2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

DOMAIN-ADAPTIVE GENERATIVE ADVERSARIAL NETWORKS FOR SKETCH-TO-PHOTO INVERSION

Yen-Cheng Liu[1], *Wei-Chen Chiu*[2], *Sheng-De Wang*[1], *and Yu-Chiang Frank Wang*[1]

[1]Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[2]National Chiao-Tung University, Hsinchiu, Taiwan

ABSTRACT

Generating photo-realistic images from multiple style sketches is one of challenging tasks in image synthesis with important applications such as facial composite for suspects. While machine learning techniques have been applied for solving this problem, the requirement of collecting sketch and face photo image pairs would limit the use of the learned model for rendering sketches of different styles. In this paper, we propose a novel deep learning model of *Domain-adaptive Generative Adversarial Networks (DA-GAN)*. The design of DA-GAN performs cross-style sketch-to-photo inversion, which mitigates the difference across input sketch styles without the need to collect a large number of sketch and face image pairs for training purposes. In experiments, we show that our method is able to produce satisfactory results as well as performing favorably against state-of-the-art approaches.

Index Terms— Image Inversion, Deep Learning, Convolutional Neural Network, Generative Adversarial Network

1. INTRODUCTION

In the past few years, deep neural networks have been successfully applied to the applications in image synthesis such as image super-resolution, inpainting, and colorization [1–3]. Recently, artistic style transfer [4–6] attracts the attention from both researchers and users, aiming at synthesizing images which preserve context information of the input while converting its style based on a different image of interest.

It is worth noting that, existing works on style transfer consider different settings, which might limit their uses and thus affect the resulting performances. For example, the style transfer work in [4] takes both the input image and the style image of interest to complete the inversion process using pre-trained networks, which are applied for calculating the similarity between the extracted visual features (in terms of context and style). To accelerate the above inversion process, the work of [5] adopts separate training and testing stages; the former requires the presence of a large number of cross-style image pairs, while the latter is applied to a given input image directly. Therefore, during the prediction stage, no style image of interest needs to be presented.

(a) (b)

Fig. 1. Example inversion results of [7]: (a) sketch inputs with the same style as that of the training sketches which have the associated ground truth photos, and (b) sketch inputs with a different style (e.g., hand-drawn sketches). Note that [7] is not able to deal with different style sketches. The inversion outputs in (b) tend to be colorized images instead of photo-realistic ones.

A recent image inversion work of [7] also adopts the above training and testing procedures, and focuses on the task of sketch to photo conversion. That is, not just to colorize the input sketches, a photo-realistic output image would be desirable. However, as noted above, such procedures require the presence of cross-domain image pairs during the training of their deep neural networks. If a different style of sketches are presented, the synthesized image output might not be necessarily photo-realistic (see Fig. 1 for examples).

Based on the recent success of generative adversarial networks (GAN) [8] and deep-learning based domain adaptation [9], we present to a novel deep learning framework for sketch-to-photo inversion. To be more specific, we propose *domain-adaptive GAN (DA-GAN)* consisting of a variational autoencoder (VAE) with a domain adversarial network. Our DA-GAN can be learned in a *semi-supervised* fashion across domains, so that one can apply the learned model for dealing with input sketches in different styles. As detailed later in Section 2, our DA-GAN exploits image pairs across sketch and photo domains for learning its image synthesis ability. In order to extend the learned model for handling input sketches in different styles, we only need to utilize additional sketch images (with the style of interest) during the training stage,

978-1-5090-6341-3/17/$31.00

without requiring the presence of the corresponding photo images.

The contributions of this paper are highlighted below:

1. We propose a novel deep-learning based model, domain-adaptive generative adversarial networks (DA-GAN), for sketch-to-photo inversion.
2. Our DA-GAN can be viewed as a semi-supervised learning model across domains, which is trained on sketch-photo image pairs and sketches of different styles (without ground truth photos available).
3. We show that, compared to recent sketch-to-photo deep learning model, our method can sufficiently associate photo images and their sketches in different styles, and produce satisfactory inversion results.

2. PROPOSED METHOD

As shown in Fig. 2, we assume there exists two distinct domain: simulated sketches x_s with corresponding photo images y_s in source domain and hand-drawn sketches x_T in absence of corresponding photo images in target domain. Our goal is to generate the photo-realistic images from two distinct style sketches and address the difficulty of lacking sketch-photo image pairs in target domain.

