photo x_r^l with a quarter size of rendered photo x_r is output by different layers of decoder Dec_b too. To enforce the similarity between rendered photo and input caricature and its discrepancy to input photo, we propose a contrastive style loss in this stage.

Second, the distortion prediction module DPM_a estimates the displacements vector v_a for the input photo x,. Given the predefined controlling points p_0, the deformed photo x^y is obtained via thin plate spline interpolation (TPS) [11],

$$x^y = TPS_{p_0,v_a}(Dec_b(c_a)), \tag{1}$$

where $v_a = DPM_a(x)$, $c_a \sim q_a(c_a|x)$, and Gaussian distribution $q_a(c_a|x) \equiv N(c_a|Enc_a(x), I)$.

Note that we obtain the exaggerating deformation in the unpaired/unsupervised fashion without any additional guidance.

B. Style Rendering

Different with AdaIN [8] based methods [7], [27], which assume that different channel is uncorrelated, we use shared-latent space assumption and weight sharing strategy, to better capture the global information in photos. Then, we propose to enforce a contrastive style loss to to decrease the artifacts.

Shared-latent space assumption. As noted in Liu *et al.* [9], for any given images x and y from different domains, there exists a shared latent code z in a shared-latent space, as shown in Fig 3. Specifically, given encoders Enc_a, Enc_b, decoders Dec_a, Dec_b, we have $z = Enc_a(x) = Enc_b(y)$, and $x = Dec_a(z), y = Dec_b(z)$. With this assumption, we have $y = Dec_b(Enc_a(x))$, $x = Dec_a(Enc_b(y))$, which map X to Y and Y to X, respectively.

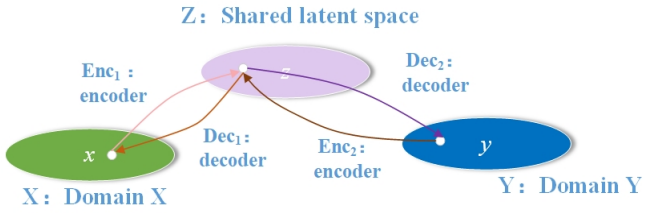

Fig. 3: The shared-latent space assumption.

Our backbone is based on variational auto-encoders (VAEs) [33]–[35] and GANs [2] [9]. For the photo domain X, encoder-decoder pair $\{Enc_a, Dec_a\}$ constitute a VAE, while decoder-discriminator pairs $\{Dec_a, D_a^l\}$, $\{Dec_a, D_a^r\}$, and $\{Dec_a, D_a\}$ constitute three GANs. With the shared-latent space assumption, we assume that the content space C is conditionally independent and Gaussian with unit variance. The encoder outputs a mean vector $Enc_a(x)$ and the distribution of the content code c_a is given by $q_a(c_a|x) \equiv N(c_a|Enc_a(x), I)$, where I is an identity matrix.

The reconstructed photo $\widetilde{x}$ is calculated by $\widetilde{x} = Dec_a(c_a \sim q_a(c_a|x))$, and rendered photo x_r is calculated by $x_r = Dec_b(c_a \sim q_a(c_a|x))$. Here the distribution of $q_a(c_a|x)$ is treated as a random vector sampled from $N(c_a|Enc_a(x), I)$.

The reparameterization trick [33] is utilized to reparameterize the non-differentiable sampling operation as a differentiable operation using auxiliary random variables. This reparameterization trick allows us to train VAEs using back-prop. Let s be a random vector with a multi-variate Gaussian distribution: $s \sim N(s|0, I)$. The sampling operations of content code $c_a \sim q_a(c_a|x)$ can be implemented via $c_a = Enc_a(x)+s$. Similarly, reconstructed caricature $\widetilde{y} = Dec_b(c_b \sim q_b(c_b|y))$, caricature content code $c_b = Enc_b(y) + s$.

