EXAMPLE-GUIDED IDENTIFY PRESERVING FACE SYNTHESIS BY METRIC LEARNING

DAIYUE WEI, XIAOMAN HU, KEKE CHEN, PATRICK P. K. CHAN*

School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
E-MAIL: weidaiyueyue@outlook.com, JacquelineHxm@outlook.com, kkechen@qq.com, patrickchan@scut.edu.cn

Abstract:

Generative adversarial networks (GANs) are commonly applied to example-guided identify preserving face synthesis. A binary classifier is used as style consistency discriminator in GAN model in order to ensure the consistency of style. However, the over-fitting problem of a binary classifier downgrade its discrimination ability on style consistency. In this paper, we propose a style consistency discriminator based on metric learning, which performs better in keeping identity information and guaranteeing consistency in style between input exemplar and result. Through separating the positive pairs form the negative, metric learning model can efficiently measure the similarity between the synthesis face and the genuine face. The experimental results indicate that the metric learning performs better than a binary classifier in terms of preserving style consistency.

Keywords:

Generative adversarial networks; Identity preserving; Metric learning

1 Introduction

Image synthesis is a popular research direction in the field of computer vision nowadays and has a wide range of applications, such as super resolution [1], image inpainting [2], video prediction [3] and text-to-image synthesis [4]. Face synthesis is one of these applications that focuses on manipulating the attributes of human face [5]. Some related research directions such as pose transformation [6], expression synthesis and transfer [7], age progression and regression [8] have been proposed for face data enhancement to alleviate the imbalance of training data in deep learning model. Facial attributes are transformed during face synthesis. The identity is usually expected to preserve for the application in face recognition and verification. However, there still remains difficult for most of the generated models to achieve the purpose of identity preservation [9].

Recently Generative Adversarial Networks (GANs) [10] achieve satisfying results in image synthesis applications [11, 12]. However, there is no way to control the types of images that are generated. As a result, the conditional generative adversarial network (cGAN) [13] is proposed to allow image generation can be conditional on a class label, *i.e.*the targeted generated of images of a given type. For example, pix2pixSC [14] is a revised version of pix2pix [15] which is one of the popular example-guided image synthesis method. Different from pix2pix, an additional style consistency discriminator is included in pix2pixSC in order to determine whether the style of generated and real samples is consistent. Therefore, the quality of synthesized image is improved.

Using a binary classifier as the style consistency discriminator of cGANs [13] may cause a loss of identity information of the input exemplar image as a binary classifier may not learn the detail of features adequately. Although the training of a binary classifier is relatively fast, it is easily to suffer from the over-fitting problem. Moreover, the background of the same person is similar, a binary classifier may tend to focus on the background as an important feature in classification rather than the person. As a result, the output images synthesized with pix2pixSC model are not highly consistent in style with the exemplar. In this paper, we investigate whether the metric learning performs better than a binary classifier in serving as the style consistency discriminator in cGAN for identify preserving face synthesis. We use metric learning based model [16] as demonstration due to its excellent performance in dealing with diverse feature similarities in face recognition [17] and person re-identification [18]. Our revised model is evaluated experimentally and compared with pix2pixSC. The results indicate that our model performs better in keeping facial identity information, synthesizing natural and realistic face images in the task.

The rest of the paper is arranged as follows. Section 2 in-

troduces the previous work related to our model. Our proposed model is discussed in detail in Section 3. The experimental results and discussion are given in Section 4. Section 5 includes the conclusion.

2 Related Work

The concepts related to this study are briefly introduced in this section.

2.1 Generative Adversarial Networks (GANs)

GANs contain the generator G and the discriminator D. During the training, G aims to maximize the error of D while D aims to minimize its error. By the army race between G and D, GAN is able to generate a realistic sample, *i.e.*the synthesized objects is similar to the real ones. Recently, GANs achieve satisfying results in many applications of computer vision, *e.g.*face editing [19], image to image translation [20], and representation disentangling [21].

Although GANs have many advantages compared with other generation models, it is still impossible to generate images with an expected mode. Therefore, a conditional GANs (cGANs) [13] which employs prior information in image generation is proposed. cGANs becomes a general solution for a series of different image to image translation problems, in which included synthesizing photos from label maps, *e.g.*pix2pix [15], pix2pixHD [22], and pix2pixSC [14].

