From attribute-labels to faces: face generation using a conditional generative adversarial network

Yaohui Wang[1,2], Antitza Dantcheva[1,2], and Francois Bremond[1,2]

[1] Inria, Sophia Antipolis, France
[2] Université Côte d'Azur, France
{yaohui.wang, antitza.dantcheva, francois.bremond}@inria.fr

Abstract. Facial attributes are instrumental in semantically characterizing faces. Automated classification of such attributes (i.e., age, gender, ethnicity) has been a well studied topic. We here seek to explore the inverse problem, namely given attribute-labels the *generation of attribute-associated faces*. The interest in this topic is fueled by related applications in law enforcement and entertainment. In this work, we propose two models for attribute-label based facial image and video generation incorporating 2D and 3D deep conditional generative adversarial networks (DCGAN). The attribute-labels serve as a tool to determine the specific representations of generated images and videos. While these are early results, our findings indicate the methods' ability to generate realistic faces from attribute labels.

Keywords: attributes · generative adversarial network · face generation

1 Introduction

While attribute extraction and classification [4, 5, 3] is a well studied topic, the inverse problem, namely face generation, given attribute-labels is a novel area of high interest, due to related applications in law enforcement and entertainment. One specific application relates to the generation of realistic faces in cases of witness description, where the descriptions are the only available evidence (*e.g.*, in the absence of facial images). Particularly, law enforcement utilizes facial hand-drawn sketches or composites, which seek to support the process of suspect-identification. Such methods for face synthesis are slow, tedious, relatively unrealistic, as well as impeding efficient face recognition (*i.e.*, matching sketches or composites with existing mugshot databases maintained by law enforcement agencies poses a heterogeneous face recognition problem, which is highly challenging). Thus, reliable and automated label-based face generation constitutes beneficial in this context.

In spite of the aforementioned applications of interest, limited research concerns attribute-based face generation.

Motivated by the above, in this work we propose to generate faces based on attribute-labels. This incorporates two steps: (i) the learning of a text feature representation that captures the important visual details, as well as (ii) given the features, the generation of compelling realistic images. We propose two approaches based on deep conditional convolutional generative adversarial network (DCGAN) [11], which was introduced to *modify* images based on attributes (image-to-image translation). We train the proposed 2D GAN with the dataset CelebA and generate faces pertained to the attribute-set *glasses*, *gender*, *hair color*, *smile* and *age*. We selected these set of attributes for the associated high descriptiveness, *e.g.*, such attributes are commonly used by humans to describe their peers. For the experiment we generate facial images, which we evaluate by 2 common GAN-quality-scores, as well as by well established face detectors and an attribute classifier. More analysis of the 2D model is presented in Wang et al. [13]. In addition we propose a 3D GAN model, trained with the UvA-NEMO[3], generating facial smiling videos pertained to the attributes *gender* and *age*. Results indicate the method's ability to generate realistic faces from attribute labels.

2 Proposed methods

2D model The proposed conditional GAN aims to fit the conditional probability $P(x|z, y)$, as depicted in Figure 1. We let z be the noise vector sampled from $\mathcal{N}(0, 1)$ with dimension $N = 100$, y be the vector representing attribute-labels (with $y_i \in \pm 1$, where i corresponds to the i_{th} attribute). We train a GAN, adding attribute-labels in both, generator and discriminator. While the generator accepts as input the combination of prior noise $p(z)$ and attributes vector y, the discriminator accepts both, real or generated images, as well as the attribute-labels. We have the objective function of our model be:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x,y\sim p_{data}}[\log D(x, y)] + \mathbb{E}_{z\sim p_z, y\sim p_y}[\log(1 - D(G(z, y), y))]. \tag{1}$$

Table 1. Architecture of 2D Generator

Operation	Kernel	Stride	Filters	Norm	Activation
Concatenation	Concatenate z and y on 1st dimension				
ConvTranspose	4×4	2×2	512	BN	ReLU
ConvTranspose	4×4	2×2	256	BN	ReLU
ConvTranspose	4×4	2×2	128	BN	ReLU
ConvTranspose	4×4	2×2	64	BN	ReLU
ConvTranspose	4×4	2×2	3	No	Tahn

