ATTRIBUTE-CONTROLLED FACE PHOTO SYNTHESIS FROM SIMPLE LINE DRAWING

Qi Guo   Ce Zhu   Zhiqiang Xia   Zhengtao Wang   Yipeng Liu

School of Electronic Engineering / Center for Robotics
University of Electronic Science and Technology of China (UESTC), Chengdu, China

ABSTRACT

Face photo synthesis from simple line drawing is a one-to-many task as simple line drawing merely contains the contour of human face. Previous exemplar-based methods are over-dependent on the datasets and are hard to generalize to complicated natural scenes. Recently, several works utilize deep neural networks to increase the generalization, but they are still limited in the controllability of the users. In this paper, we propose a deep generative model to synthesize face photo from simple line drawing controlled by face attributes such as hair color and complexion. In order to maximize the controllability of face attributes, an attribute-disentangled variational auto-encoder (AD-VAE) is firstly introduced to learn latent representations disentangled with respect to specified attributes. Then we conduct photo synthesis from simple line drawing based on AD-VAE. Experiments show that our model can well disentangle the variations of attributes from other variations of face photos and synthesize detailed photorealistic face images with desired attributes. Regarding background and illumination as the style and human face as the content, we can also synthesize face photos with the target style of a style photo.

Index Terms— Photo synthesis, simple line drawing, face attributes, deep generative model

1. INTRODUCTION

Face sketch-photo synthesis has been well developed in recent years for its widely application on law enforcement. An experienced artist draws a sketch about the suspect according to the description of the witness. Then the sketch is transformed to a photorealistic face image. At last, the synthesized face photo is compared to the images from face datasets to find the suspect.

The majority of prior works utilize sketches with rich details that are difficult to draw. Liang et al. [1] conduct face sketch-photo synthesis from simple line drawing. As a rough kind of sketch, simple line drawing, merely containing the contour of human face, is easy to be obtained and modified for ordinary people, which also makes photo synthesis from simple line drawing a quite tough task. In most cases, witnesses can remember some visual attributes (hair color, complexion, etc.) of suspects apart from the outline. We take advantage of these face attributes to add rich details to the outline of the face. A simple line drawing with different face attributes correspond to different faces.

The prior works of face sketch-photo synthesis can be divided into two categories: traditional methods and deep learning methods.

Most of traditional methods are based on an image patch dictionary of exemplars [1, 2, 3, 4, 5, 6, 7, 8]. With the number of training data grows and the complicacy of chosen representation of image patch increases, the testing time grows linearly. Moreover, these exemplar-based methods require both training and testing faces constrained under the same conditions, such as the same race, front pose and similar background. These restrictions hinder them from generalizing to more complicated natural scenes. Even though Peng et al. [8] utilize multiple representations of image to avoid suffering from these limitations at the cost of test time consumption, they can scarcely synthesize natural face image with complex backgrounds and diverse poses.

Recently, deep learning is explosively used in computer vision for its powerful generalization in highly complex tasks. A trained deep neural network is quite fast in testing stage due to its feed-forward framework, which allows users to obtain the results in real time. Convolutional Sketch Inversion (CSI) [9] and Scribbler [10] apply deep learning into sketch-photo synthesis. The architecture in [9] is a one-to-one convolutional neural network, namely, it generates only one photo from one sketch. Scribbler [10] makes use of generative adversarial networks (GANs) [11, 12, 13] conditioned on sketched boundaries and sparse color strokes to generate realistic faces, cars and bedrooms. It allows users to scribble over the sketch to indicate preferred color for objects. Complex as human faces, simple color strokes are not enough to generate rich details.
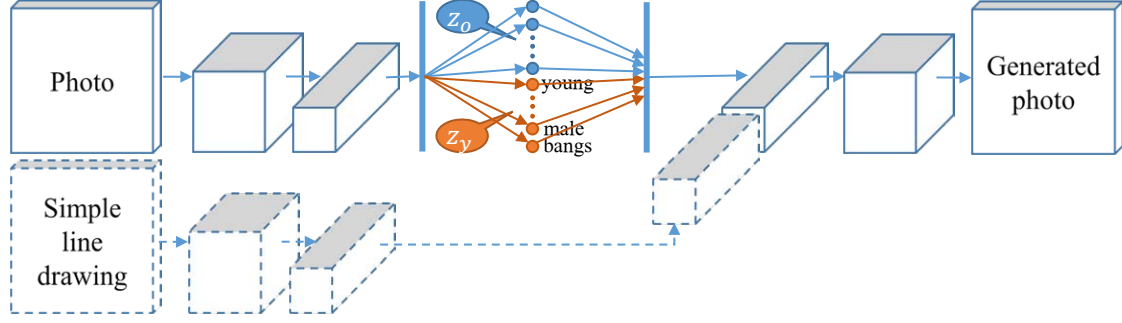
Fig. 1: Model architecture. Solid lines depict the basic structure of AD-VAE. Dashed lines show another channel of convolutional neural network which takes simple line drawing as input.