We now introduce the reconstructed loss within domain,

$$\begin{aligned}\mathcal{L}_{rec} =& \| \widetilde{x} - x \|_1 + \| \widetilde{y} - y \|_1 \\ =& \| Dec_a(c_a) - x \|_1 + \| Dec_b(c_b) - y \|_1.\end{aligned} \tag{2}$$

Based on the shared-latent space assumption, we take a weight-sharing approach [9]. In detail, the last two layers' weights of encoder Enc_a and encoder Enc_b, which extract high-level features, are sharing. Analogously for the first two layers of decoder Dec_a and decoder Dec_b which are vital to decode the high-level representation. The KL loss is introduced to enforce the outputs of these two encoders share a common content space,

$$\mathcal{L}_{KL} = KL(q_a(c_a|x) \parallel p_s(c)) + KL(q_b(c_b|y) \parallel p_s(c)), \tag{3}$$

where KL divergence term is defined as:

$$KL(p \parallel q) = \sum_{i=1}^{n} p(x_i) \cdot log\frac{p(x_i)}{q(x_i)}, \tag{4}$$

the KL divergence term penalizes the deviation of the distribution of the content space from the prior distribution. We use Gaussian distribution to model q_a and q_b, where the prior distribution is a standard Gaussian distribution $p_s(c) \sim N(0, I)$.

In the photo-to-caricature translation stream, decoder Dec_b decodes the content code of domain X. Note that our decoders have two branches, which generate images with 128-scale and 256-scale, noted as Dec_a^l and Dec_a^r respectively. To generate plausible rendered photo, the adversarial loss is introduced,

$$\begin{aligned}\mathcal{L}_{adv}^{styG} =& \mathbb{E}[log(1 - D_b^l(x_r^l))] + \mathbb{E}[log(1 - D_b^r(x_r))] + \\ & \mathbb{E}[log(1 - D_a^l(y_r^l))] + \mathbb{E}[log(1 - D_a^r(y_r))],\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{adv}^{styD} =& -\mathbb{E}[logD_b^l(x^l)] - \mathbb{E}[logD_b^r(x)] \\ & -\mathbb{E}[logD_a^l(y^l)] - \mathbb{E}[logD_a^r(y)] \\ & -\mathbb{E}[log(1 - D_b^l(x_r^l))] - \mathbb{E}[log(1 - D_b^r(x_r))] \\ & -\mathbb{E}[log(1 - D_a^l(y_r^l))] - \mathbb{E}[log(1 - D_a^r(y_r))],\end{aligned} \tag{6}$$

where low resolution auxiliary rendered photo $x_r^l = Dec_a^l(Enc_a(c_a \sim q_a(c_a|x) + s \sim N(0, I)))$, rendered photo $x_r = Dec_a^r(Enc_a(c_a \sim q_a(c_a|x) + s \sim N(0, I)))$, $\mathbb{E}$ means expectation. y_r^l, y_r, y^l and y in the caricature domain can be obtained in the same manner.

To preserve the identity information, inspired by Gatys *et al.* [36], we define the content loss to enforce the content similarity between the photo x and rendered photo x_r,

$$\mathcal{L}_{cont} = \| \xi(x) - \xi(x_r) \| + \| \xi(y) - \xi(y_r) \|, \tag{7}$$

where $\xi(\cdot)$ is a pretrained VGG net [37].

Contrastive Style Loss To enforce the similarity of the rendered photo to caricatures and its discrepancy to the photos, we propose a constrastive style loss function to pull the texture of rendered photo x_r to the input caricature and push it away from the input photo. First, given two feature maps m and n, the style distance [36] can be defined as,

$$d(m, n) = \frac{1}{4 * n_c * n_h * n_w} \sum_{1}^{n_c} (G_{ij}^m - G_{ij}^n)^2, \tag{8}$$

where G^m is the gram matrix, which represents the dot product of the of feature m in any two channels. G^n is the gram matrix of feature n extracted by a pretrained VGG net. Gram matrix measures the characteristics of each feature dimension

and their relationships. Therefore, it can enlarge the texture details. Then we define our contrastive style function as,

$$\begin{aligned} Ctr(i_1, i_2, l) =& \frac{1}{2}[l \cdot d(i_1, i_2)^2 + \\ & (1-l)max(mg - d(i_2, i_1), 0)^2], \end{aligned} \tag{9}$$

where $l \in \{0, 1\}$ is the label of the images pair $\{i_1, i_2\}$, the style distance between the image pair decreases when $l = 1$, and increases when $l = 0$. $mg > 0$ is a margin, indicating that only the images from two domains with a style distance between 0 and mg are considered. The whole contrastive style loss is,

$$\begin{aligned} \mathcal{L}_{ctr} =& \alpha_1 Ctr(x_r, x, 0) + \alpha_2 Ctr(x_r, y, 1) + \\ & \alpha_3 Ctr(y_r, y, 0) + \alpha_4 Ctr(y_r, x, 1). \end{aligned} \tag{10}$$

where α_i is the hyper parameters.

