Generating Photo-Realistic Training Data to Improve Face Recognition Accuracy

Daniel Sáez Trigueros[a], Li Meng[a,*], Margaret Hartnett[b]

[a]*School of Engineering and Technology, University of Hertfordshire, Hatfield AL10 9AB, UK*
[b]*Hartnett Innovations Ltd*

Abstract

Face recognition has become a widely adopted biometric in forensics, security and law enforcement thanks to the high accuracy achieved by systems based on convolutional neural networks (CNNs). However, to achieve good performance, CNNs need to be trained with very large datasets which are not always available. In this paper we investigate the feasibility of using synthetic data to augment face datasets. In particular, we propose a novel generative adversarial network (GAN) that can disentangle identity-related attributes from non-identity-related attributes. This is done by training an embedding network that maps discrete identity labels to an identity latent space that follows a simple prior distribution, and training a GAN conditioned on samples from that distribution. A main novelty of our approach is the ability to generate both synthetic images of subjects in the training set and synthetic images of new subjects not in the training set, both of which we use to augment face datasets. By using recent advances in GAN training, we show that the synthetic images generated by our model are photo-realistic, and that training with datasets augmented with those images can lead to increased recognition accuracy. Experimental results show that our method is more effective when augmenting small datasets. In particular, an absolute accuracy improvement of 8.42% was

*Corresponding author
Email address: l.1.meng@herts.ac.uk (Li Meng)

achieved when augmenting a dataset of less than 60k facial images.

Keywords: image generation, generative adversarial learning, face and gesture recognition, machine learning

---

1. Introduction

Recent progress in machine learning has made possible the development of face recognition systems that can match face images as good as or better than humans. For this reason, these systems have become a valuable tool in forensics and security. However, state-of-the-art face recognition systems based on convolutional neural networks (CNNs) need to be trained with very large datasets of face images. In this work we aim to reduce the data requirement of face recognition systems by synthesising artificial face images.

Image synthesis is a widely studied topic in computer vision. In particular, face image synthesis has gained a lot of attention because of its diverse practical applications. These include facial image editing (Larsen et al. 2016; Yan et al. 2016; Perarnau et al. 2016; Brock et al. 2017; Zhang et al. 2017; Antipov et al. 2017; Choi et al. 2018; Shu et al. 2017; Lample et al. 2017), face de-identification (Meden et al. 2017, 2018; Brkic et al. 2017; Wu et al. 2019), data augmentation (Masi et al. 2016; Banerjee et al. 2017; Osadchy et al. 2017; Masi et al. 2019; Zhao et al. 2018; Kortylewski et al. 2018; Mokhayeri et al. 2018), face frontalisation (Zhu et al. 2013, 2014; Hassner et al. 2015; Zhu et al. 2015; Yim et al. 2015; Tran et al. 2018; Huang et al. 2017) and artistic applications (e.g. video games and advertisements).

In this work, we focus on the applicability of face image synthesis for data augmentation. It is widely known that training data is one of the most important factors that affect the accuracy of deep learning models. The datasets used for training need to be large and contain sufficient variation to allow the resulting models to learn features that generalise well to unseen samples. In the case of

face recognition, the datasets must contain many different subjects, as well as many different images per subject. The first requirement enables a model to learn inter-class discriminative features that can generalise to subjects not in the training set. The second requirement enables a model to learn features that are robust to intra-class variations. Even though there are several public large-scale datasets (Sun et al. 2014; Yi et al. 2014; Parkhi et al. 2015; Guo et al. 2016; Nech & Kemelmacher-Shlizerman 2017; Bansal et al. 2017; Cao et al. 2018) that can be used to train CNN-based face recognition models, these datasets are nowhere near the size or quality of commercial datasets. For example, the largest publicly available dataset contains about 10M images of 100K different subjects (Guo et al. 2016), whereas Google's FaceNet (Schroff et al. 2015) was trained with a private dataset containing between 100M and 200M face images of about 8M different subjects. Another issue is the presence of long-tail distributions in some publicly available datasets, i.e. datasets in which there are many subjects with very few images. Such unbalanced datasets can make the training process difficult and result in models that achieve lower accuracy than those trained with smaller but balanced datasets (Zhao et al. 2018; Zhou et al. 2015). In addition, some publicly available datasets (e.g. Guo et al. 2016) contain many mislabelled samples that can decrease face recognition accuracy if not discarded from the training set. Since collecting large-scale, good quality face datasets is a very expensive and labour-intensive task, we propose a method for generating photo-realistic face images that can be used to effectively increase the depth (number of images per subject) and width (number of subjects) of existing face datasets.

An approach that has recently gained popularity for augmenting face datasets is the use of 3D morphable models (Blanz & Vetter 2003). In this approach, new faces of existing subjects can be synthetized by fitting a 3D morphable model to existing images and modifying a variety of parameters to generate new poses and expressions (Masi et al. 2016, 2019; Zhao et al. 2018; Mokhayeri et al. 2018). It is also possible to generate images with other variations using this approach. For example, Mokhayeri et al. (2018) incorporated a reflectance

model to generate images under different lighting conditions; and Kortylewski et al. (2018) randomly sampled 3D face shapes and colours to generate faces of new subjects. The main drawback of methods based on 3D morphable models is that the generated images often look unnatural and lack the level of detail found in real images. Another recent approach based on blending small triangular regions from different training images was proposed in Banerjee et al. (2017). Although this method seemed to produce photo-realistic faces, the authors limited their work to frontal face images. In contrast, our approach makes use of generative adversarial networks (GANs) (Goodfellow et al. 2014; Schmidhuber 2020), which have recently been shown to produce photo-realistic in-the-wild images often indistinguishable from real images (Karras et al. 2018). Another advantage of using GANs is that they are end-to-end trainable models that do not require any domain-specific processing, as opposed to methods based on 3D modelling or face triangulation.