2.2 Metric Learning

The goal of metric learning is to minimize the intra-class variations and maximize the inter-class variations. Due to the success of deep learning, deep metric learning strategies have been successfully applied to many computer vision related application, *e.g.*landmark recognition [23] and pose estimation [24] recently. Triplet losses is one of the commonly used objective function in metric learning. Triplet losses consider the relative distance between the samples in the same and different classes. There are some examples of combining deep metric learning with GAN [25]. A triplet network is used as the discriminator in GANs. This method makes use of the good capability of representation learning of the discriminator to increase the predictive quality of the model. A novel framework to train GANs based on distance metric learning, which aimed at improving the performance and stability of GANs [26].

3 Proposed Model

Metric learning model separate the positive samples from the negative ones in a distance space, which directly corresponds to a similarity measurement. It maps images of each person into a cluster with a distance from others. It is the reason why the metric learning can be better to learn the discriminant characteristics for person identification. Our work applies the metric learning to pix2pixSC [14], which is example-guided image synthesis aims to synthesize an image from a semantic label map and an exemplary image indicating style. Pix2pixSC is firstly introduced and then the metric learning based discriminator in our model is introduced.

3.1 pix2pixSC

Given an input label map x and a guidance exemplar I, pix2pixSC synthesizes the result image $y : (x, I) \rightarrow y$. y is expected to be style consistent with the exemplar and semantically consistent with the label map. pix2pixSC consists of 1) a generator G, which takes label map x, exemplar I and its corresponding label map $F(I)$ as the input and syntheses y; 2) a discriminator D_R for differentiating the real and fake images synthesized by G; 3) a discriminator D_{SC} to determine whether y is style consistent with I.

The loss function of D_{SC} is defined as:

$$\begin{aligned} L_{SCAdv}(G, D_{SC}) =& E_{y_1^S, y_2^S}[logD_{SC}(y_1^S, y_2^S)]+ \\ & E_{y_1^N, y_2^N}[log(1 - D_{SC}(y_1^N, y_2^N))]+ \\ & E_{(x,I)}[log(1 - D_{SC}(I, G(x, I, f(I))))] \end{aligned} \tag{1}$$

where y_1^S and y_2^S are a pair of style consistent real images, and y_1^N and y_2^N are inconsistent in style. Figure 1 illustrates an example on the images with consistent and inconsistent style.

3.2 Contrastive and Triplet Loss in Metric Learning

Two commonly used loss, including contrastive and triplet loss [27–29], of metric learning are applied to revise pix2pixSC. Contrastive loss, shown as follows, maximizes the distance between two objects if they belong to different classes. Otherwise, their distance is minimized.

$$\begin{aligned} & L_{contrastive}(x_0, x_1) \\ & = 1/2y \left\| f(x_0) - f(x_1) \right\|_2^2 \\ & + 1/2(1 - y)max(0, m - \left\| f(x_0) - f(x_1) \right\|_2)^2 \end{aligned} \tag{2}$$

FIGURE 1. The style of the images is consistent in the first row but not consistent in second row.

.

where x_0 and x_1 are objects, f is the mapping function, y denotes the target label, and m is the margin which is the target distance between two objects of different classes.

On the other hand, triplet loss takes a triplet including an anchor sample, a positive and negative sample. The positive (negative) sample means the sample belongs to the same class (different classes) as the anchor. It aims to maximize the distance between the anchor and negative sample, and minimize the distance between the anchor and positive sample. The triplet loss is defined as:

$$\begin{aligned} &L_{triplet}(x_a, x_p, x_n) = \\ &max(0, m + \|f(x_a) - f(x_p)\|_2^2 - \|f(x_a) - f(x_n)\|_2^2) \end{aligned} \tag{3}$$

where x_a,x_p,x_n are the anchor, positive and negative samples.

3.3 Our proposed style consistency discriminator

The metric learning based style consistency discriminator is devised for pix2pixSC. Our proposed metric learning based style consistency discriminator aims to maximize the distance between objects with inconsistent style, minimize the distance between the objects with consistent style, and maximize the distance between the ones from exemplar and synthetic image. The loss function L_{SCmAdv} is defined as:

$$\begin{aligned} &L_{SCmAdv} = \\ &1/2(1-y)max(0, m - \|D_{SCm}(G(x, I, f(I)) - D_{SCm}(I)\|_2)^2 \\ &+ max(0, m + \left\|D_{SCm}(y_1^S) - D_{SCm}(y_2^S)\right\|_2^2 - \\ &\left\|D_{SCm}(y_1^N) - D_{SCm}(y_2^N)\right\|_2^2) \end{aligned} \tag{4}$$

where D_{SCm} denotes our proposed style consistency discriminator based on metric learning. By using the proposed style consistency discriminator, we expect the synthetic result to become consistent with the given exemplar in terms of style.