2.1. Preliminary: Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)

We now briefly review the generative adversarial network (GAN) [8] for the sake of completeness of this paper. Aiming at synthesizing or recovering signals with similar data distribution to the training ones, GAN utilizes a generative function $G(\mathbf{z}) \in \mathbb{R}^p \rightarrow \mathbb{R}^{m\times n\times 3}$ and a discriminative function $D(\mathbf{d}) \in \mathbb{R}^{m\times n\times 3} \rightarrow \mathbb{R}$, which optimizes a two-player minimax game with the following objective function:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{d}\sim p_{data}(\mathbf{d})}[\log D(\mathbf{d})] \\ + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D(G(\mathbf{z})))],\end{aligned} \tag{1}$$

where $\mathbf{d} \in \mathbb{R}^{m\times n\times 3}$ is a color image sampled from the distribution of training data, and $\mathbf{z} \in \mathbb{R}^p$ is a randomly sampled vector from the prior distribution $p_{\mathbf{z}}$. With (1) optimized, the generator G will be applied to produce images in $\mathbb{R}^{m\times n\times 3}$ lying within the distribution of the training data, while D can discriminate between the observed data with the same/different distributions as that of the training ones.

With the goal of learning the model to generate the samples similar with training data, [10] introduce variational autoencoder, which aims at maximizing the probability of each X in training set under the generative process,

$$P(X) = \int P(X|z;\theta)dz, \tag{2}$$

where θ denotes the network parameters. The key idea of VAE is to learn an auxiliary distribution Q, such that $E_{z\sim Q(z|X)}P(X|z)$ attempts to sample values of z that is more likely to produce X. The relation between $P(X)$ and $E_{z\sim Q(z|X)}P(X|z)$ can be expressed as:

$$\begin{aligned}\log P(X) - \mathcal{KL}\left[Q(z|X)\|P(z|X)\right] \\ = E_{z\sim Q}\left[\log P(X|z)\right] - \mathcal{KL}\left[Q(z|X)\|P(z)\right]\end{aligned} \tag{3}$$

where $P(z)$ denotes a prior which is often assumed as Gaussian random variable. For practical implementation of Eqn. 3, $P(z|X)$ and $Q(X|z)$ can be viewed as the encoder and decoder respectively, which can be optimized via standard gradient-based method under the reparameterization trick [10]. The first term in right-hand side of Eqn. 3 is regular reconstruction objective of autoencoder; the last term is KL-divergence regularization, which urges approximate posterior similar with the prior.

2.2. Architecture

As illustrated in Fig. 2, our proposed domain-adaptive GAN (DA-GAN) consists of two network models: *inversion* and *domain adversarial networks*, as we detail below.

2.2.1. Inversion Network (Inv-Net)

The inversion network (*Inv-Net*) in DA-GAN adopts the architecture of variational autoencoder (VAE) [10] as a image synthesis model for converting sketch inputs into the associated photos, while the derived representation in the latent space is further utilized in DA-Net for discrimination purposes.

The reason why we adopt the idea of VAE as our *Inv-Net* is that, in addition to recovering the desirable output for each sketch input (as the standard autoencoder does), we aim to learn a joint distribution of latent features across domains such that the decoder is able to generate photo-realistic images based on arbitrary sketch inputs which are not included during training.

Based on the above motivation, we follow VAE [10] and adopt its design for the latent feature ρ in our DA-GAN.

2.2.2. Domain Adversarial Network (DA-Net)

In order to handle sketch inputs from target domain during training stage (e.g., the hand-drawn ones without corresponding photo images shown in Fig. 2), we share encoder across domains to transfer the feature encoding of sketches from source to target domain, and further narrow the gap of feature distributions between two domains based on the strength of GANs [8]. The encoder of *Inv-Net* can also be viewed as generator (G) of GAN conditioned on the sketches; the domain