The contrastive style loss in Eq. (10) pulls the style distance between rendered photo x_r and input caricature y, while pushing its distance away from the input photo x apart, which can lead to more plausible caricature style texture of rendered photo. The rendered caricature y_r is optimized in the same manner.

C. Distortion Prediction Module

Despite of the plausible texture rendering, the other key issue in photo-to-caricature translation is the exaggerating deformation. To obtain exaggerating deformation in a completely unsupervised fashion without any guidance, we propose a Distortion Prediction Module (DPM) is used in this stage. DPM accepts diverse distortion via a random-perturbed input photo.

DPM is a subnetwork with four fully connected layers. DPM_a accepts photo as input, to output the corresponding displacements vector $v_a = \{v_a^1, v_a^2, \ldots, v_a^n\}$. we predefine the controlling points as $p_0 = \{p^1, p^2, \ldots, p^n\}$, where both displacement vector v_a^i and controlling point p^i are 2-D vectors in $u - v$ space. We employ the thin plate spline interpolation (TPS) [11] for our deformation, due to its remarkable performance in warping [7],

$$x^y = TPS_{p_0, v_a}(x_r). \tag{11}$$

Now the rendered photo with exaggerating deformation is obtained. To make sure DPM can predict meaningful results, some constraints are needed to enforce decoder Dec_b^r to generate plausible results. Discriminator D_b is added in this stage to introduce the adversarial loss,

$$\mathcal{L}_{adv}^{warpG} = \mathbb{E}[log(1 - D_b(x^y))] + \mathbb{E}[log(1 - D_a(y^x))], \tag{12}$$

$$\begin{aligned} \mathcal{L}_{adv}^{warpD} =& -\mathbb{E}[D_b(x^y)] - \mathbb{E}[D_a(y^x)] \\ & -\mathbb{E}[log(1 - D_b(x^y))] - \mathbb{E}[log(1 - D_a(y^x))]. \end{aligned} \tag{13}$$

Preventing from undesirable deformation, identity loss is introduced:

$$\mathcal{L}_{idt} = \mathbb{E} \parallel x^y - x \parallel + \mathbb{E} \parallel y^x - y \parallel . \tag{14}$$

We set a small weight to this strong constraint to avoid generated effect from being too similar to the original image.

D. Total Loss

The proposed method is optimized in a two-stage strategy. First (section III-B), we optimize the style network via,

$$\begin{aligned} \min_{Enc_a, Enc_b, Dec_a, Dec_b} =& \lambda_r \mathcal{L}_{rec} + \lambda_K \mathcal{L}_{KL} + \lambda_a \mathcal{L}_{adv}^{styG} + \\ & \lambda_c \mathcal{L}_{cont} + \lambda_{ctr} \mathcal{L}_{ctr}, \\ \min_{D_a^l, D_a^r, D_b^l, D_b^r} =& \lambda_a \mathcal{L}_{adv}^{styD}. \end{aligned} \tag{15}$$

Second, we optimize the distortion prediction modules via,

$$\begin{aligned} \min_{DPM_a, DPM_b} =& \lambda_a \mathcal{L}_{adv}^{warpG} + \lambda_i \mathcal{L}_{idt}, \\ \min_{D_a, D_b} =& \lambda_a \mathcal{L}_{adv}^{warpD}. \end{aligned} \tag{16}$$

Note that we are under unsupervised setting, and support bidirectional translation.

IV. Experiments

We evaluate our method on the benchmark photo-caricature dataset Webcaricature [12] [38] consisting of 6042 caricatures and 5974 photos from 252 identities.

A. Training Details

We employ Adam algorithm [39] to iteratively update our model according to the defined the loss functions Eq. (15) and (16), until achieving a minimal expectation. Hyper parameters β_1 and β_2 representing exponential decay rates is fixed as $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

There are a random pair of photo and caricature in each mini-batch. We train our model for 100,000 and then 50,000 steps for the first and second stages respectively. Learning rate is 0.0001. We first optimize our encoders, decoders, and discriminators except D_a and D_b via Eq. (15), then optimize DPMs, D_a and D_b via Eq. (16) with the same learning rate. We empirically set the hyper parameters $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \lambda_r, \lambda_K, \lambda_a, \lambda_c, \lambda_{ctr}, \lambda_i, mg\}$ as $\{0.5, 0.5, 1, 1, 10, 1, 1, 1, 0.5, 8, 2.0\}$. The implementation platform is pytorch 1.4.0, python 3.6 on one Geforce GTX 1080 Ti GPU.