Many methods based on GANs have been proposed for manipulating attributes of existing face images, including age (Yan et al. 2016; Zhang et al. 2017; Antipov et al. 2017), facial expressions (Yan et al. 2016; Choi et al. 2018; Zhou & Shi 2017; Ding et al. 2018), and other attributes such as hairstyle, glasses, makeup, facial hair, skin colour or gender (Larsen et al. 2016; Perarnau et al. 2016; Brock et al. 2017; Choi et al. 2018; Shu et al. 2017; Lample et al. 2017; Shen & Liu 2017; Lu et al. 2018; He et al. 2017). While these methods can be used to increase the depth of a dataset, it remains unclear how to increase the width of a dataset, i.e. how to generate faces of new subjects. Our proposed GAN is able to generate faces from a latent representation $\boldsymbol{z}$ that has two gaussian distributed components $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{nid}$ encoding identity-related attributes and non-identity-related attributes respectively. In this way, face images of new subjects can be generated by fixing the identity component $\boldsymbol{z}_{id}$ and varying the non-identity component $\boldsymbol{z}_{nid}$. The method most closely related to ours is the *semantically decomposed GAN* (SD-GAN) proposed in Donahue et al. (2018). SD-GANs are trained with pairs of real images from the same subject and pairs of images generated with the same identity-related attributes but different non-

identity-related attributes. A discriminator learns to reject pairs of images when either they do not look photo-realistic or when they do not appear to belong to the same subject. One of the main differences of our method with respect to SD-GANs is that it allows the generation of face images of subjects that exist in the training set. In other words, our method can increase both the width and the depth of a given face dataset. Furthermore, our proposed GAN is arguably simpler to implement than SD-GAN and easier to incorporate into other GAN architectures.

To demonstrate the efficacy of our method, we trained several CNN-based face recognition models with different combinations of real and synthetic data. In most cases, the models trained with a combination of real and synthetic data outperformed the models trained with real data alone.

Our main contributions can be summarised as:

- A novel face image synthesis method based on GANs that allows the disentangling of identity-related attributes from non-identity-related attributes.
- A data augmentation approach that uses the proposed GAN to increase the depth and width of existing face datasets.
- A experimental demonstration that the proposed data augmentation approach can be used to increase the accuracy of a face recognition algorithm trained with real images alone.

The rest of this paper is organised as follows. Section 2 provides the background needed to understand our proposed GAN. Section 3 explains each part of our proposed GAN and the loss functions used for training. Section 4 discusses our experimental results, both in terms of the quality of the synthetic images generated by our proposed GAN and the accuracy achieved by datasets augmented with the synthetic images. Finally, our conclusions are presented in Section 5.

2. Background

Generative adversarial networks (GANs) generate data by sampling from a probability distribution p_{model} that is trained to match a true data generating distribution p_{data}. This is done by mapping a vector of random latent variables $\boldsymbol{z} \sim p_{\boldsymbol{z}}$ to a sample $G(\boldsymbol{z})$ through a generator network G, where $p_{\boldsymbol{z}}$ is a prior distribution that can be easily sampled (e.g. Gaussian or uniform). The generator is trained to fool a discriminator network D that tries to determine whether a sample is real or generated (i.e. synthetic). Thus, the generator and discriminator are trained with opposing optimisation objectives. While the discriminator is trained to maximise the probability of correctly classifying both real and generated samples, the generator is trained to minimise the probability that the generated samples are classified as such. Formally, the standard GAN optimisation objective can be expressed as follows:

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x} \sim p_{data}}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}}[\log(1 - D(G(\boldsymbol{z})))] \tag{1}$$

As training progresses, the discriminator gets better at distinguishing real from generated samples and the generator gets better at producing realistic samples that can fool the discriminator. The training is considered completed when the generator and the discriminator reach an equilibrium, i.e. when the generator and the discriminator stop improving. In practice, since GANs rarely reach an equilibrium, it is common to simply stop the training process whenever there is no noticeable improvement in the visual quality of the generated samples.

The training of GANs is often unstable and can lead to the mode collapse problem (this happens when the generator maps different values of $\boldsymbol{z}$ to the same output sample (Goodfellow 2016)). Although some works have proposed heuristics that reduce this effect Radford et al. (2015); Salimans et al. (2016), a full understanding of the training dynamics of GANs remains an open research question. Based on the idea that optimising the training objective in (1) can be interpreted as minimising the Jensen-Shannon divergence between the true data generating distribution p_{data} and the model distribution p_{model}

(Goodfellow et al. 2014), Arjovsky et al. (2017) proposed a novel training objective for GANs that minimises the Wasserstein distance instead of the Jensen-Shannon divergence. This GAN variation, which was named the Wasserstein GAN (WGAN), was shown to be more stable and to reduce the mode collapse problem (Arjovsky et al. 2017). An improved formulation of this approach (Gulrajani et al. 2017) is considered one of the current state-of-the-art techniques for training GANs.

Many works on GANs have adopted a family of architectures known as *deep convolutional GANs* (DCGANs) (Radford et al. 2015). DCGANs follow a set of guidelines that were proposed for stable training and good image quality. More recently, Karras et al. (2018) proposed a new methodology for training GANs that consist of progressively growing both the spatial resolution of real and generated images and the number of layers of the discriminator and generator networks (PGGAN). In this manner, the training is very stable at the start since low-resolution images are easier to generate than high-resolution images due to their lower dimensionality and hence diversity. As training progresses, and the resolution of the images is increased, the generator gradually learns to generate images with finer detail. In contrast, standard GANs that are tasked with learning high-resolution images from the outset are typically more unstable. Using the PGGAN approach together with several proposed heuristics, Karras et al. (2018) were able to generate impressive photo-realistic 1024×1024 images with a $5.4\times$ speedup factor with respect to the standard GAN training approach.

Conditional versions of GANs allow the generation of samples with specific attributes. The first conditional GAN was introduced in Mirza & Osindero (2014) and consisted of feeding a label $\boldsymbol{y}$ encoding some attribute(s) of the data to both the generator and the discriminator. An alternative type of conditional GAN called auxiliary classifier GAN (AC-GAN) was proposed in (Odena et al. 2017). Instead of feeding the label $\boldsymbol{y}$ to the discriminator, AC-GANs use an auxiliary classifier in the discriminator that is tasked with predicting the label $\boldsymbol{y}$ that has been fed to the generator. The use of an auxiliary classifier to predict $\boldsymbol{y}$ was shown in Odena (2016) and Salimans et al. (2016) to improve

image quality even when $\boldsymbol{y}$ was not fed to the generator.

An approach related to conditional GANs is the InfoGAN model proposed in Chen et al. (2016). The goal of an InfoGAN is to disentangle data attributes in an unsupervised way. This is done by maximising the mutual information between a subset $\boldsymbol{c}$ of the latent variables $\boldsymbol{z}$ and the generated image $G(\boldsymbol{z})$. InfoGANs can be implemented with an auxiliary network in the discriminator that is trained to predict $\boldsymbol{c}$. As shown in Chen et al. (2016), this method ensures that the latent variables $\boldsymbol{c}$ encode meaningful data attributes that are not lost during the generation process. Our proposed GAN incorporates elements of both conditional GANs and InfoGANs to disentangle identity-related attributes from non-identity-related attributes.