4 Experiments

The method with our proposed metric learning based style consistency discriminator is evaluated and compared with [14] experimentally in this section.

4.1 Experimental setting

27 high definition videos are downloaded from YouTube. Each video contains a single broadcaster who is Chinese and between 20 - 60 year old. Face alignment algorithm of [30] is applied to localize facial landmarks, which are used to create face sketch as input label map. Facial regions are then cropped and resized to size 256×256 pixels as in [14].

14 videos are randomly selected as the training set. Three frames are randomly selected from each video as exemplars in order to ensure the facial expression of the broadcaster in the frames is different. A label map is generated for each exemplar according to the method used in [14]. Two kinds of test sets are used for evaluation. The first test set contains the persons in the training set. Other four frames are extracted from each of 14 videos used in the training set. Another test set contains the person who never seem in training. Four frames are extracted from each of the rest of the videos (*i.e.*13 videos). Similarly, a label map is extracted from each sample in test sets. Examples of samples are showed in Figure 2. Our proposed style consistency discriminator with the metric learning is implemented by using convolutional-PReLU layers for encoding images into embedding vector with the exception that PReLU is not applied in the last convolutional layer. All networks are trained from scratch using the Adam solver with a batch size of 1 on a computer with NVIDIA GeForce GTX 1080 GPU.

Photorealism is evaluated by Frechet Inception Distance (FID) [31] which is widely used for implicit generative models. It calculates the FID between exemplars and synthesized images distributions. A smaller FID is often favored by the human subjects. The method used in [14] is applied to evaluate the Semantic Consistency by mapping the synthesized image back to face sketch and comparing consistency with input label map. A lower value is better. A subjective human perceptual study [32] is used to evaluate the style consistency.

4.2. Result and Discussion

Table 1 shows the quantification evaluation. FID of our model is lower than that of pix2pixSC, which indicates our

FIGURE 2. Examples of training samples. Each row represents one Triplet. From left to right images are anchor, positive and negative samples respectively.

Evaluation Metrics	ours	pix2pixSC
FID	122	137
Semantic Consistency	0.0083	0.0129
Human Perceptual Study	82.6%	17.4%

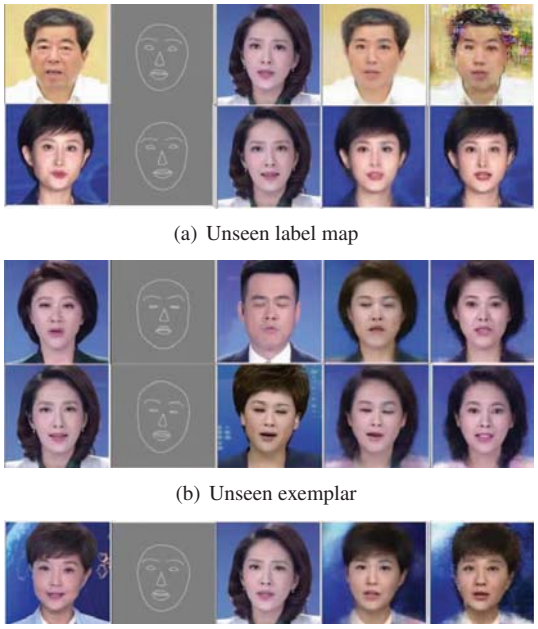
TABLE 1. Quantification evaluation on the results of our method and [14]

model performs better in generating photorealism face. Although the semantic consistency is not intentionally optimized, our method still gets better results (0.0083) than pix2pixSC (0.0129). Also, in human perceptual study, 82.6% of people agree that our results are better in identity preserving and clarity comparing to 17.4% obtained by pix2pixSC.

Figures 3 and 4 show the results in the first and second test sets. Each row represents one sample. From left to right, the images are input style sample, input label maps, the real image of label maps, the result of our model, and the result of [14].

Figures 3 shows pix2pixSC fails to generate a photorealistic image in the first test set. The photos generated by pix2pixSC are less photorealistic than those generated by our model. This may be because the metric learning model learns better identity discriminant features to constrain the style of the synthesized image.

Figures 4(a), 4(b) and 4(c) show the results when the label map, exemplar and both of them are not seem in training. Our model generalizes well in the case of the input label map is



(a) Unseen label map

(b) Unseen exemplar

(c) Unseen label map and exemplar

FIGURE 4. Example of results when the label map, exemplar and both of them are not seem in training. From left to right, the images are input exemplar, input label maps, the real image of label maps, the result of our model, and pix2pixSC

unseen in training. The performance of our method is better than the one of pix2pixSC, shown in Figure 4(a). In the case of the exemplar is unseen in training, pix2pixSC cannot get correct results in terms of semantic consistency with input label map, while our model still output reasonable images. The final scenario, both input label map and exemplar are not included in training. The performance of both our method and pix2pixSC drops significantly comparing to the previous ones. However, the overall result of our model still slightly better than pix2pixSC, *i.e.*our result has clearer facial outline.