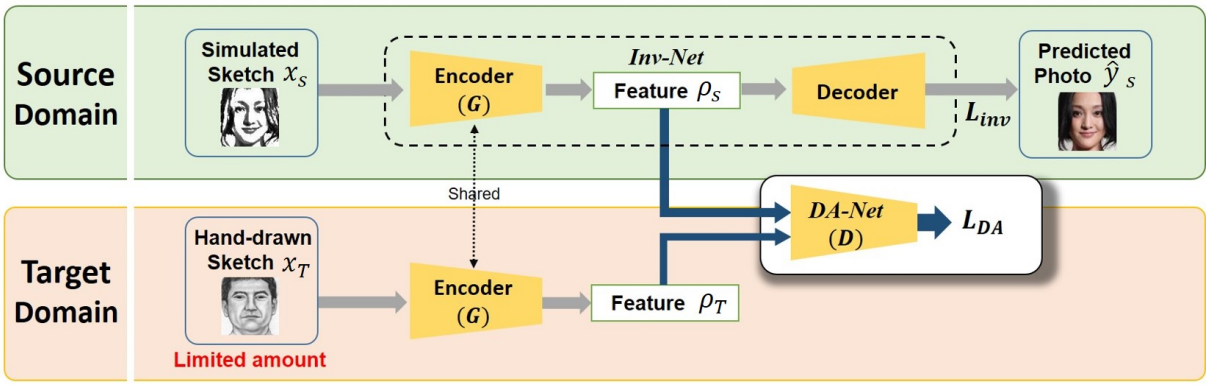

Fig. 2. Our proposed Domain-Adaptive Generative Adversarial Network (DA-GAN) for sketch-to-photo inversion.

Algorithm 1: Learning of DA-GAN

Data: Face images y_s and the corresponding simulated sketches x_s; a set of hand-drawn sketches x_T
Result: Configurations of *Inv-net* and D

```
for Iters. of DA-GAN do
    for Iters. of Inv-Net do
        Update whole Inv-net to minimize (7) on x_s and y_s
    for Iters. of G do
        Update encoder of Inv-Net to maximize (8) on x_s and x_T
    for Iters. of D do
        Update DA-Net to minimize (8) on x_s and x_T
```

adversarial network (*DA-Net*) can be viewed as the discriminator.

Different from producing realistic images demonstrated in regular GANs, our DA-GAN is designed to mitigate the gap of the feature distributions from two distinct domains (i.e. source and target domain) with adversarial learning of discriminating the feature domain in D and generating domain-invariant feature in G.

As depicted in Fig. 2, while we can provide sketch-photo image pairs for learning our DA-GAN, not all the sketch inputs (e.g., the hand-drawn ones) have the corresponding photo-realistic outputs. In absence of such ground-truth information, the introduction of *DA-Net* for hand-drawn sketches would transfer the knowledge across sketch image domains, with the goal to eliminate the difference between input domains for recovering satisfactory photo-realistic outputs.

2.3. Objective

With the goal of preserving identity-aware information of input sketches, we consider the *perceptual loss* [5] L_p as the loss function of the *Inv-Net*:

$$L_p = \|\Phi(y_s) - \Phi(\hat{y_s})\|_F^2, \tag{4}$$

which calculates the element-wise reconstruction error between the synthesized output $\hat{y}_s$ and its ground-truth photo y_s in pre-trained VGG19 layer (*conv3_4*) Φ. The use of perceptual loss engenders the checkboard artifacts due to pooling layer in pre-trained model, so we then introduce total-variation loss L_{tv} [11]:

$$L_{tv} = \sum_{i,j} \sqrt{(\hat{y}_{i+1,j} - \hat{y}_{i,j})^2 + (\hat{y}_{i,j+1} - \hat{y}_{i,j})^2}, \tag{5}$$

where $\hat{y}_{i,j}$ denotes pixel with position (i, j) in synthesized image $\hat{y}_s$.

Since our inversion network adopts VAE for performing image conversion, we follow the setting of [10] and also include the standard VAE loss L_{VAE} and additional total variation loss L_{tv} as the loss functions for the inversion network:

$$L_{VAE} = L_p + \mathcal{KL}(q_T(z|x_s)||p(z)), \tag{6}$$

where x_s is the input simulated sketches, $\mathcal{KL}(q_T(z|x_s)||p(z))$ denotes Kullback-Leibler divergence over the prior $p(z)$ and the auxiliary distribution $q_T(z|x_s)$.

With the above loss functions, the overall objective function (i.e., the total loss function L_{inv}) of our *Inv-Net* can be calculated as follows:

$$L_{inv} = \lambda_{VAE} L_{VAE} + \lambda_{tv} L_{tv}. \tag{7}$$

As for the objective function of the *DA-Net*, we have the loss function L_{DA} defined as follows:

$$L_{DA} = \mathbb{E}[\log(1 - D(\rho_T))] + \mathbb{E}[\log(D(\rho_S))], \tag{8}$$

where ρ_S and ρ_T are the latent features produced by the simulated and hand-drawn sketches (i.e., source and target-domain sketches), respectively. In (8), D calculates the probability of

Table 1. Structure of our proposed DA-GAN.