In this paper, we propose a deep generative model to synthesize face photo from simple line drawing controlled by face attributes. First, an attribute-disentangled variational auto-encoder (AD-VAE) is introduced to learn latent representations disentangled with respect to specified face attributes. We regard face attributes as some high-level variations of face images and specify a factored set of latent variables in variational auto-encoder (VAE) [14, 15] to capture these variations utilizing the binary attribute labels. The remaining latent variables are to learn other factors of variation such as illumination and background. Then we conduct face photo synthesis from simple line drawing by adding another channel of convolutional neural network to the AD-VAE which takes simple line drawing as input.

Different from attribute-conditioned variational auto-encoder (AC-VAE) [16], AD-VAE adds an inference from input images to the attribute variables, which can turn discrete variables to be continuous. This supervised manner is similar to [17], but we adopt a generative way. [18] force some latent variables to specifically represent active transformations in 3D datasets by organizing their training data changing in only one single scene variable for each mini-batch. However, such data organization is unmanageable for many natural image databases.

Regarding background and illumination as the style and human face as the content, simple line drawing and face attributes control the content of the generated face image, and the other latent variables control the style of the generated face image. Given a simple line drawing, the proposed method can generate different face photos with desired attributes and random styles. Given a simple line drawing and a style photo, we can also synthesize photo with target style.

2. METHODS

2.1. Variational auto-encoder

Given an input $x \in \mathbb{R}^N$ and its corresponding latent variables $z \in \mathbb{R}^M$, the basic structure of VAE consists of two networks: an encoder $q_\phi(z|x)$ (recognition model) to approx-

imate posterior inference and a decoder $p_\theta(x|z)$ (generative model) to map the latent variables to data space. Due to the intractable posterior $p_\theta(x|z)$, we maximize the following variational lower bound of log-likelihood $log p_\theta(x)$:

$$\mathcal{L}(x;\phi,\theta) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + E_{q_\phi(z|x)}[log p_\theta(x|z)] \tag{1}$$

where $p_\theta(z)$ is the prior distribution of the latent variables z, generally a simple isotropic unit Gaussian. D_{KL} is Kullback-Leibler divergence.

2.2. Attribute-disentangled variational auto-encoder

As shown in Figure 1, the latent variables z is split into two parts: $z_y \in \mathbb{R}^L$ and $z_o \in \mathbb{R}^K$. We regard some face attributes as a part of variations of face images and specify a factored set of variables z_y to capture these information. Each dimension of z_y represents one single attribute. The remaining factors of variation, such as position and background, are captured by z_o. z_y and z_o are independent and the dimensions of z_y are also independent with each other.