5 Conclusion

This study aims to improve the style consistency discrimination ability of pix2pixSC, which is one of the commonly used example-guided identify preserving face synthesis method. Due to the limitation of a binary classifier, the metric learning is used as the style consistency discriminator for pix2pixSC. The experimental results suggest that the images generated by our method is more photorealistic and also more style-consistent than the ones generated by pix2pixSC, *i.e.*the metric learning improves the style consistency discrimination ability of

(a) (b)

FIGURE 3. Example of results with two different label maps on the training samples. From left to right, the images are input exemplar, input label maps, the real image of label maps, the result of our model, and pix2pixSC

pix2pixSC.

Acknowledgements

This work is supported by Natural Science Foundation of Guangdong Province, China (No.2018A030313203), the Fundamental Research Funds for the Central Universities South China University of Technology (No.2018ZD32).

References

[1] S. Y. Kim, J. Oh, and M. Kim, "Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video," *arXiv: Image and Video Processing*, 2019.

[2] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, "Image inpainting with learnable bidirectional attention maps," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[3] J. Wang, B. Hu, Y. Long, and Y. Guan, "Order matters: Shuffling sequence generation for video prediction," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[4] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L. D. Jorio, and T. Fevens, "Dual adversarial inference for text-to-image synthesis," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[5] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets." *arXiv: Computer Vision and Pattern Recognition*, 2018.

[6] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," *computer vision and pattern recognition*, pp. 8398–8406, 2018.

[7] Z. Shao, H. Zhu, J. Tang, X. Lu, and L. Ma, "Explicit facial expression transfer via fine-grained semantic representations," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[8] Q. Li, Y. Liu, and Z. Sun, "Age progression and regression with spatial attention modules," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[9] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation." *arXiv: Computer Vision and Pattern Recognition*, 2019.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," 2014.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *arXiv: Neural and Evolutionary Computing*, 2018.

[12] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv: Learning*, 2015.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv: Learning*, 2014.

[14] M. Wang, G. Yang, R. Li, R. Liang, S. Zhang, P. A. Hall, and S. Hu, "Example-guided style consistent image synthesis from semantic labeling." *arXiv: Computer Vision and Pattern Recognition*, 2019.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks." *arXiv: Computer Vision and Pattern Recognition*, 2016.

[16] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *computer vision and pattern recognition*, pp. 815–823, 2015.

[17] Y. Li, "Massface: an efficient implementation using triplet loss for face recognition." *arXiv: Computer Vision and Pattern Recognition*, 2019.

[18] Y. Zhang, Q. Zhong, L. Ma, D. Xie, and S. Pu, "Learning incremental triplet margin for person re-identification," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[19] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[20] P.-W. Wu, Y.-J. Lin, C.-H. Chang, E. Y. Chang, and S.-W. Liao, "Relgan: Multi-domain image-to-image translation via relative attributes." *arXiv: Computer Vision and Pattern Recognition*, 2019.

[21] T. Hinz and S. Wermter, "Image generation and translation with disentangled representations." *arXiv: Computer Vision and Pattern Recognition*, 2018.

[22] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *computer vision and pattern recognition*, pp. 8798–8807, 2018.

[23] A. Boiarov and E. Tyantov, "Large scale landmark recognition via deep metric learning," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[24] R. Mitra, N. B. Gundavarapu, S. C. Babu, P. Bindal, A. Sharma, and A. Jain, "3d human pose estimation under limited supervision using metric learning," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[25] M. Zieba and L. Wang, "Training triplet networks with gan," *arXiv: Machine Learning*, 2017.

[26] Z. Dou, "Metric learning-based generative adversarial network." *arXiv: Machine Learning*, 2017.

[27] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," *arXiv: Learning*, 2014.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[29] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," *computer vision and pattern recognition*, pp. 1386–1393, 2014.

[30] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[31] S. Mishra, D. Stoller, E. Benetos, B. L. Sturm, and S. Dixon, "Gan-based generation and automatic selection of explanations for neural networks." *arXiv: Machine Learning*, 2019.

[32] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis." *arXiv: Computer Vision and Pattern Recognition*, 2018.