B. Comparisons with State-of-the-Arts

We first qualitatively compare our caricature generation method with four state-of-the-art methods as shown in Fig. 4. We can see that, CycleGAN [6] and StarGAN [22] tend to produce very photo-like images with trivial texture changes. This may be due to that these two methods are based on cycle consistency and multi-classification so that they retain too much photo's information. UNIT [9] and MUNIT [27] demonstrate the most visually appealing texture styles due to the weigh-sharing strategy and AdaIN [8], respectively. However, UNIT [9] generates too much artifacts without any consideration on geometric deformation. AdaIN [8] assumes feature maps in different channels are uncorrelated, which ignores the global information and results in artifacts too. All these four methods focus only on transferring the texture styles, while fail to deform the faces into caricatures. Although they try to compensate the distortion by using texture, it easily

Fig. 4: Comparison with five state-of-art methods. (a) input photos, (b)-(f) Generation results of five state-of-art methods. (g) Our results.

results in difficult training and mode collapse. WarpGAN [7] aims to generate plausible caricature, and take both texture and deformation into consideration. However, it also generates many unexpected artifacts as MUNIT [27]. Since these two are both based on AdaIN [8]. Our results in Fig. 4 (g) present more plausible style and deformation with less artifacts, benefit from the texture rendering enforced by contrastive style loss on the weight-sharing strategy. In addition, the distortion prediction module allows each input photo an exaggerating deformation.

C. Distortion Diversity

We introduce a scale factor α during deployment to allow customization of the exaggeration extent. We scale the displacement of control points i.e. warp vector v by α to control how much the face shape will be exaggerated. The bigger α, the larger degree of deformation. Fig. 5 shows the generated caricatures with different scale factors. It is clear that the deformation increases as the scale factor ascends. In addition, we sample different noises from standard Gaussian distribution with clamping in $[-0.1, 0.1]$, and simply add it to input. DPM accepts a random-perturbed version of inputs, which cannot change the overall information but lead to more diverse deformation. Fig. 6 demonstrates the diverse exaggeration results with additional random Gaussian noises.

D. Ablation Study

To verify the contribution of each component in our model, we evaluate three variants of the proposed model, by removing $\mathcal{L}_{ctr}$ and $\mathcal{L}_{cont}$ from the first stage, and removing $\mathcal{L}_{idt}$ from the second stage respectively for ablation study. Fig. 7 and Fig. 8 demonstrate several qualitative results of each variant. Note that we only use the contrastive style loss $\mathcal{L}_{ctr}$ and the content loss $\mathcal{L}_{cont}$ in the first stage, therefore, as shown in

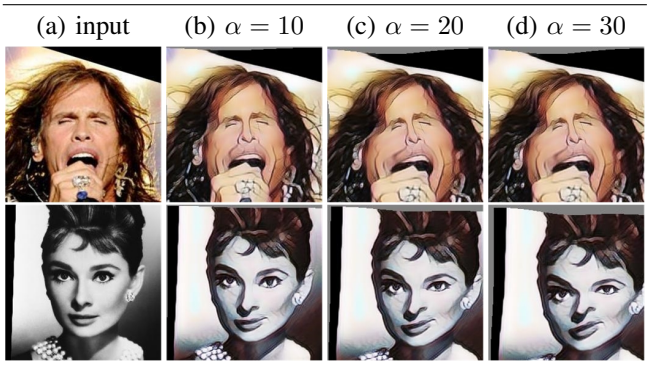

Fig. 5: Examples of diverse distortions against the scale factor α.