3. Proposed Method

In this section, we first explain our choice of GAN architecture and type of conditional GAN, and then our proposed modifications to disentangle identity-related attributes from non-identity-related attributes. Finally, we explain how we use the proposed GAN for augmenting existing face datasets.

3.1. Conditional PGGAN

We follow the same architecture and training method proposed in the PGGAN work (Karras et al. 2018) (we use the open-source code released by the authors and keep the default training settings unless otherwise stated) and add our modifications to it. The training starts by generating 4×4 images. The number of layers in the generator and discriminator is then gradually increased from 1 to 11 (each time the resolution is doubled, two convolutional layers are added) until 128×128 images are generated. We do not generate higher resolution images because the network that we use for face recognition in our experiments takes 100×100 images as input. Higher resolution images could be generated by simply increasing the number of layers in the PGGAN generator and in the face recognition network accordingly. Following Karras et al.

(2018), instead of using the standard GAN objective proposed in (1), we use the WGAN training objective proposed in Gulrajani et al. (2017):

$$\min_{G} \max_{D \in \mathcal{D}} \mathbb{E}_{\boldsymbol{x} \sim p_{data}}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}}[D(G(\boldsymbol{z}))] \tag{2}$$

where $\mathcal{D}$ is the set of 1-Lipschitz functions (for a full derivation of (2) see Arjovsky et al. (2017)). When posed as a minimisation problem and adding a gradient penalty that enforces the Lipschitz constraint (the gradient of 1-Lipschitz functions are bounded to 1), the WGAN loss (Gulrajani et al. 2017) becomes:

$$L_{img} = \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}}[D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{x} \sim p_{data}}[D(\boldsymbol{x})] + \lambda \, \mathbb{E}_{\tilde{\boldsymbol{x}} \sim p_{\tilde{\boldsymbol{x}}}}[\|(\nabla_{\tilde{\boldsymbol{x}}} D(\tilde{\boldsymbol{x}})\|_2 - 1)^2] \tag{3}$$

where $p_{\tilde{\boldsymbol{x}}}$ is a distribution of samples interpolated from generated samples $G(\boldsymbol{z})$ and real samples $\boldsymbol{x}$, and λ is a weight controlling the contribution of the gradient penalty to the loss. Following Gulrajani et al. (2017), we set the value of λ to 10.

To make this model conditional, we use the AC-GAN method, i.e. the generator is conditioned on identity labels $\boldsymbol{y}$ and an auxiliary network D_c in the discriminator D is trained to predict $\boldsymbol{y}$. Since the identity labels $\boldsymbol{y}$ are categorical, D_c is trained as a classifier with cross-entropy loss:

$$L_c = -\mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}, \boldsymbol{y} \sim p_{\boldsymbol{y}}}[\log D_c(G(\boldsymbol{z} \mid \boldsymbol{y}))] - \mathbb{E}_{\boldsymbol{x} \sim p_{data}}[\log D_c(\boldsymbol{x})] \tag{4}$$

Note that the cross-entropy loss is applied to the real images $\boldsymbol{x}$ and the synthetic images $G(\boldsymbol{z} \mid \boldsymbol{y})$. This loss encourages the generator to generate images with the correct identity labels $\boldsymbol{y}$ so that they can be predicted by the discriminator instead of being ignored during the generation process.

3.2. Identity Latent Space

The model described in Section 3.1 can generate images of subjects with identities $\boldsymbol{y}$ existing in the training set. However, it is not possible to generate images of subjects with new identities. For this reason, we propose the use of an embedding network E to map the discrete identity labels $\boldsymbol{y}$ to a vector of

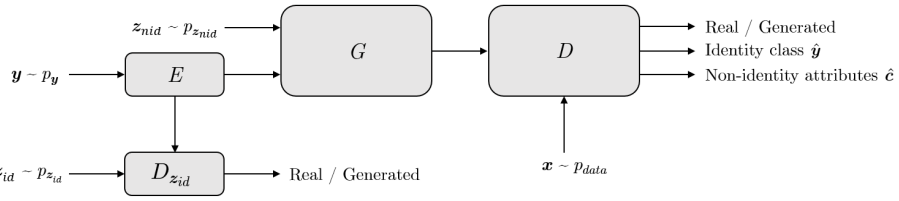


Figure 1: During training, the generator G takes as an input a vector of latent variables $E(\boldsymbol{y})$ encoding the identity-related attributes of subject $\boldsymbol{y}$ and a vector of latent variables $\boldsymbol{z}_{nid}$ encoding non-identity-related attributes. The discriminator $D_{\boldsymbol{z}_{id}}$ is used to match $E(\boldsymbol{y})$ to the prior distribution $p_{\boldsymbol{z}_{id}}$. The discriminator D is used to encourage G to generate photo-realistic images and has two auxiliary outputs predicting the identity of the subject $\boldsymbol{y}$ and the non-identity-related attributes $\boldsymbol{z}_{nid}$.

latent variables $E(\boldsymbol{y})$. Since the goal is to learn a continuous latent space of identities that we can sample from, $E(\boldsymbol{y})$ is trained to follow an easy to sample prior distribution $p_{\boldsymbol{z}_{id}}$. In our case, we assume that the identity latent space is normally distributed and choose to use a Gaussian prior $p_{\boldsymbol{z}_{id}} = \mathcal{N}(0, 1)$. This can be done by using another discriminator $D_{\boldsymbol{z}_{id}}$ that is trained to match the posterior distribution defined by $E(\boldsymbol{y})$ to the Gaussian prior distribution $p_{\boldsymbol{z}_{id}}$ using adversarial training, as proposed in Makhzani et al. (2016). Note that, alternatively, this can be done by adding a Kullback–Leibler divergence (KLD) term to the loss to match the posterior to the prior, as done in the variational autoencoder framework (Kingma & Welling 2013). In our experiments, we did not notice differences between these two methods and decided to use the discriminator approach due to its additional flexibility (Makhzani et al. 2016). To incorporate the identity latent space into the conditional PGGAN model from Section 3.1, the generator network G is conditioned on the latent representation of the labels, i.e. $G(\boldsymbol{z} \mid E(\boldsymbol{y}))$. A diagram of our proposed GAN is shown in Fig. 1. The aforementioned modifications to the standard AC-GAN architecture are shown on the left side of Fig. 1.