Table 2. Architecture of 2D Discriminator

Operation	Kernel	Stride	Filters	Norm	Activation
Conv	4×4	2×2	64	No	LeakyReLU
Concatenation	Replicate y and concatenate to 1st conv. layer				
Conv	4×4	2×2	128	SN	LeakyReLU
Conv	4×4	2×2	256	SN	LeakyReLU
Conv	4×4	2×2	512	SN	LeakyReLU
Conv	4×4	1×1	1	No	Sigmoid

3D model We expand the above presented 2D model onto 3D, in order to create a conditional 3DGAN (Figure 2), generating videos. In both, generator and discriminator, the convolutional kernels have been expanded onto three dimensions (H, W, T),

[3] http://www.uva-nemo.org/

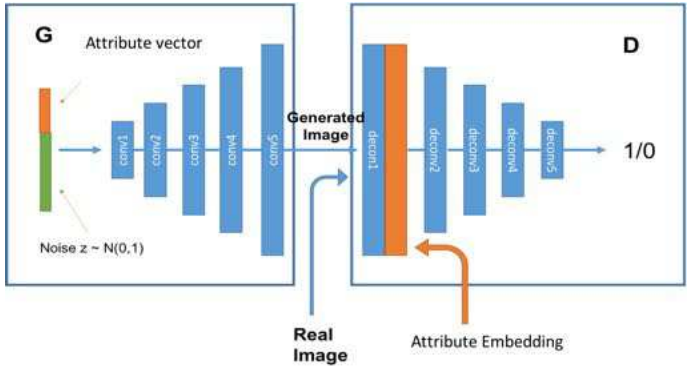


Fig. 1. Architecture of proposed 2D method consisting of two modules, a discriminator D and a generator G. While D learns to distinguish between real and fake images, classifying based on attribute-labels, G accepts as input both, noise and attribute-labels in order to generate realistic face images.

where H, W and T represent the height, the width and the temporal step of the receptive fields in each kernel.

We feed the attribute vectors into 3D model in a similar manner as the 2D model. Specifically, in the generator we concatenate the attribute vector with the noise vector. In the discriminator, the feature map after the first layer has the dimension of (H, W, C, T), each $(H, W, C, t), t \in T$ containing spatial-temporal features of a certain time period. Our goal is to generate face videos based on attributes, so we proceed to provide each spatial-temporal feature map with the same attribute embedding. Based on this, we insert an attribute embedding onto the spatial-temporal feature map from the first layer of the discriminator, creating a new feature map with the dimension $(H, W, C + y, T)$, where y is the dimension of the attribute vector.

Table 3. Architecture of 3D Generator

Operation	Kernel	Stride	Filters	Norm	Activation
Concatenation	Concatenate z and y on 1st dimension				
ConvTranspose3D	$2 \times 4 \times 4$	$2 \times 4 \times 4$	512	BN	ReLU
ConvTranspose3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	256	BN	ReLU
ConvTranspose3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	128	BN	ReLU
ConvTranspose3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	64	BN	ReLU
ConvTranspose3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	3	No	Tahn

Table 4. Architecture of 3D Discriminator

Operation	Kernel	Stride	Filters	Norm	Activation
Conv3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	64	No	LeakyReLU
Concatenation	Replicate y and concatenate to 1st conv. layer				
Conv3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	128	SN	LeakyReLU
Conv3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	256	SN	LeakyReLU
Conv3D	$4 \times 4 \times 4$	$2 \times 2 \times 2$	512	SN	LeakyReLU
Conv3D	$2 \times 4 \times 4$	$1 \times 1 \times 1$	1	No	Sigmoid

3 Experiments

2D model We train our network with the benchmark dataset CelebA [10] comprising of 202,599 face images annotated with 40 *binary* attribute labels. We generate images given five attributes glasses, gender, hair color, smile and age), in total 2048 images.