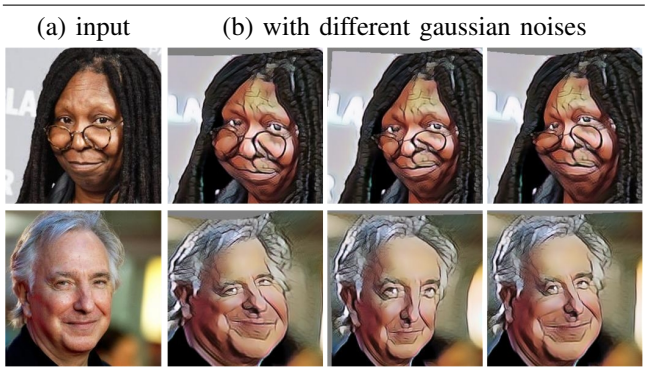

Fig. 6: Examples of diverse distortions against different additional Gaussian noises.

Fig. 7: Ablation study on contrastive style loss $\mathcal{L}_{ctr}$ and content loss $\mathcal{L}_{cont}$. Note that all the results are the rendered images in the first stage.

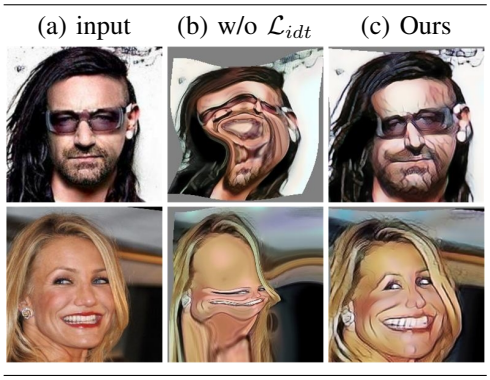

Fig. 8: Ablation study on identity loss $\mathcal{L}_{idt}$.

Fig. 7, we evaluate their contributions through the rendered images without deformation for better comparison. It is clear that, Fig. 7 (b) indicates that model with content loss $\mathcal{L}_{cont}$ preserves more identity details between the generated images and the original ones, since their deep features are pulled together by $\mathcal{L}_{cont}$. Fig. 7 implies that the contrastive style loss $\mathcal{L}_{ctr}$ plays a important role in image stylization, while it will result in poor style generation without $\mathcal{L}_{ctr}$ Fig. 8 illustrates the results of the variant without identity loss $\mathcal{L}_{idt}$, from which we can see, it can lead to less unexpected deformation.

E. User Study

We investigate the user study against two competitors, WarpGAN [7] and UNIT [9]. 42 caricatures synthesized by each of the three methods are presented to 29 subjects, who are told to vote the best caricature considering both geometry and style aspects. The results are shown in Table II reports

Method	Proportion
UNIT [9]	24.39%
WarpGAN [7]	24.56%
Ours	51.05%

TABLE II: Voting results of our user study.

Probe	Rank-1 accuracy
Photos	100%
Hand-drawnings	8.46%
WarpGAN [7]	34.18%
Ours	34.56%

TABLE III: Rank-1 face recognition accuracy of four different matching protocols using a state-of-art face recognition model SphereFace [40].

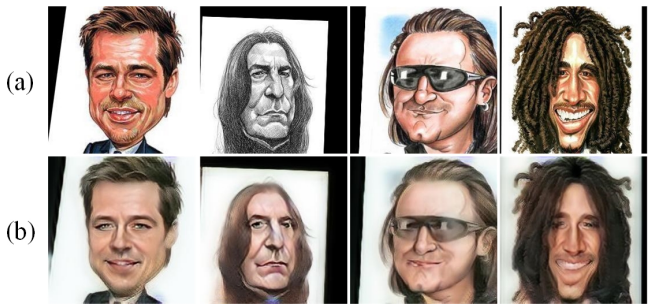

Fig. 9: Caricature-to-photos synthesizing without deformation. (a) Input caricatures. (b) Transformed photos.

the voting ratios of each method. Our method gets the highest voting against the other two competitors, which verifies the performance of our method on photo-caricature generation.

F. Face Recognition

To evaluate the ability of identity maintaining of our method, we evaluate the face recognition task on the whole dataset via the widely used face recognition model SphereFace [40]. Specifically, we select one photo for each 252 identity in training and testing sets as gallery, while randomly selecting 4000 photos, hand-drawn caricatures, caricatures synthesized by WarpGAN and our methods respectively as probe. Table III reports the Rank-1 accuracy of the recognition results in four probe scenarios. It is clear that, caricatures are harder to preserve the identity information comparing to the photos due to the large distortions and huge style changes. Our method still achieves comparable accuracy as WarpGAN and significantly beats results probed by the hand-drawings, which verifies the ability of identify preservation of our method.