While the prior $p_\theta(z_o)$ remains to be an isotropic unit Gaussian, we choose a conditional distribution $p_\theta(z_y|y)$ as the prior of z_y instead:

$$p_\theta(z_o) \sim N(0, I)$$
$$p_\theta(z_y^i|y^i) \sim N(y^i, \sigma)$$

where y^i refers to i_{th} binary attribute label and σ is the standard deviation of $p_\theta(z_y|y)$. Then the variational lower bound described in Eq. 1 can be rewritten as:

$$
\begin{aligned}
L(x, y; \phi, \theta) = & -\sum_{i=1}^{L} \alpha^i D_{KL}(q_\phi(z_y^i|x)||p_\theta(z_y^i|y^i)) \\
& - \beta D_{KL}(q_\phi(z_o|x)||p_\theta(z_o)) \\
& + E_{q_\phi(z_o|x)q_\phi(z_y|x)}[log p_\theta(x|z_o, z_y)]
\end{aligned}
\tag{2}
$$

where $q_\phi(z_o|x)$, $q_\phi(z_y|x)$ and $p_\theta(x|z_o, z_y)$ are multivariate Gaussian distributions parameterized by deep neural networks.

|(a) Mouth_slightly_open | (b) Eyeglasses | (c) Bangs |
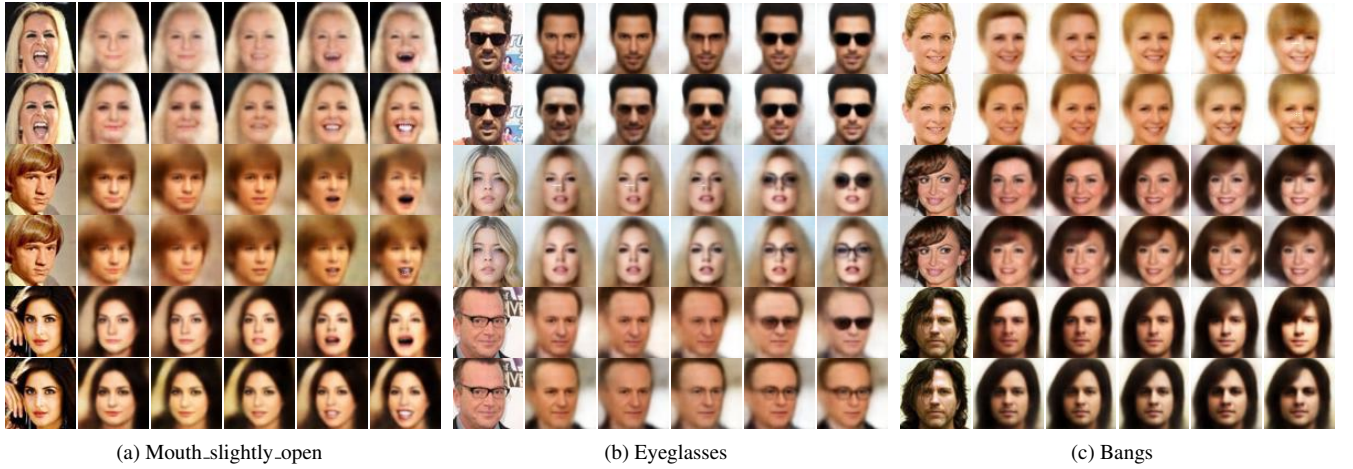
Fig. 2: Manipulating attribute variables. For each group, the first column is ground truth. For each ground truth, the first row is generated by AD-VAE and the second row is generated by AC-VAE.

The first term of above formula is discriminative because we specify the mean of prior $p_\theta(z_y|y)$ to be the binary attribute label y of the input image. Hence $q_\phi(z_y|x)$ also can serve as a classifier of face attributes by distinguishing the sign of the predicted mean of the posterior. In order to augment the quality of disentanglement with respect to the specified face attributes, we set a regularization coefficient vector $\alpha \in \mathbb{R}^L$ with large values to the discriminative KL term of the attributes. Larger α will lead to higher classification accuracy but worse reconstruction fidelity. Following β-VAE [19], we also set another coefficient β to constrain the capacity of other latent variables to hinder these variables from encoding variations of face attributes. We choose a smaller value of β than α.