Figure 3 illustrates generated samples of the proposed approach. We observe that the model succeeds specifically in generation of local-attribute-labels based faces (*e.g.*, glasses, smile). The generated glasses appear to be similar, which is a limitation associated to DCGAN and the related loss function we use.

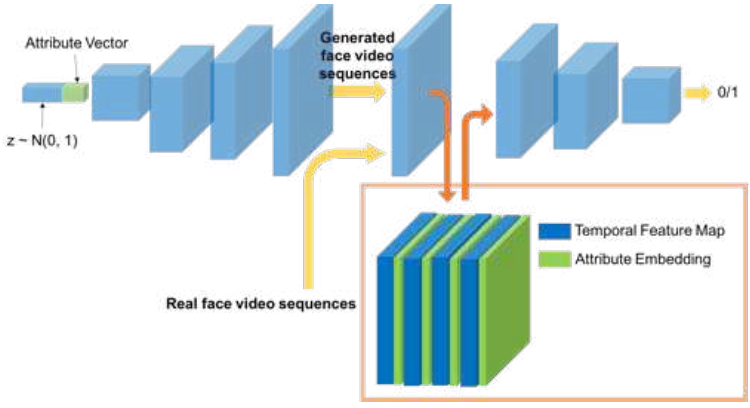


Fig. 2. Architecture of proposed 3D model for face video generation

3D model To train our 3DGAN model, we use the UvA-NEMO dataset [7]. It contains 1240 smile videos (597 spontaneous and 643 posed) from 400 subjects. Each subject has two attributes, gender and age. We label the subjects into two categories, adolescent (under 25 years old) and adults (above 25 years old). The generated smile video samples are portrayed in Figure 4.

4 Evaluation

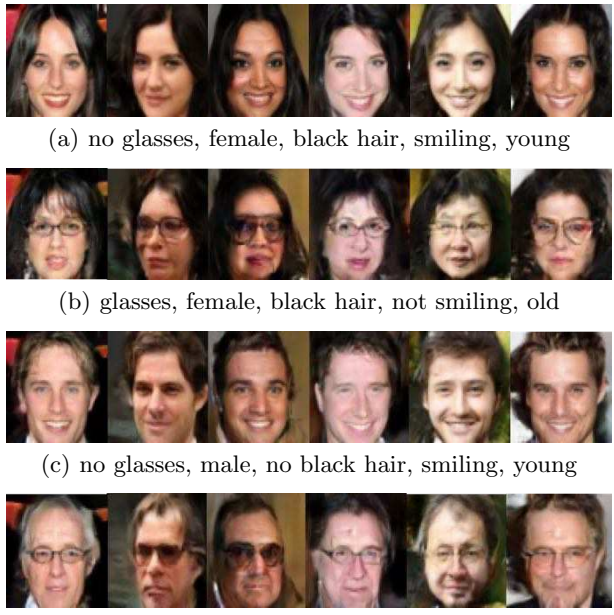
To evaluate how realistic our generated images are, in this section, we report the results based on the pre-trained face and attribute classification models. Then we proceed to present evaluation results of two quality metrics, namely Inception Score (IS) [12] and Fréchet Inception Distance (FID) [9], which have been widely used in image quality evaluation for GANs.

We obtain face detection results of up to 96% by DFace [6]) and gender true classification rate of up to 81.8% by the attribute classifier Face++ [8]. For age, note a shift in estimated age for the old / young labels, but the shift is not profound.

4.1 GAN quality metrics

We proceed to compute two commonly used GAN - quality measures, namely the Inception Score and the Fréchet Inception Distance, that we firstly proceed to describe. *Inception Score (IS)* is a metric for automated quality evaluation of images originated from generative models. The score is computed by an Inception V3 Network pre-trained on ImageNet and calculates a statistic of the network's outputs, when applied to generated images. *Fréchet Inception Distance (FID)* is an improvement of Inception Score. IS has the limitation of not utilizing statistics of real world samples are not used, and compared to the statistics of synthetic samples. Overcoming that, the Frćhet distance measures the distance between a generated image set and a source dataset.