We choose to train the embedding network E to learn a stochastic mapping with Gaussian noise rather than a deterministic mapping. This is because in the deterministic case, E can only use the stochasticity of the identity labels in the

training set (which is fixed and typically very limited) to map the posterior distribution defined by $E(\boldsymbol{y})$ to the Gaussian prior distribution $p_{\boldsymbol{z}_{id}}$. Therefore, a deterministic mapping might not yield a smooth posterior distribution. In contrast, the additional randomness introduced by Gaussian noise in the stochastic mapping can alleviate this issue. With a stochastic mapping, the output of the embedding network E is a vector of means and variances that is used to produce samples that must be indistinguishable from samples from the Gaussian prior distribution $p_{\boldsymbol{z}_{id}}$. In order to allow backpropagation through the sampling operation, the reparametrization trick proposed in Kingma & Welling (2013) is used.

To train the embedding network E and the discriminator network $D_{\boldsymbol{z}_{id}}$, we again use the WGAN loss:

$$L_e = \mathbb{E}_{\boldsymbol{y}\sim p_y}[D_{\boldsymbol{z}_{id}}(E(\boldsymbol{y}))] - \mathbb{E}_{\boldsymbol{z}_{id}\sim p_{\boldsymbol{z}_{id}}}[D_{\boldsymbol{z}_{id}}(\boldsymbol{z}_{id})] + \lambda_e \, \mathbb{E}_{\tilde{\boldsymbol{y}}\sim p_{\tilde{\boldsymbol{y}}}}[(\|\nabla_{\tilde{\boldsymbol{y}}} D_{\boldsymbol{z}_{id}}(\tilde{\boldsymbol{y}})\|_2 - 1)^2] \quad (5)$$

where, in this case, $p_{\tilde{\boldsymbol{y}}}$ is a distribution of samples interpolated from the latent representation of the labels $E(\boldsymbol{y})$ and Gaussian samples $\boldsymbol{z}_{id}$. In a similar manner to (3), λ_e is set to 10.

3.3. Mutual Information Loss

In our experiments, we observed that the increased dimensionality of $\boldsymbol{z}_{id}$ with respect to the discrete labels $\boldsymbol{y}$ might cause some or all of the non-identity-related attributes to be encoded by $\boldsymbol{z}_{id}$ instead of $\boldsymbol{z}_{nid}$. For this reason, we force $\boldsymbol{z}_{nid}$ to encode meaningful non-identity-related attributes by using a mutual information loss, as proposed in Chen et al. (2016). As mentioned in Section 2, this can be achieved through an auxiliary network D_{mi} in the discriminator D that is trained to predict $\boldsymbol{c} \subset \boldsymbol{z}_{nid}$. Since we do not have any prior knowledge about the latent variables $\boldsymbol{c}$, we treat them as continuous variables and train D_{mi} as a regressor using minimum squared error (MSE):

$$L_{mi} = \mathbb{E}_{\boldsymbol{z}_{nid}\sim p_{\boldsymbol{z}_{nid}}, \boldsymbol{y}\sim p_{\boldsymbol{y}}}[\|\boldsymbol{c} - D_{mi}(G(\boldsymbol{z}_{nid} \mid E(\boldsymbol{y})))\|_2^2] \quad (6)$$

Balancing the mutual information loss from (6) and the cross-entropy loss from (4) is key to disentangling identity-related attributes from non-identity-related attributes.

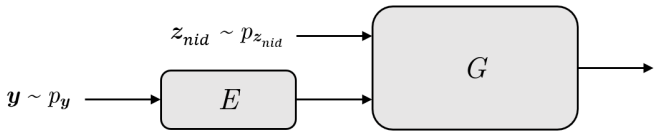
3.4. Proposed GAN

Fig. 1 shows a diagram of our proposed GAN during training. It should be noted that, in practice, the auxiliary networks D_c and D_{mi} share all layers with the discriminator D. Hence, the last layer of D is split into three components that are trained with different loss functions (adversarial loss L_{img} for the real/generated classifier, cross-entropy loss L_c for the identity classifier and MSE loss L_{mi} for the non-identity-related attributes regressor). The architectures of the generator G and the discriminator D are the same as those in PGGAN (Karras et al. 2018). The embedding network E contains an embedding layer that maps the discrete identity labels $\boldsymbol{y}$ to real-valued vectors, followed by two fully-connected layers. The discriminator network $D_{\boldsymbol{z}_{id}}$ contains three fully-connected layers. Both the dimensionality of the latent variables $\boldsymbol{z}_{id}$ encoding the identity-related attributes and the dimensionality of the latent variables $\boldsymbol{z}_{nid}$ encoding the non-identity-related attributes are fixed to 64. In our experiments, we did not notice any major difference between making $\boldsymbol{c}$ a subset of $\boldsymbol{z}_{nid}$ and simply making $\boldsymbol{c}$ equal to $\boldsymbol{z}_{nid}$. For simplicity, we adopted the latter option.

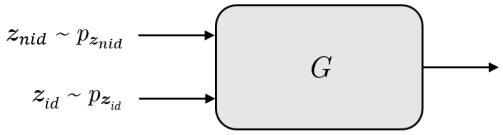
The proposed GAN is trained with the following overall loss:

$$L = L_{img} + \alpha L_c + \beta L_e + \gamma L_{mi} \tag{7}$$

where α, β and γ are weights controlling the contributions of L_c, L_e and L_{mi} to the loss relative to the contribution of L_{img}. Note that (7) is the loss used when training the discriminators D and $D_{\boldsymbol{z}_{id}}$. The generator G and the embedding network E are trained with the same loss as (7) except that the adversarial losses L_{img} and L_e have a negative sign. After extensive experimentation, we set $\alpha = 1$, $\beta = 1$ and $\gamma = 50$. These weights are highly dependent on our specific architecture and should be tuned as necessary for different architectures. In our experiments using the PGGAN generator and discriminator we mainly needed



(a)



(b)

Figure 2: (a) Generation of images of subjects $\boldsymbol{y}$ in the training set with identity-related attributes $E(\boldsymbol{y})$ and random non-identity-related attributes $\boldsymbol{z}_{nid}$. (b) Generation of images of new subjects with random identity-related-attributes $\boldsymbol{z}_{id}$ and random non-identity-related attributes $\boldsymbol{z}_{nid}$.

to carefully tune the γ parameter associated with the mutual information loss L_{mi}. For example, if we increased the dimensionality of the latent variables $\boldsymbol{z}_{id}$ encoding the identity-related attributes, we risk encoding some non-identity-related attributes in that space. Therefore, we would need to counteract this effect by adjusting γ to increase the contribution of the mutual information loss L_{mi}.