2.3. Photo Synthesis from Simple Line Drawing

In order to synthesize face photo x from simple line drawing s controlled by face attributes y, we maximize the variational lower bound of the conditional log-likelihood $log p_\theta(x|s)$ instead:

$$
\begin{aligned}
L(x,y,s;\phi,\theta) = &-\sum_{i=1}^{L} \alpha^i D_{KL}(q_\phi(z_y^i|x)||p_\theta(z_y^i|y^i)) \\
&- \beta D_{KL}(q_\phi(z_o|x)||p_\theta(z_o)) \\
&+ E_{q_\phi(z_o|x)q_\phi(z_y|x)}[log p_\theta(x|z_o,z_y,s)]
\end{aligned}
\tag{3}
$$

z_o then mainly captures variations of background and illumination and we regard these variations as the style of face photos. As shown in Figure 1, another channel of convolutional neural network is added to the AD-VAE which takes simple line drawing as input and its feature maps are concatenated to the decoder.

3. EXPERIMENTS

3.1. Dataset

We conduct our experiments on CelebA dataset [20]. CelebA consists of 202599 face images annotated with 40 binary attributes such as male, young, smiling, etc. FDoG [21] filter is employed on CelebA to simulate the simple line drawing data. We binarize the synthetic simple line drawings with random thresholds on the training stage to avoid overfitting to the particular style. Both photos and simple line drawings are cropped and resized to 64×64. We use 182637 image pairs for training and the remaining 19962 pairs for testing. Among the training data, 10% are used for cross-validation. 38 attributes are selected without wearing necklace and wearing necktie.

We also test on ZJU-VIPA Line Drawing Face Database [1] which is build on CUHK Face Sketch Database (CUFS) [4].

3.2. Attribute Manipulation

In order to demonstrate the qualitative disentanglement with respect to the face attributes, we manipulate the attribute variables by varying desired attribute variable smoothly and keeping all other latent variables fixed. We compare AD-VAE to attribute-conditioned variational auto-encoder (AC-VAE) [16]. For fair model comparison, we train both AC-VAE and AD-VAE on CelebA with 38 selected binary attributes.

As shown in Figure 2, with the desired attribute variable changes, the corresponding images generated by AD-VAE transform more visibly and naturally than those generated by AC-VAE. For example, when we increase the attribute variable of mouth_slightly_open, both two models generate faces with their mouths open, but AC-VAE uses teeth to fill the

Fig. 3: Attribute controlled photo synthesis

mouth no matter whether the teeth of ground truth are visible or not. While AC-VAE can hardly remove the sunglasses of input face, AD-VAE generates realistic eyes to replace the sunglasses. These results show our proposed AD-VAE can effectively separate the variations of specified attributes and other variations of face images.

3.3. Photo Synthesis from Simple Line Drawing

In this experiment, we choose some attributes not contained in simple line drawing to guide the photo synthesis. As shown in Figure 3, controlled by face attributes, we can modify the hair color and complexion of the synthesized face photos. In Figure 4, two types of simple line drawings with different stroke weights are used to generate face photos. Compared to the CSI [9] with content loss, our proposed method can synthesize more photorealistic and natural faces images even though the contours of human faces are not complete.
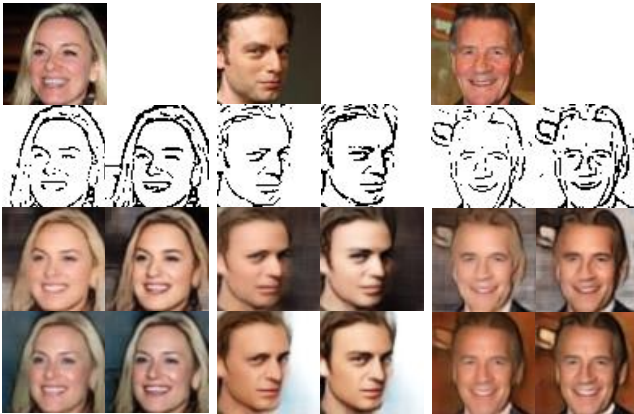

Fig. 4: Results comparison. From top to bottom: ground truth, two types of simple line drawings, CSI, our method