We obtain $IS = 2.2$ and $FID = 43.8$. For IS, higher values indicate a higher quality, while for FID the opposite is the case. As a comparison Wasserstein GAN (WGAN)[1] reportedly achieves an $IS = 8.42$ and $FID = 55.2$; the Boundary Equilibrium Generative Adversarial Network BEGAN[2] obtains an $IS = 5.62$ and $FID = 71.4$.



(a) no glasses, female, black hair, smiling, young

(b) glasses, female, black hair, not smiling, old

(c) no glasses, male, no black hair, smiling, young

(d) glasses, male, no black hair, not smiling, old

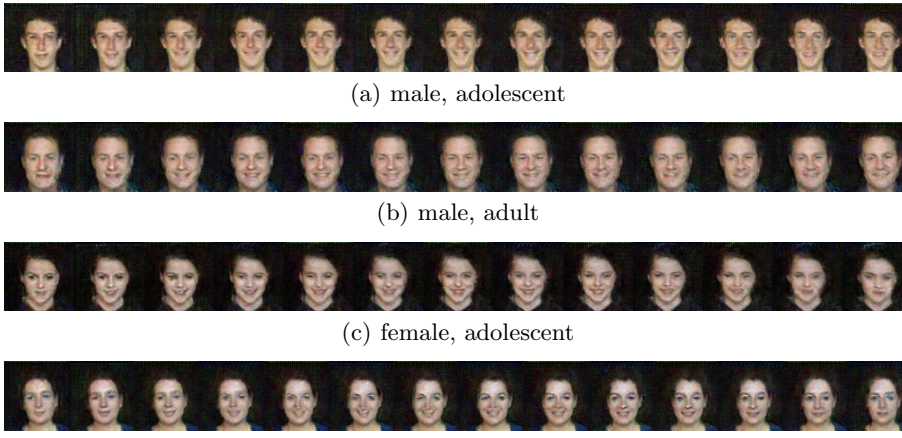
Fig. 3. Example images generated by the proposed 2D model.

5 Conclusions

In this work we presented 2D and 3D models for attribute based facial images and video generation, both based on DCGAN. Results, evaluated by a face detector, an attribute estimator and benchmark quality scores, suggest the models' ability to generate realistic faces from attribute labels. The presented approaches can be instrumental in the visualization of witness descriptions. Future work will involve experiments related to matching of generated faces with existing faces.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
2. Berthelot, D., Schumm, T., Metz, L.: Began: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)
3. Chen, C., Dantcheva, A., Ross, A.: Impact of facial cosmetics on automatic gender and age estimation algorithms. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP) (2014)
4. Dantcheva, A., Elia, P., Ross, A.: What else does your biometric data reveal? a survey on soft biometrics. IEEE Transactions on Information Forensics and Security pp. 1–26 (2015)
5. Dantcheva, A., Velardo, C., D'Angelo, A., Dugelay, J.L.: Bag of soft biometrics for person identification. Multimedia Tools and Applications **51**(2), 739–777 (2011)



(a) male, adolescent

(b) male, adult

(c) female, adolescent

(d) male, adult

Fig. 4. Chosen output samples from 3DGAN

6. DFace: Deeplearning face, https://github.com/kuaikuaikim/dface, 2018 (2018), `https://github.com/kuaikuaikim/DFace`
7. Dibeklioğlu, H., Salah, A.A., Gevers, T.: Are you really smiling at me? spontaneous versus posed enjoyment smiles. In: European Conference on Computer Vision. pp. 525–538. Springer (2012)
8. Face++: Face++ api, https://www.faceplusplus.com.cn/, 2018 (2018), `https://www.faceplusplus.com.cn/`
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR **abs/1706.08500** (2017), `http://arxiv.org/abs/1706.08500`
10. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738 (2015)
11. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355 (2016)
12. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 2234–2242. Curran Associates, Inc. (2016), `http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf`
13. Wang, Y., Dantcheva, A., Bremond, F.: From attributes to faces: a conditional generative adversarial network for face generation. In: International Conference of the Biometrics Special Interest Group (BIOSIG) (2017)