3.5. Data Augmentation Approach

Once the model is trained, we can generate multiple images of the same subject by feeding the generator with a fixed vector of identity-related attributes and different vectors of non-identity-related attributes. Our model allows generation of images of subjects in the training set by feeding the generator with the latent representation of their labels $E(\boldsymbol{y})$ obtained by mapping $\boldsymbol{y}$ through E, as shown in Fig. 2a. Since $E(\boldsymbol{y})$ is trained to follow a Gaussian distribution $p_{\boldsymbol{z}_{id}}$, we can also feed the generator with a random sample $\boldsymbol{z}_{id}$ to generate images of new subjects, as shown in Fig. 2b.

Through this approach, we can increase the depth of the face dataset available to train a face recognition model by generating images of subjects existing in the training set and its width by generating images of new subjects. In Section 4.2 we present our experiments comparing a face recognition model trained with the original face dataset with a face dataset augmented using this approach.

4. Experiments

In this section, we start by providing a qualitative analysis of the synthetic images generated by our proposed GAN. Next, we explore the feasibility of augmenting face datasets with synthetic images, both in terms of width and depth. The augmented datasets are used to train CNN-based face recognition models (henceforth referred to as discriminative models) to determine whether they achieve a higher accuracy than models trained with real images alone.

We use the Face-Resnet architecture as our discriminative model, a popular CNN architecture based on residual blocks that has been used in other face recognition works (Ranjan et al. 2017; Hasnat et al. 2017; Wu et al. 2017). This architecture is composed of 27 convolutional layers, 4 pooling layers and 1 final fully-connected layer. Convolutional layers use 3×3 kernels and are followed by PReLU activation functions. See Ranjan et al. (2017) or Hasnat et al. (2017) for more details. The network is trained with softmax loss and optimised using stochastic gradient descent with momentum. The initial learning rate is set to 0.01 and decreased during training whenever the accuracy on the validation set stops improving. The input to the network are 100×100 RGB images. When training with datasets augmented with synthetic images we make sure that on each training batch the number of real and synthetic images is roughly the same.

We use three different subsets of the curated version of the VGGFace dataset (Parkhi et al. 2015) to train the discriminative models. The number of subjects and images of each subset is specified in Table 1. We chose this dataset because it contains a good number of images per subject (on average, close to 300 in each subset), which helps the training of our proposed GAN.

4.1. Qualitative Analysis of Generated Images

We first train our proposed GAN using the VGGFace$^{\text{large}}$ dataset. Fig. 3 shows synthetic images of subjects in the training set generated by our trained model using the method shown in Fig. 2a. The identity-related attributes $E(\boldsymbol{y})$ have been fixed for each row and the non-identity-related attributes $\boldsymbol{z}_{nid}$ have been fixed for each column. The highlighted images in the first column are real images from the training set. Note how many of the synthetic images are as photo-realistic as the real images shown in the first column of Fig. 3. Fig. 4 shows synthetic images of new subjects generated by our trained model using the method shown in Fig. 2b. As in Fig. 3, the identity-related attributes $\boldsymbol{z}_{id}$ have been fixed for each row and the non-identity-related attributes $\boldsymbol{z}_{nid}$ have been fixed for each column. Note how in both Figs. 3 and 4, the images in each row appear to belong to the same subject since the identity-related attributes have been fixed. In contrast, the images in each column display common attributes that do not affect the identity of the subjects (e.g. head pose, facial expression and background) since the non-identity-related attributes have been fixed. From this we can conclude that our proposed GAN is able to effectively disentangle identity-related attributes from non-identity-related attributes.

To test whether our method can generate images of subjects not present in the training set, we need to make sure that the synthetic images of new subjects indeed display identities that do not exist in the training set. Figs. 5a to 5c show a comparison between synthetic images of three new subjects (shown in the top row of each of Figs. 5a to 5c) and synthetic images of their most similar

Dataset	Number of subjects	Number of images
VGGFace$^{\text{large}}$	2558	734,665
VGGFace$^{\text{medium}}$	800	227,466
VGGFace$^{\text{small}}$	200	58,952

Table 1: VGGFace Subsets Used for Training our Models

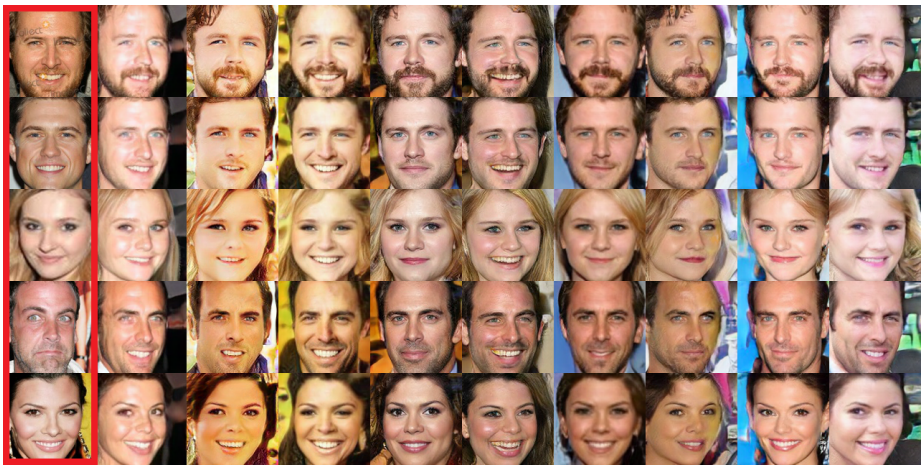


Figure 3: Synthetic images of subjects in the training set generated by our proposed GAN using the method shown in Fig. 2a. The identity related attributes have been fixed for each row and the non-identity related attributes have been fixed for each column. Note that the highlighted images in the first column are real images from the training set.

subject in the training set (shown in the bottom row of each of Figs. 5a to 5c). These figures were created by measuring the average image difference between synthetic images of each new subject in Figs. 5a to 5c and synthetic images of each subject in the training set. Since we were only interested in comparing the identity of the subjects, we averaged over image differences between synthetic images generated with the same non-identity-related attributes $\boldsymbol{z}_{nid}$, as observed in the columns of each pair of rows in Figs. 5a to 5c. We can see how even though the synthetic images of new subjects look similar to the synthetic images of their most similar subject in the training set, it is possible to visually differentiate them as two different identities. Hence, we can conclude that our proposed GAN is able to successfully generate images of subjects not present in the training set.

We can also show how our model has not overfit the training images by applying linear interpolation between two random vectors of identity-related attributes $\boldsymbol{z}_{id}$ and two random vectors of non-identity-related attributes $\boldsymbol{z}_{nid}$. Fig. 6 shows how the transition between synthetic images generated from inter-

Figure 4: Synthetic images of new subjects generated by our proposed GAN using the method shown in Fig. 2b. The identity related attributes have been fixed for each row and the non-identity related attributes have been fixed for each column.

polated vectors is smooth and visually consistent with what would be expected from mixing attributes of two face images. This suggests that our model is able to generate images with enough diversity and it is not just learning to replicate the training images.

