A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 1

# From attribute-labels to faces: text based face generation using conditional generative adversarial networks

**Abstract:** Recent advances in computer vision have aimed at *extracting* and *classifying* auxiliary biometric information such as age, gender, as well as health attributes, referred to as soft biometrics or *attributes*. We here seek to explore the inverse problem, namely *face generation based on attribute labels*, which is of interest due to related applications in law enforcement and entertainment. Particularly, we propose a method based on deep conditional generative adversarial network (DCGAN), which introduces additional data (*e.g.*, labels) towards determining specific representations of generated images. We present experimental results of the method, trained on the dataset CelebA, and validate these based on two GAN-quality-metrics, as well as based on three face detectors and one commercial off the shelf (COTS) attribute classifier. While these are early results, our findings indicate the method's ability to generate realistic faces from attribute labels.

**Keywords:** Attributes, soft biometrics, generative adversarial networks

## 1 Introduction

Recently *attributes* or *soft biometrics* [JDN04, DER15, Ni15] such as gender, age, ethnicity, height and weight have gained popularity due to their semantic interpretation, *i.e.,* they can provide a description that can be readily understood by humans; for example the description "young, female, tall". While the possibility to extract and classify such attributes has been shown, 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 descriptions, where the descriptions are the only available evidence (*e.g.*, in the absence of facial images). Currently in this context, law enforcement utilizes facial composites, which seek to support the process of suspect-identification. Composites are either hand drawn (sketches) or computer-generated, see Fig. 1. Both methods are slow, tedious relatively unrealistic, as well as impeding efficient face recognition (*i.e.*, match such composites with existing mugshot databases maintained by law enforcement agencies). Thus, reliable and automated label-based face generation would be a valuable asset in similar scenarios.

Additionally, label-based face generation accepts the application of visualization of classic fictional characters from novels. Often such characters are verbosely depicted, *e.g.*, "[His] jaw was long and bony, his chin a jutting 'v' under the more flexible v of his mouth. His nostrils curved back to make another, smaller, 'v.' His yellow-grey eyes were horizontal." [To12]. However there exists the hypothesis that humans are not able to imagine faces, which they have not seen before, and hence an automated visualization based on similar detailed descriptions might be beneficial for an enhanced reading experience.

Fig. 1: Composite images from FACES 4.0 [Sw14, FA14].

In spite of the aforementioned applications of interest, limited research concerns attribute-based face generation. In this context, particularly generative adversarial networks (GANs) and variational auto-encoders (VAE) are instrumental towards such face generation.

Motivated by the above, in this paper 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 an approach based on deep conditional convolutional generative adversarial network (DCGAN) [Pe16], which was introduced to *modify* images based on attributes (image-to-image translation, see Section 1). We train the proposed GAN with the dataset CelebA and generate faces pertained to two attribute-sets: (a) glasses and gender, as well as (b) glasses, gender, hair color, smile and age. We selected these two sets of attributes for the associated high descriptiveness, *e.g.*, such attributes are commonly used by humans to describe their peers. For the experiment (a) we generate 256 images and for (b) 2048 images, which we evaluate by 2 common GAN-quality-scores, as well as by three well established face detectors and an attribute classifier. Results indicate the method's ability th generate realistic faces from attribute labels.

**Related work** Generative adversarial networks (GANs), as introduced by Goodfellow *et al.* [Go14], incorporate two networks, a *Generator*, which generates new data instances and a *Discriminator*, which evaluates them for authenticity. The generator accepts noise as input and generates new samples of data in line with the observed training data. GANs have succeeded in applications such as image generation, image translation, super-resolution imaging, as well as face image synthesis.

Conditional GANs enhance the GAN-concept by providing both, the discriminator and generator with additional class information, in order to generate samples conditioned on different classes. Such *text-to-image translation* has been beneficial in domain transfer, super-resolution imaging, as well as image editing. Notable approaches include the conditional generative adversarial networks (cGANs) work by Mirza and Osindero [MO14] and the invertible conditional GAN proposed by Perarnau *et al.* [Pe16].

Limited literature concerns attribute-labels based face generation. The work of Gauthier [Ga14] tackles the topic by proposing a conditional GAN in this context and losely validating the approach *qualitatively* on Labeled Faces in the Wild (LFW). Deviating from our work, Gauthier added the attribute-information at the last layer of the *Discrimator*, did not use the set "real images and fake labels" for training, as well as lacked proficient validation. We note here that we tested their architecture and obtained poor results. A further

related work presents the VAE-based framework Attribute2Image [Ya16] proposed by Yan *et al.*

As stated above, while limited research concerns attribute-based face generation, extensive research relates to attribute based *image-to-image translation*. In this context, *image-to-image translation* refers to the alteration of a particular aspect of a given image to another, *e.g.*, changing the hair color. Notable very recent attribute-translation GANs include CycleGAN [CZS17], StarGAN [Ch17], as well as MakeupGAN [Ch18].