*Inversion Network (**Inv-Net**)*			
Encoder		Decoder	
Layer	Activation Size	Layer	Activation Size
Input	$1 \times 96 \times 96$	Input	$256 \times 6 \times 6$
9×9 Conv.	$16 \times 96 \times 96$	Res. 512	$256 \times 6 \times 6$
5×5 Conv.	$32 \times 48 \times 48$	Res. 512	$256 \times 6 \times 6$
3×3 Conv.	$64 \times 24 \times 24$	3×3 Conv.	$128 \times 12 \times 12$
3×3 Conv.	$128 \times 12 \times 12$	3×3 Conv.	$64 \times 24 \times 24$
Res. 512	$512 \times 6 \times 6$	3×3 Conv.	$32 \times 48 \times 48$
Res. 512	$512 \times 6 \times 6$	3×3 Conv.	$16 \times 96 \times 96$
Res. 512	$512 \times 6 \times 6$	3×3 Conv.	$3 \times 96 \times 96$

*Domain Adversarial Network (**DA-Net**)*	
Layer	Activation Size
3×3 Conv.	$256 \times 6 \times 6$
3×3 Conv.	$16 \times 6 \times 6$
3×3 Conv.	$16 \times 6 \times 6$
3×3 Conv.	$32 \times 6 \times 6$
3×3 Conv.	$64 \times 6 \times 6$
3×3 Conv.	$64 \times 6 \times 6$
Fully Connected	1000
Fully Connected	1

observing the recovered latent feature generated from a real domain (i.e., hand-drawn sketches), which means the task of D is to distinguish whether the query feature (ρ_S or ρ_T) generated from source or target domain.

The learning of DA-GAN is now summarized in Algorithm 1, which includes the learning of generator G, *Inv-Net*, and the domain adversarial network *DA-Net*. We note that, learning standard GANs simply alternates between the optimization of G and D, which tends to be dominated by the learning of either network if the parameters are not carefully selected. As for DA-GAN, the learning of G is embedded in that of *Inv-Net*, which provides additional regularization during the training stage. In other words, the learning of DA-GAN is observed more stable and less sensitive to the optimization parameters (the ratio of updating times of G and D). Later in Sect. 3, we will provide the designs of both *Inv-Net* and *Domain Adversarial Network* for our DA-GAN.

3. EXPERIMENT

3.1. Implementation Details

We list the architecture and configuration of our DA-GAN in Table 1. All activation functions in our framework are Leaky ReLU with slope 0.2. To learn the DA-GAN, we use mini-batch SGD with batch size 8 and apply ADAM [12] for updating both *Inv-net* and *Domain Adversarial Network*.

3.2. Datasets and Settings

In our experiments of photo-to-sketch inversion, we consider the face image dataset of CelebA [13]. In order to learn our DA-GAN, we follow the procedure of simulating sketches in [7, 14], which synthesize the sketch images of the above face photos by Sobel edge detection and contrast enhancement, and the resulting (simulated) sketch and photo pairs are viewed as the training data. As a result, a total of 160,000 and 42,599 image pairs are the training and test sets, respectively.

To evaluate the robustness of our approach, we further consider the CUFSF dataset [15], which contains hand-drawn sketches and their ground truth face images. It is worth noting, since our DA-GAN is able to handle sketch images of different styles without ground truth face photos during training, we include 720 sketch images from CUFSF into the training set, while the remaining 240 sketches are viewed as the test data. In our work, we follow the setting in [7] and crop facial region of each image, which is resized into 96×96 pixels.

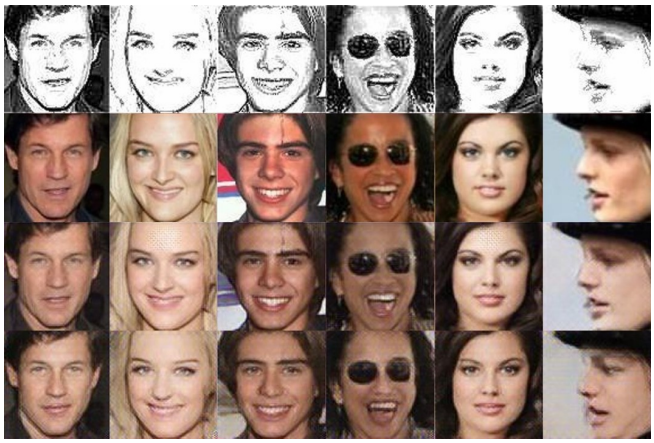

Fig. 3. Example results of photo inversion using **simulated sketches**. (Top to bottom rows: simulated sketches, ground-truth photos, inverted outputs of [7], and our inverted outputs)