It is worth noting that the quality of our images is lower than those presented in the PGGAN work (Karras et al. 2018). We attribute this to the lower quality and resolution of the VGG dataset with respect to the CelebA-HQ dataset used in Karras et al. 2018. Moreover, more recent GAN methods like Brock et al. (2018); Karras et al. (2019) (published after we completed our work) might be able to generate higher quality images than the PGGAN-based approach considered in this work.

4.2. Augmenting Datasets with Synthetic Images

In order to evaluate the quality of datasets augmented with synthetic images, we train several discriminative models with different combinations of real and synthetic images and evaluate them against models trained with real images alone. Each augmented dataset is created by adding synthetic images to one of

(a)

(b)

(c)

Figure 5: Comparison between synthetic images of new subjects and synthetic images of their most similar subject in the training set. The top row of each of (a), (b), (c) contains synthetic images of a new subject and the bottom row of each of (a), (b), (c) contains synthetic images of their most similar subject in the training set. Note that the non-identity-related attributes only vary across the rows of (a), (b), (c) to restrict the comparison to the identity of the subjects.

the VGGFace subsets shown in Table 1. The synthetic images are generated using our proposed GAN trained with the same dataset that we want to augment. For example, if we want to augment VGGFace$^{\text{small}}$, we add synthetic images generated by our proposed GAN trained with VGGFace$^{\text{small}}$. In this way, we can realistically assess whether we can improve the performance of a discriminative model by augmenting its training set using our proposed method. All our discriminative models are evaluated using the IJB-A dataset (Klare et al. 2015). In particular, we use the verification protocol described in Klare et al. (2015)

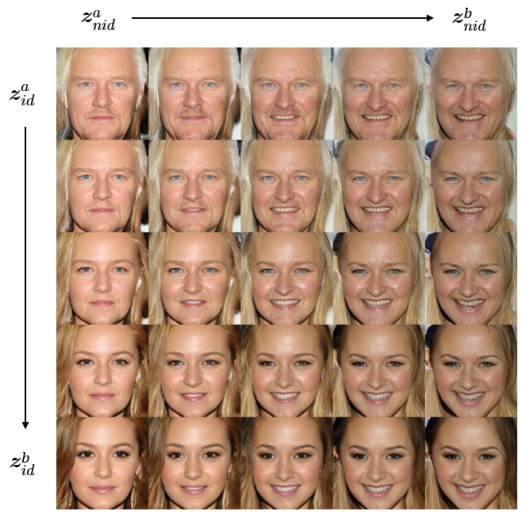


Figure 6: Synthetic images generated by interpolating between two random vectors of identity related attributes $\boldsymbol{z}^a_{id}$, $\boldsymbol{z}^b_{id}$ and two random vectors of non-identity related attributes $\boldsymbol{z}^a_{nid}$, $\boldsymbol{z}^b_{nid}$.

and report the true acceptance rate when the false acceptance rate is fixed to 0.01. We chose the IJB-A dataset for evaluation because it contains challenging images that do not overlap with any of the subjects in the VGGFace dataset.

In Table 2 we show the accuracy of models trained with depth-augmented datasets, i.e. datasets augmented by increasing the number of images per subject with synthetic images. For each subject in the training set we generate multiple synthetic images by fixing the vector of identity-related attributes $E(\boldsymbol{y})$ and randomly sampling different vectors of non-identity-related attributes $\boldsymbol{z}_{nid}$. We augment each VGGFace subset with 250, 500 and 1,500 images, which roughly correspond to a real-synthetic ratio of 1:1, 1:2 and 1:3 respectively. We can see how, in general, the accuracy of the models trained with depth-augmented datasets increases with respect to the models trained without synthetic images. In particular, we obtained maximum accuracy improvements of +1.44%, +2.22% and +4.51% when adding 500 synthetic images per subject to the VGGFace$^{\text{large}}$, VGGFace$^{\text{medium}}$ and VGGFace$^{\text{small}}$ datasets respectively. These results are consistent with the intuition that adding synthetic images

	Number of synthetic images per subject			
Training set	0	250	500	1000
VGGFace$^{\text{large}}$	67.58%	66.65%	**69.02**%	67.74%
VGGFace$^{\text{medium}}$	50.32%	52.25%	**52.54**%	51.97%
VGGFace$^{\text{small}}$	30.64%	33.30%	**35.15**%	32.95%

Table 2: Accuracy of discriminative models trained with depth-augmented datasets. The reported accuracy corresponds to the TAR@FAR=0.01 obtained when evaluating the models on the IJB-A dataset.

	Number of synthetic subjects			
Training set	0	500	1000	1500
VGGFace$^{\text{large}}$	67.58%	65.76%	66.32%	**68.77**%
VGGFace$^{\text{medium}}$	50.32%	52.15%	**54.55**%	54.32%
VGGFace$^{\text{small}}$	30.64%	38.06%	**39.06**%	38.81%

Table 3: Accuracy of discriminative models trained with width-augmented datasets. The reported accuracy corresponds to the TAR@FAR=0.01 obtained when evaluating the models on the IJB-A dataset.

to smaller datasets should result in greater improvement than adding them to larger datasets which contain more real images. The results shown in Table 2 also suggest that there is an optimal balance between the number of real and synthetic images per subject in a given dataset. Indeed, adding 1,000 synthetic images per subject to the VGGFace$^{\text{large}}$, VGGFace$^{\text{medium}}$ and VGGFace$^{\text{small}}$ datasets resulted in lower accuracy than adding 500 synthetic images per subject since the proportion of real images per subject becomes smaller.