Finally, *face sketch synthesis* (FSS) has gained interest in the design community (*e.g.*, face artistic styles synthesis [Zh18, Zh16]) and is being used for digital entertainment, as well as law enforcement (sketch based face recognition [Kl14]). A recent VAE and GAN-based face synthesis approach from sketches is the work of Di and Patel [DP17].

## 2 Proposed method

The proposed conditional GAN aims to fit the conditional probability $P(x|z,y)$, as depicted in Figure 2 and specified in the Tables 1 and 2. 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:

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

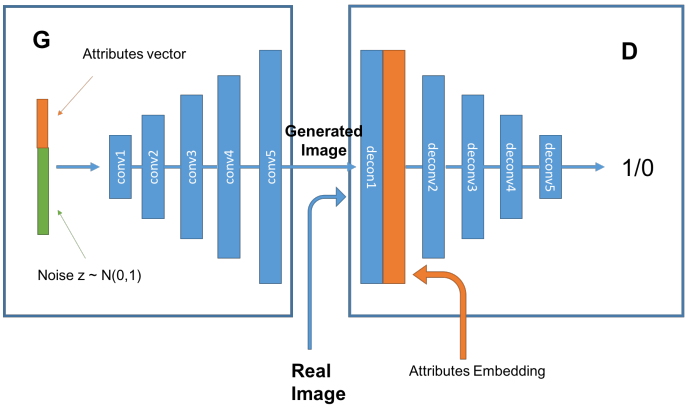


Fig. 2: Architecture of proposed method consisting of two modules, a discriminator $D$ and a generator $G$. $D$ learns to distinguish between real and fake images, classifying based on attribute-labels. $G$ accepts as input both, noise and attribute-labels in order to generate realistic face images.

### 2.1 Implementation details

Our model is based on the Pytorch implementation of DCGAN. The generator accepts as the input the concatenation of attribute-label-vector $y$ and noise $z$. On the discriminator-side, we stack the attribute-label-vector with the feature map of the first convolutional layer. We tested the insertion of the attribute-label-vector at each discriminator-layer and obtained best results at the first layer-level. We train the model with the Adam optimizer at a learning rate of 0.0002 and a mini batch size of 128. The input and generated image sizes are both $64 \times 64$.

Tab. 1: Architecture of Generator

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

Tab. 2: Architecture of Discriminator

| Operation | Kernel | Stride | Filters | BN | Activation |
|---|---|---|---|---|---|
| Conv | $4 \times 4$ | $2 \times 2$ | 64 | No | LeakyReLU |
| Concatenation | Replicate y and concatenate to 1st conv. layer | | | | |
| Conv | $4 \times 4$ | $2 \times 2$ | 128 | Yes | LeakyReLU |
| Conv | $4 \times 4$ | $2 \times 2$ | 256 | Yes | LeakyReLU |
| Conv | $4 \times 4$ | $2 \times 2$ | 512 | Yes | LeakyReLU |
| Conv | $4 \times 4$ | $1 \times 1$ | 1 | No | Sigmoid |

Algorithm 1 summarizes the training procedure. When we train the discriminator, we provide two types of negative samples. Firstly the generated images with real labels, secondly real images with fake labels. We observe that such training enforces the discriminator to learn from diverse samples and at the same time enables the generation of realistic samples.

**Algorithm 1** Attributes based Conditional DCGAN

**Input:** minibatch $x$, matching label $y$, mismatching label $\hat{y}$, number of training epochs $N$

1: **for** $n = 1$ to $N$ **do**
2: $z \sim \mathcal{N}(0,1)^Z$ {Draw sample of random noise}
3: $y \leftarrow \mathscr{Y}$ {Draw matching attributes label}
4: $\hat{y} \leftarrow \hat{\mathscr{Y}}$ {Draw mismatching attributes label}
5: $\hat{x} \leftarrow G(z,y)$ {Forward through generator}
6: $s_r \leftarrow D(x,y)$ {Forward through discriminator with real image, right attribute labels}
7: $s_w \leftarrow D(x,\hat{y})$ {Forward through discriminator with real image, wrong attribute labels}
8: $s_g \leftarrow D(\hat{x},y)$ {Forward through discriminator with generated image, right attribute labels}
9: $\mathscr{L}_D \leftarrow \log(s_r) + (\log(1-s_w) + \log(1-s_g))/2$ {loss function of discriminator }
10: $D \leftarrow D - \alpha \partial \mathscr{L}_D / \partial D$ {Update discriminator}
11: $\mathscr{L}_G \leftarrow \log(s_f)$ {loss function of generator}
12: $G \leftarrow G - \alpha \partial \mathscr{L}_G / \partial G$ {Update generator}
13: **end for**

## 3 Experiments

We use the benchmark dataset CelebA [Li15] comprising of 202,599 face images annotated with 40 *binary* attribute labels. We perform two experiments: generation of images given (a) two attributes

(glasses, gender) and (b) five attributes (glasses, gender, hair color, smile and age). For (a) we generate 256 images and for (b) 2048 images.