We exchange z_o of several photos in ZJU-VIPA Line Drawing Face Database and CelebA dataset to synthesize face images with target styles in Figure 5. As background and illumination are irrelevant to the human faces, this target style photo synthesis can eliminate these disturbances for further sketch-based face recognition.
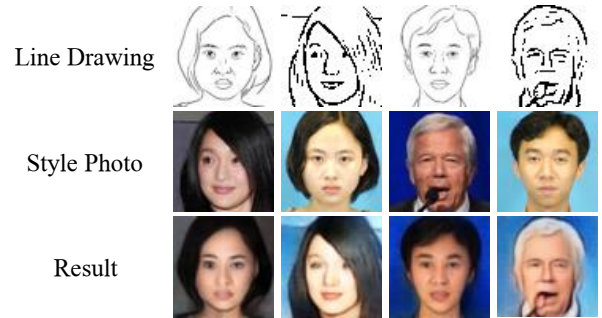

Fig. 5: Target style photo synthesis

4. CONCLUSION

This paper proposed a deep generative model to synthesize face photo from simple line drawing controlled face attributes. First, an attribute-disentangled variational auto-encoder (AD-VAE) is introduced to disentangle variations of face attributes from other variations of face images. Then we synthesized face photo from simple line drawing based on AD-VAE. Experiments showed our proposed method could learn interpretable representations of face images and generate face images with rich details and desired attributes even though the simple line drawing is not complete. We also did a target style photo synthesis that could help face recognition and face verification in further step.

5. REFERENCES

[1] Y. Liang, M.L. Song, L. Xie, J.J. Bu, and C. Chen, "Face sketch-to-photo synthesis from simple line drawing," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–5.

[2] W. Liu, X.O. Tang, and J.Z. Liu, "Bayesian tensor inference for sketch-based facial photo hallucination.," in *IJCAI*, 2007, pp. 2141–2146.

[3] B. Xiao, X.B. Gao, D.C. Tao, and X.L. Li, "A new approach for face recognition by sketches in photos," *Signal Processing*, vol. 89, no. 8, pp. 1576–1588, 2009.

[4] X.G. Wang and X.O. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[5] X.B. Gao, N.N. Wang, D.C. Tao, and X.L. Li, "Face sketch–photo synthesis and retrieval using sparse representation," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 8, pp. 1213–1226, 2012.

[6] H. Zhou, Z.H. Kuang, and K.Y.K Wong, "Markov weight fields for face sketch synthesis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1091–1097.

[7] N.N. Wang, D.C. Tao, X.B. Gao, X.L. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 9, pp. 1364–1376, 2013.

[8] C.L. Peng, X.B. Gao, N.N. Wang, D.C. Tao, X.L. Li, and J. Li, "Multiple representations-based face sketch–photo synthesis," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 11, pp. 2201–2215, 2016.

[9] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. AJ. van Gerven, "Convolutional sketch inversion," in *European Conference on Computer Vision*. Springer, 2016, pp. 810–824.

[10] P. Sangkloy, J.W. Lu, F. Chen, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," *Computer Vision and Pattern Recognition, CVPR*, 2017.

[11] I.J. 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.

[12] E.L. Denton, S. Chintala, R. Fergus, et al., "Deep generative image models using at laplacian pyramid of adversarial networks," in *Advances in neural information processing systems*, 2015, pp. 1486–1494.

[13] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[14] D.J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 1278–1286.

[15] D.P. Kingma and M. Welling, "Auto-encoding variational bayes," in *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[16] X.C. Yan, J.M. Yang, K. Sohn, and H.L Lee, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

[17] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," in *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[18] T.D. Kulkarni, W.F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Advances in Neural Information Processing Systems*, 2015, pp. 2539–2547.

[19] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-vae: Learning basic visual concepts with a constrained variational framework," in *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

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

[21] H. Kang, S. Lee, and C.H. Chui, "Coherent line drawing," in *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*. ACM, 2007, pp. 43–50.