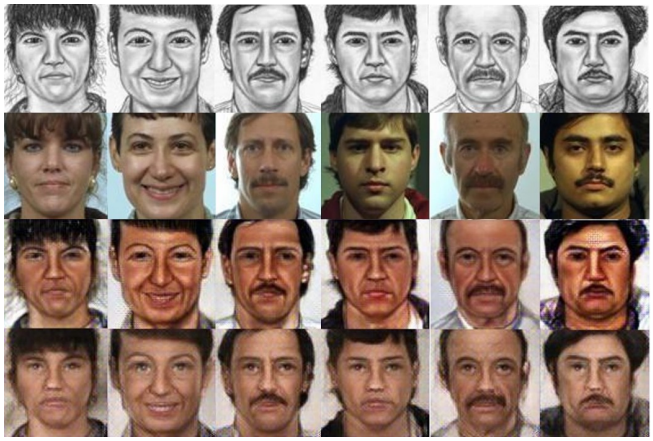

Fig. 4. Example results of photo inversion using **hand-drawn sketches**. (Top to bottom rows: hand-drawn sketches, ground-truth photos, inverted outputs of [7] and ours)

Table 2. Quantitative evaluation of image inversion using hand-drawn sketches as the inputs.

Approach	Perceptual Loss	PSNR	SSIM
Ours	**241.39**	**22.58**	**0.410**
[7]	273.34	22.01	0.378

3.3. Qualitative Evaluation

In Fig. 3, we visually assess the results of image conversion, using the simulated sketches as the inputs. From the example results in this figure, it can be seen that there is essentially no significant difference between the results produced by ours and the state-of-the-art approach of [7]. This is because that, the deep-learning based method of [7] trains their model on the image pairs of simulated sketches and photos, and thus it is also expected to perform promising inversion results as ours does. Both [7] and our DA-GAN were able to perform inversion for sketches with lighting, expression, and pose variations (and even for those with occlusion).

On the other hand, we present and compare the results of image conversion in Fig. 4, using hand-drawn sketches as the test inputs. Even without the ground-truth photo of hand-drawn sketches during training stage, our DA-GAN demonstrate the adaptability to multiple-style sketches and produces photo-realistic outputs and reduced sketch artifacts like wrinkles, black eyes, or uneven skin tone.

3.4. Quantitative Evaluation

We now perform quantitative evaluation on our proposed method. In particular, we consider the metrics of *perceptual loss* L_p in (4) and peak signal-to-noise ratio (i.e., smaller L_p or larger PSNR indicate better performances). Table 2 lists and compares the performances between [7] and our DA-GAN. We note that, PSNR might not be an ideal metric to be calculated, since the hand-drawn sketches and their ground-truth photos might not be perfectly aligned in CUFSF. Nevertheless, from Table 2, we see that our DA-GAN also outperforms the state-of-the-art approach of [7], which verifies the effectiveness and robustness of our approach for inverting different styles of sketches in practical scenarios.

4. CONCLUSION

In this paper, we presented Domain-adaptive Generative Adversarial Networks (DA-GAN), which is a deep learning model for rendering photo-realistic outputs from face sketches in different styles. Our DA-GAN is trained on image pairs of simulated sketches and also sketches of different styles (e.g., hand-drawn sketches), and thus can be viewed as a cross-domain semi-supervised deep learning model. By suppressing the difference between input sketches of different styles in a GAN based network, our DA-GAN further utilized the architecture of variational autoencoder (VAE) for performing image inversion. In the experiments, we provided qualitative and quantitative evaluation, which supported the effectiveness and robustness of DA-GAN.

5. REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.

[2] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color! large-scale texturing of 3d reconstructions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[6] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[7] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. van Gerven, "Convolutional sketch inversion," in *Proceedings of the European Conference on Computer Vision Workshop*. Springer, 2016.

[8] 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 (NIPS)*, 2014.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[10] M. W. Diederik P Kingma, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[11] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention.," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.