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

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

Fig. 3: Chosen output samples based on 5 attributes

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. Further, based on Figure 4, we found binary labels might be too limiting in describing age. Such labels can guide the model to generate 'younger' or 'older' faces, not exactly 'young' and 'old' faces. In addition, CelebA depicts celebrities, with face alterations inflicted by makeup and plastic surgery, which might affect the reliability of the pair "age-label and image".

We witnessed a 'mode collapse' in both experiments after 100 epochs, which has been reported as the main limitation of GANs. One explanation for that is that the generator *deceived* the discriminator by generating similar fake images.

### 3.1 Evaluation

To evaluate how realistic our generated images are, in this section, firstly we report the results based on the pre-trained face and attribute detection models. Then we proceed to present evaluation results of two quality metrics, namely Inception Score (IS) [Sa16] and Fréchet Inception Distance (FID) [He17], which have been widely used in image quality evaluation for GANs.

**Face Detection** We report face detection results based on 3 different face detectors (Face ++ [Fa18], DFace [DF18], dlib [dl18]) in Table 3. We observe that DFace has the highest detection rates (for experiment (a) 89%, for (b) 96%), while Face++ has the lowest rates ((a) 58.2%, (b) 53%). This might be due to the fact that DFace was trained using the CelebA dataset (generated faces are of a similar probability distribution as the training data).

**Attribute Estimation** We present in Table 4 gender classification results from Face ++. We note the higher true gender classification accuracy in the experiment (b) (generating 5 attributes), which indicates that higher prior information allows the model to generate better targeted images. Figure 4 illustrates the boxplot of estimated age for age estimation by Face ++. Here, we note a shift in estimated age for the old / young labels, but the shift is not profound.

**GAN quality metrics** We proceed with 2 commonly used GAN - quality measures, namely the Inception Score and the Fréchet Inception Distance, that we firstly proceed to describe.

Tab. 3: Face detection results of generated faces. We report the number of detected faces for the three face detectors: dlib, DFace and Face ++. In experiment (a) we generate faces labeled with two attributes corresponding to *gender* and *glasses*. Experiment (b) involves 5 attributes corresponding to *gender*, *glasses*, *age*, *smile*, *hair color*.

| **Number of Attributes** | **Generated Images** | **Detected Faces dlib** | **Detected Faces DFace** | **Detected Faces Face ++** |
|---|---|---|---|---|
| (a) 2 | 256 | 193 | 228 | 149 |
| (b) 5 | 2048 | 1794 | 1966 | 1085 |

Tab. 4: Gender estimation accuracy of generated faces as computed by Face++. In experiment (a) we generate faces labeled with two attributes corresponding to *gender* and *glasses*. Experiment (b) involves 5 attributes corresponding to *gender*, *glasses*, *age*, *smile*, *hair color*.

| **Number of Attributes** | **Detected Faces** | **True Gender Classification Accuracy (%)** | **False Male Rate (%)** | **False Female Rate (%)** |
|---|---|---|---|---|
| 2 | 149 | 66.4 | 74.6 | 59.7 |
| 5 | 1085 | 81.8 | 94.7 | 67.2 |

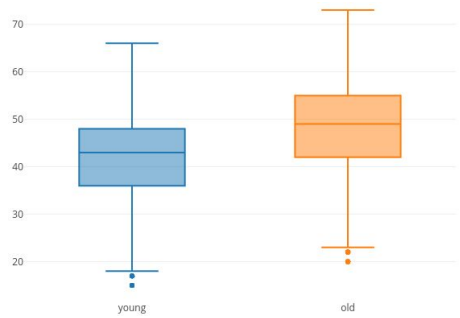


Fig. 4: Boxplot of the estimated age pertained to generated faces as computed using the Face ++ attribute classifier.

*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. The Inception Score is given by

$$IS(G) = \exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x)\|p(y))), \tag{2}$$

where $x \sim p_g$ indicates that $x$ is an image sampled from $p_g$, $D_{KL}(p\|q)$ is the KL-divergence between the distributions $p$ and $q$. The score examines generated images for included meaningful objects, as well as a low label entropy.

*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, and is calculates as

$$d^2((m,C),(m_\omega,C_\omega)) = \|m-m_\omega\|_2^2 + Tr(C+C_\omega - 2(CC_\omega)_{\frac{1}{2}}), \tag{3}$$

where $(m,C)$, $(m_\omega,C_\omega)$ represent the mean and covariance of the two respective distributions.