In Table 3 we show the accuracy of models trained with width-augmented datasets, i.e. datasets augmented by increasing the number of subjects with synthetic images of new subjects. For each new subject, we generate 500 synthetic images (since this was the best number of synthetic images per subject obtained

in Table 2) by fixing a randomly sampled vector of identity-related attributes $\boldsymbol{z}_{id}$ and randomly sampling different vectors of non-identity-related attributes $\boldsymbol{z}_{nid}$. Again, we observe improvement in most cases. In particular, we obtained a maximum accuracy improvement of +1.19% when adding 1,500 synthetic subjects to the VGGFace$^{\text{large}}$ dataset; and +4.23% and +8.42% when adding 1,000 synthetic subjects to the VGGFace$^{\text{medium}}$ and VGGFace$^{\text{small}}$ datasets respectively. In this case, we also observe that adding synthetic images to the VGGFace$^{\text{large}}$ dataset does not significantly change the recognition accuracy. This can be explained by the fact that this dataset already contains a large number of real subjects. In contrast, we observe a large improvement when increasing the number of synthetic subjects in the VGGFace$^{\text{medium}}$ and VGGFace$^{\text{small}}$ datasets, as the number of real subjects in neither of these datasets is very large. We also observe that there seems to be an optimal balance between the number of real and synthetic subjects in a given dataset. Indeed, as shown in Table 3, adding 1,500 synthetic subjects to the VGGFace$^{\text{medium}}$ and VGGFace$^{\text{small}}$ does not result in higher accuracy than adding 1,000 synthetic subjects since the proportion of real subjects becomes smaller. Note that in the case of the VGGFace$^{\text{large}}$ dataset, more synthetic subjects can be added since there is still a good balance between real and synthetic subjects. However, as mentioned earlier, this dataset already contains a large number of real subjects. Hence, it is expected that no significant improvement in recognition accuracy will be obtained by adding more synthetic subjects.

Looking at the results shown in Tables 2 and 3, we can conclude that augmenting datasets with synthetic images is mainly beneficial for small and medium datasets. Moreover, the accuracy improvement obtained when training with width-augmented and depth-augmented datasets is relative to the number of subjects and number of images per subject of each augmented dataset. For example, the VGGFace$^{\text{small}}$ dataset contains 200 subjects and an average of 295 images per subject. Thus, it is reasonable that the accuracy is improved by a larger margin when adding synthetic images of new subjects (+8.42%) than when adding synthetic images of existing subjects (+4.51%). Note that

we also tried to combine both approaches by simultaneously augmenting the depth and width of the VGGFace subsets. The results were similar to those obtained by training with the width-augmented datasets. We hypothesise that the improvement in these particular datasets is dominated by the addition of synthetic images of new subjects, given that the number of real images per subject is already quite large.

5. Conclusions

In this paper, we have studied the feasibility of augmenting face datasets with photo-realistic synthetic images. In particular, we have presented a new type of conditional GAN that can generate photo-realistic face images from two latent vectors encoding identity-related attributes and non-identity-related attributes respectively. By fixing the latent vector of identity-related attributes and varying the latent vector of non-identity-related attributes, our proposed GAN can generate images of subjects with fixed identities but different attributes, such as facial expression and head pose. The introduction of an embedding network to map discrete identity labels to a continuous latent space of identities allows us to both generate images of subjects in the training set and generate images of new subjects not in the training set. Our experiments have shown the effectiveness of the disentangled representation and the high visual quality of the generated images. To demonstrate the benefit of augmenting datasets with our method, we have trained several CNN-based face recognition models with different combinations of real and synthetic images. In most cases, the discriminative models trained with a combination of real and synthetic images have outperformed the discriminative models trained with real images alone. According to our experimental results, our method is particularly effective when augmenting datasets with a moderate number of subjects and/or images per subject.

Since our proposed generative model is based on GANs, it is easy to adapt to other applications by simply training the model with images of other kind (e.g. animals, cars, etc.) Moreover, by only adding a few simple modifications

to the standard AC-GAN architecture, our method can be easily extended. For example, in our particular case, the proposed GAN could be extended to control the non-identity-related attributes explicitly by conditioning the GAN on specific attributes. This would allow face datasets to be augmented in a tailored manner, e.g. by adding synthetic images of subjects with sunglasses, facial hair, different ages, etc. We hope that this and other ideas derived from our work will contribute to the development of new data augmentation techniques that can facilitate the development of high-accuracy face recognition systems, and accelerate their adoption by the forensics and security communities.

Acknowledgments

This research did not receive any specific grant from funding agencies in public, commercial, or not-for-profit sectors.

References

Antipov, G., Baccouche, M., & Dugelay, J.-L. (2017). Face aging with conditional generative adversarial networks. In *Proc. IEEE Int. Conf. Image Process.* (pp. 2089–2093).

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proc. Int. Conf. Mach. Learn.* (pp. 214–223).

Banerjee, S., Bernhard, J. S., Scheirer, W. J., Bowyer, K. W., & Flynn, P. J. (2017). SREFI: Synthesis of realistic example face images. In *Proc. IEEE Int. Joint Conf. Biometrics* (pp. 37–45).

Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R., & Chellappa, R. (2017). UMDFaces: An annotated face dataset for training deep networks. In *IEEE Int. Joint Conf. Biometrics* (pp. 464–473).

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell*, *25*, 1063–1074.

Brkic, K., Sikiric, I., Hrkac, T., & Kalafatic, Z. (2017). I know that person: generative full body and face de-identification of people in images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops* (pp. 1319–1328).

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Brock, A., Lim, T., Ritchie, J. M., & Weston, N. (2017). Neural photo editing with introspective adversarial networks. In *Proc. Int. Conf. Learn. Representations* (pp. 1–15).

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.* (pp. 67–74).

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. Ann. Conf. Neural Inform. Process. Syst.* (pp. 2172–2180).

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (pp. 8789–8797).

Ding, H., Sricharan, K., & Chellappa, R. (2018). ExprGAN: Facial expression editing with controllable expression intensity. In *Proc. AAAI Conf. Artif. Intell.* (pp. 6781–6788).

Donahue, C., Balsubramani, A., McAuley, J., & Lipton, Z. C. (2018). Semantically decomposing the latent spaces of generative adversarial networks. In *Proc. Int. Conf. Learn. Representations*. ArXiv preprint, 22 Feb 2018; arXiv:1705.07904v3.

Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks, . ArXiv preprint, 3 Apr2017; arXiv:1701.00160v4.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Proc. Ann. Conf. Neural Inform. Process. Syst* (pp. 2672–2680).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein GANs. In *Proc. Ann. Conf. Neural Inform. Process. Syst.* (pp. 5767–5777).

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proc. Eur. Conf. Comput. Vis.* (pp. 87–102).

Hasnat, A., Bohné, J., Milgram, J., Gentric, S., & Chen, L. (2017). DeepVisage: Making face recognition simple yet with powerful generalization skills. In *Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1682-1691* (pp. 1682–1691).

Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 4295–4304).

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2017). AttGAN: Facial attribute editing: Only change what you want, . ArXiv preprint, 25 Jul 2018; arXiv:1711.10678v3.

Huang, R., Zhang, S., Li, T., & He, R. (2017). Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *Proc. IEEE Int. Conf. Comput. Vis.* (pp. 2439–2448).