G. Caricature-to-Photo Translation

Benefit from our symmetric structure, our method supports the bidirectional translation between photos and caricatures. Fig. 9 demonstrates additional results of caricature-to-photo synthesizing. However, our distortion modules are not well suitable for the caricature-to-photo translation since caricature's shape is extremely irregular. Results show that the proposed method can learn the skin texture with fewer artifacts.

V. Conclusion

In this paper, we propose a symmetric architecture for photo-to-caricature translation. We divide this process into

two stages, style rendering and geometry distortion. In the first stage, we propose a novel contrastive style loss to better render the texture style and decrease the artifacts. In the second stage, we propose a distortion prediction module for diverse unsupervised deformation. Comprehensive experiments verify the effectiveness of our method comparing with the existing methods.

VI. Acknowledgements

This work is partially funded by Beijing Natural Science Foundation (Grant No. JQ18017), Youth Innovation Promotion Association CAS (Grant No. Y201929), the National Natural Science Foundation of China (61976002) and the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033).

References

[1] X. Han, K. Hou, D. Du, Y. Qiu, Y. Yu, K. Zhou, and S. Cui, "Caricatureshop: Personalized and photorealistic caricature sketching," *arXiv preprint arXiv:1807.09064*, 2018.

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

[3] Z. Zheng, C. Wang, Z. Yu, N. Wang, H. Zheng, and B. Zheng, "Unpaired photo-to-caricature translation on faces in the wild," *Neurocomputing*, vol. 355, pp. 71–81, 2019.

[4] W. Li, W. Xiong, H. Liao, J. Huo, Y. Gao, and J. Luo, "Carigan: caricature generation through weakly paired adversarial learning," *arXiv preprint arXiv:1811.00445*, 2018.

[5] K. Cao, J. Liao, and L. Yuan, "Carigans: Unpaired photo-to-caricature translation," *arXiv preprint arXiv:1811.00222*, 2018.

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

[7] Y. Shi, D. Deb, and A. K. Jain, "Warpgan: Automatic caricature generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 762–10 771.

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

[9] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[10] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[11] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[12] J. Huo, W. Li, Y. Shi, Y. Gao, and H. Yin, "Webcaricature: a benchmark for caricature recognition," *arXiv preprint arXiv:1703.03230*, 2017.

[13] T. Lewiner, T. Vieira, D. Martínez, A. Peixoto, V. Mello, and L. Velho, "Interactive 3d caricature from harmonic exaggeration," *Computers & Graphics*, vol. 35, no. 3, pp. 586–595, 2011.

[14] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 637–647, 2017.

[15] R. He, B.-G. Hu, and X.-T. Yuan, "Robust discriminant analysis based on nonparametric maximum entropy," in *Asian Conference on Machine Learning*. Springer, 2009, pp. 120–134.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "pix2pixhd: High-resolution image synthesis and semantic manipulation with conditional gans."

[18] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual gan," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 205–220.

[19] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in neural information processing systems*, 2017, pp. 465–476.

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

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

[22] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[23] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," in *Advances in neural information processing systems*, 2017, pp. 752–762.

[24] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 783–790.

[25] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 3–21.

[26] Y. Wang, J. van de Weijer, and L. Herranz, "Mix and match networks: encoder-decoder alignment for zero-pair image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5467–5476.

[27] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.

[28] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[29] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "St-gan: Spatial transformer generative adversarial networks for image compositing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9455–9464.

[30] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky, "Deepwarp: Photorealistic image resynthesis for gaze manipulation," in *European conference on computer vision*. Springer, 2016, pp. 311–326.

[31] R. Wu, X. Tao, X. Gu, X. Shen, and J. Jia, "Attribute-driven spontaneous motion in unpaired image translation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5923–5932.

[32] N. S. Detlefsen and S. Hauberg, "Explicit disentanglement of appearance and perspective in generative models," *arXiv preprint arXiv:1906.11881*, 2019.

[33] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[34] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and variational inference in deep latent gaussian models," in *International Conference on Machine Learning*, vol. 2, 2014.

[35] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[36] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] J. Huo, Y. Gao, Y. Shi, and H. Yin, "Variation robust cross-modal metric learning for caricature recognition," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 340–348.

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

[40] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.