*IS* and *FID* for the both experiments are summarized in Table 5. For IS, higher values indicate a higher quality, while for FID the opposite is the case. Again, we observe that an increased prior information (experiment (b) improves the generated image quality. As a comparison Wasserstein GAN (WGAN) reportedly has $IS = 8.42$ and $FID = 55.2$; the Boundary Equilibrium Generative Adversarial NetworksBEGAN $IS = 5.62$ and $FID = 71.4$.

Tab. 5: Inception Score (IS) and Fréchet Inception Distance (FID) quantitatively validating generated images for the two experiments pertained to 2 and 5 attributes.

| **Number of Attributes** | *IS* | *FID* |
|---|---|---|
| 2 | 1.94 | 63.8 |
| 5 | 2.20 | 43.8 |

## 4 Conclusions

In this work we presented preliminary results on attribute based face generation based on a DCGAN-approach. Results, evaluated by three well established face detectors, an attribute estimator and benchmark quality scores, suggest the method's ability to generate realistic faces from attribute labels. The presented approach can be instrumental in the visualization of witness descriptions. Future work will involve the enhancement of the architecture by an Adversarial Auto-encoder (AAE) in order to improve image quality.

## References

[Ch17] Choi, Yunjey; Choi, Minje; Kim, Munyoung; Ha, Jung-Woo; Kim, Sunghun; Choo, Jaegul: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. arXiv preprint arXiv:1711.09020, 2017.

[Ch18] Chang, Huiwen; Lu, Jingwan; Yu, Fisher; Finkelstein, Adam: PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In: Conference on Computer Vision and Pattern Recognition. 2018.

[CZS17] Chu, Casey; Zhmoginov, Andrey; Sandler, Mark: CycleGAN: a Master of Steganography. arXiv preprint arXiv:1712.02950, 2017.

[DER15] 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.

[DF18] DeepLearning Face, https://github.com/kuaikuaikim/DFace, 2018.

[dl18] dlib, http://dlib.net/, 2018.

[DP17] Di, Xing; Patel, Vishal M: Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs. arXiv preprint arXiv:1801.00077, 2017.

[FA14] FACES 4.0, http://www.iqbiometrix.com, 2018.

[Fa18] Face++ API, https://www.faceplusplus.com.cn/, 2018.

[Ga14] Gauthier, Jon: Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014(5):2, 2014.

[Go14] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680, 2014.

[He17] Heusel, Martin; Ramsauer, Hubert; Unterthiner, Thomas; Nessler, Bernhard; Klambauer, Günter; Hochreiter, Sepp: GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. CoRR, abs/1706.08500, 2017.

[JDN04] Jain, A. K.; Dass, S. C.; Nandakumar, K.: Soft biometric traits for personal recognition systems. In: Proc. of ICBA. 2004.

[Kl14] Klum, Scott J; Han, Hu; Klare, Brendan F; Jain, Anil K: The FaceSketchID system: Matching facial composites to mugshots. IEEE Transactions on Information Forensics and Security, 9(12):2248–2263, 2014.

[Li15] Liu, Ziwei; Luo, Ping; Wang, Xiaogang; Tang, Xiaoou: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3730–3738, 2015.

[MO14] Mirza, Mehdi; Osindero, Simon: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[Ni15] Nixon, Mark S; Correia, Paulo L; Nasrollahi, Kamal; Moeslund, Thomas B; Hadid, Abdenour; Tistarelli, Massimo: On soft biometrics. Pattern Recognition Letters, 68:218–230, 2015.

[Pe16] Perarnau, Guim; van de Weijer, Joost; Raducanu, Bogdan; Álvarez, Jose M: Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355, 2016.

[Sa16] Salimans, Tim; Goodfellow, Ian; Zaremba, Wojciech; Cheung, Vicki; Radford, Alec; Chen, Xi; Chen, Xi: 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.

[Sw14] Sweegers, Carly CG; Takashima, Atsuko; Fernández, Guillén; Talamini, Lucia M: Neural mechanisms supporting the extraction of general knowledge across episodic memories. Neuroimage, 87:138–146, 2014.

[To12] Tolkien, John Ronald Reuel: The Lord of the Rings: One Volume. Houghton Mifflin Harcourt, 2012.

[Ya16] Yan, Xinchen; Yang, Jimei; Sohn, Kihyuk; Lee, Honglak: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision. Springer, pp. 776–791, 2016.

[Zh16] Zhang, Shengchuan; Gao, Xinbo; Wang, Nannan; Li, Jie: Robust face sketch style synthesis. IEEE Transactions on Image Processing, 25(1):220–232, 2016.

[Zh18] Zhang, Mingjin; Li, Jie; Wang, Nannan; Gao, Xinbo: Compositional model-based sketch generator in facial entertainment. IEEE transactions on cybernetics, 48(3):904–915, 2018.