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *Proc. Int. Conf. Learn. Representations*. ArXiv preprint, 25 May 2016; arXiv:1511.05644v2, supplemental material: `https://github.com/tkarras/progressive_growing_of_gans`.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4401–4410).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. In *Proc. Int. Conf. Learn. Representations*. ArXiv preprint, 7 Jun 2017; arXiv:1703.09507v3.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., & Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 1931–1939).

Kortylewski, A., Schneider, A., Gerig, T., Egger, B., Morel-Forster, A., & Vetter, T. (2018). Training deep face recognition systems with synthetic data, . ArXiv preprint, 16 Feb 2018; arXiv:1802.05891.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L. et al. (2017). Fader networks: Manipulating images by sliding attributes. In *Proc. Ann. Conf. Neural Inform. Process. Syst.* (pp. 5969–5978).

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *Proc. Int. Conf. Mach. Learn.* (pp. 1558–1566).

Lu, Y., Tai, Y.-W., & Tang, C.-K. (2018). Conditional CycleGAN for attribute guided face image generation. In *Proc. Eur. Conf. Comput. Vis.* (pp. 282–297).

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2016). Adversarial autoencoders. In *Proc. Int. Conf. Learn. Representations*. ArXiv preprint, 25 May 2016; arXiv:1511.05644.

Masi, I., Chang, F.-J., Choi, J., Harel, S., Kim, J., Kim, K., Leksut, J., Rawls, S., Wu, Y., Hassner, T. et al. (2019). Learning pose-aware models for pose-

invariant face recognition in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, *41*, 379–393.

Masi, I., Tran, A. T., Hassner, T., Leksut, J. T., & Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? In *Proc. Eur. Conf. Comput. Vis.* (pp. 579–596).

Meden, B., Emeršič, Ž., Štruc, V., & Peer, P. (2018). k-Same-Net: k-anonymity with generative deep neural networks for face deidentification. *Entropy*, *20*. Issue 1, No. 60, 2018; doi: 10.3390/e20010060.

Meden, B., Mallı, R. C., Fabijan, S., Ekenel, H. K., Štruc, V., & Peer, P. (2017). Face deidentification with generative deep neural networks. *IET Signal Processing*, *11*, 1046–1054.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets, . ArXiv preprint, 6 Nov 2014; arXiv:1411.1784.

Mokhayeri, F., Granger, E., & Bilodeau, G.-A. (2018). Domain-specific face synthesis for video face recognition from a single sample per person. *IEEE Trans. Inf. Forensics Security*, *14*, 757–772.

Nech, A., & Kemelmacher-Shlizerman, I. (2017). Level playing field for million scale face recognition. In *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 3406–3415).

Odena, A. (2016). Semi-supervised learning with generative adversarial networks. In *Proc. Int. Conf. Mach. Learning, Workshop Data-Efficient Mach. Learn.*. ArXiv preprint, 22 Oct 2016; arXiv:1606.01583v2.

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. In *Proc. 34th Int. Conf. Mach. Learn.* (pp. 2642–2651).

Osadchy, M., Wang, Y., Dunkelman, O., Gibson, S., Hernandez-Castro, J., & Solomon, C. (2017). GenFace: Improving cyber security using realistic

synthetic face generation. In *Proc. Int. Conf. Cyber Security Cryptography and Mach. Learn.* (pp. 19–33).

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proc. British Mach. Vis. Conf* (pp. 41.1–41.12).

Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible conditional GANs for image editing. *Proc. Ann. Conf. Neural Inform. Process. Syst., Workshop Adversarial Training*, . ArXiv preprint, 19 Nov 2016, arXiv:1611.06355.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint, 7 Jan 2016; arXiv:1511.06434v2*, .

Ranjan, R., Castillo, C. D., & Chellappa, R. (2017). L2-constrained softmax loss for discriminative face verification, . ArXiv preprint, 7 Jun 2017; arXiv:1703.09507v3.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In *Proc. Ann. Conf. Neural Inform. Process. Syst.* (pp. 2234–2242).

Schmidhuber, J. (2020). Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). *Neural Networks*, *127*, 58–66. DOI: 10.1016/j.neunet.2020.04.008.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 815–823).

Shen, W., & Liu, R. (2017). Learning residual images for face attribute manipulation. In *Proc. IEEE Conf. on Comp. Vis. and Pattern Recognit.* (pp. 1225–1233).

Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., & Samaras, D. (2017). Neural face editing with intrinsic image disentangling. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 5444–5453).

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 1891–1898).

Tran, L. Q., Yin, X., & Liu, X. (2018). Representation learning by rotating your faces. *IEEE Trans. Pattern Anal. Mach. Intell., accepted for publication*, . Early access, 03 September 2018; doi: 10.1109/TPAMI.2018.2868350.

Wu, Y., Liu, H., Li, J., & Fu, Y. (2017). Deep face recognition with center invariant loss. In *Proc. ACM Multimedia Thematic Workshops* (pp. 408–414).

Wu, Y., Yang, F., & Ling, H. (2019). Privacy-protective-GAN for privacy preserving face de-identification. *J. Comput. Sci. and Technol.*, . Vol. 34, issue 1, pp. 47-60, 2019.

Yan, X., Yang, J., Sohn, K., & Lee, H. (2016). Attribute2image: Conditional image generation from visual attributes. In *Proc. Eur. Conf. Comput. Vis.* (pp. 776–791). Springer.

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch, . ArXiv preprint, 28 Nov 2014; arXiv:1411.7923.

Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., & Kim, J. (2015). Rotating your face using multi-task deep neural network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 676–684).

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 4352–4360).

Zhao, J., Xiong, L., Li, J., Xing, J., Yan, S., & Feng, J. (2018). 3D-aided dual-agent GANs for unconstrained face recognition. *IEEE Trans. Pattern*

Anal. Mach. Intell., . Accepted for publication, early access, 23 July 2018; doi: 10.1109/TPAMI.2018.2858819.

Zhou, E., Cao, Z., & Yin, Q. (2015). Naive-deep face recognition: Touching the limit of LFW benchmark or not?, . ArXiv preprint, 20 Jan 2015; arXiv:1501.04690.

Zhou, Y., & Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *Proc. Int. Conf. Affective Comput. and Intell. Interaction* (pp. 370–376).

Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (pp. 787–796).

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2013). Deep learning identity-preserving face space. In *Proc. IEEE Int. Conf. Comput. Vis.* (pp. 113–120).

Zhu, Z., Luo, P., Wang, X., & Tang, X. (2014). Multi-view perceptron: a deep model for learning face identity and view representations. In *Proc. Ann. Conf. Neural Inform. Process. Syst.* (pp. 217–225).