# Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model

Baris Gecer[1], Binod Bhattarai[1], Josef Kittler[2], and Tae-Kyun Kim[1]

[1] Department of Electrical and Electronic Engineering, Imperial College London, UK
{b.gecer,b.bhattarai,tk.kim}@imperial.ac.uk
https://labicvl.github.io/
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, UK
j.kittler@surrey.ac.uk
https://www.surrey.ac.uk/centre-vision-speech-signal-processing

**Abstract.** We propose a novel end-to-end semi-supervised adversarial framework to generate photorealistic face images of new identities with a wide range of expressions, poses, and illuminations conditioned by synthetic images sampled from a 3D morphable model. Previous adversarial style-transfer methods either supervise their networks with a large volume of paired data or train highly under-constrained two-way generative networks in an unsupervised fashion. We propose a semi-supervised adversarial learning framework to constrain the two-way networks by a small number of paired real and synthetic images, along with a large volume of unpaired data. A set-based loss is also proposed to preserve identity coherence of generated images. Qualitative results show that generated face images of new identities contain pose, lighting and expression diversity. They are also highly constrained by the synthetic input images while adding photorealism and retaining identity information. We combine face images generated by the proposed method with a real data set to train face recognition algorithms and evaluate the model quantitatively on two challenging data sets: LFW and IJB-A. The generated images by our framework consistently improve the performance of deep face recognition networks trained with the Oxford VGG Face dataset, and achieve comparable results to the state-of-the-art.

## 1 Introduction

Deep learning has shown a great improvement in performance of several computer vision tasks [41,22,17,18,13,14,66] including face recognition [37,47,63,34,62] in the recent years. This was mainly thanks to the availability of large-scale datasets. Yet the performance is often limited by the volume and the variations of training examples. Larger and wider datasets improve the generalization and overall performance of the model [47,1].

The process of collecting and annotating training examples for every specific computer vision task is laborious and non-trivial. To overcome this challenge, additional synthetic training examples along with limited real training examples can be utilised to train the model. Some of the recent works such as 3D face reconstruction [42], gaze

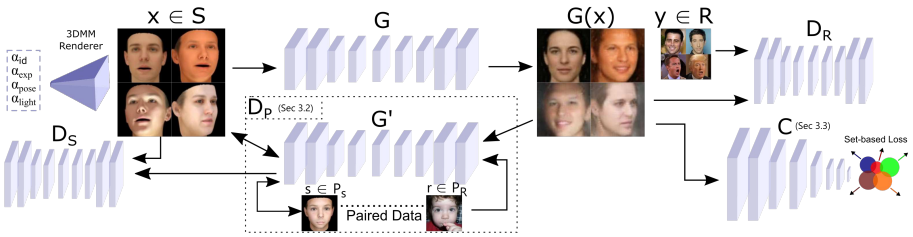



Fig. 1: Our approach aims to synthesize photorealistic images conditioned by a given synthetic image by 3DMM. It regularizes cycle consistency [71] by introducing an additional adversarial game between the two generator networks in an unsupervised fashion. Thus the under-constraint cycle loss is supervised to have correct matching between the two domains by the help of a limited number of paired data. We also encourage the generator to preserve face identity by a set-based supervision through a pretrained classification network.

estimation [69,61], human pose, shape and motion estimation [58] *etc*. use additional synthetic images generated from 3D models to train deep networks. One can generate synthetic face images using a 3D morphable model (3DMM) [3] by manipulating identity, expression, illumination, and pose parameters. However, the resulting images are not photorealistic enough to be suitable for in-the-wild face recognition tasks. It is because the information of real face scans is compressed by the 3DMM and the graphical engine that models illumination and surface is not perfectly accurate. Thus, the main challenge of using synthetic data obtained from 3DMM model is the discrepancy in the nature and quality of synthetic and real images which poses the problem of domain adaptation [38]. Recently, adversarial training methods [48,51,12] have become popular to mitigate such challenges.

Generative Adversarial Network (GAN), introduced by Goodfellow *et al*. [20], and its variants [39,28,2,15] are quite successful in generating realistic images. However, in practice, GANs are likely to stuck in mode collapse for large scale image generation. They are also unable to produce images that are 3D coherent and globally consistent [20]. To overcome these drawbacks, we propose a semi-supervised adversarial learning framework to synthesize photorealistic face images of new identities exhibiting extensive data variation supplied by a 3DMM. We address these shortcomings by exciting a generator network with synthetic images sampled from 3DMM and transforming them into photorealistic domain using adversarial training as a bridge. Unlike most of the existing works that excite their generators with a noise vector [39,2], we feed our generator network by synthetic face images. Such a strong constraint naturally helps in avoiding the mode collapse problem, one of the main challenges faced by the current GAN methods. Fig. 1 shows a general overview of the proposed method. We discuss the proposed method in more details in Sec. 3.

In this paper, we address the challenge of generating photorealistic face images from 3DMM rendered faces of different identities with arbitrary poses, expressions, and illuminations. We formulate this problem as a domain adaptation problem *i.e*. aligning

the 3DMM rendered face domain into realistic face domain. One of the previous works closest to ours [26] addresses the style transfer problem between a pair of domains with classical conditional GAN. The major bottleneck of this method is that it requires a large number of paired examples from both domains which are hard to collect. CycleGAN [71], another recent method and closest to our work, proposes a two-way GAN framework for unsupervised image-to-image translation. However, the cycle consistency loss proposed in their method is satisfied as long as the transitivity of the two mapping networks is maintained. Thus, the resulting mapping is not guaranteed to produce the intended transformation. To overcome the drawbacks of these methods [26,71], we propose to use a small amount of paired data to train an inverse mapping network as a matching aware discriminator. In the proposed method, the inverse mapping network plays the role of both the generator and the discriminator. To the best of our knowledge, this is the first attempt for adversarial semi-supervised style translation for an application with such limited paired data.

Adding realism to the synthetic face images and preserving their identity information is a challenging problem. Although synthetic input images, 3DMM rendered faces, contain distinct face identities, the distinction between them vanishes as a result of the inherent non-linear transformations induced by the discriminator to encourage realism. To tackle such a problem, prior works either employ a separate pre-trained network [65] or embed Identity labels (id) [55] into the discriminator. Unlike existing works, which are focused on generating new images of existing identities, we are interested in generating multiple images of new identities. Therefore, such techniques are not directly applicable to our problem. To address this challenge, we propose to use set-based center [59] and pushing loss functions [19] on top of a pre-trained face embedding network. This will keep track of the changing average of embeddings of generated images belonging to the same identity (i.e. centroids). In this way identity preservation becomes adaptive to the changing feature space during the training of the generator network unlike softmax layer that converges very quickly at the beginning of the training before meaningful images are generated.

Our contributions can be summarized as follows:

- We propose a novel end-to-end adversarial training framework to generate photorealistic face images of new identities constrained by synthetic 3DMM images with identity, pose, illumination and expression diversity. The resulting synthetic face images are visually plausible and can be used to boost face recognition as additional training data or any other graphical purposes.
- We propose a novel semi-supervised adversarial style transfer approach that trains an inverse mapping network as a discriminator with paired synthetic-real images.
- We employ a novel set-based loss function to preserve consistency among unknown identities during GAN training.

## 2 Related Works

In this Section we discuss the prior art that are closely related to the proposed method.

*Domain Adaptation.* As stated in the introduction, our problem of generating photo-realistic face images from 3DMM rendered faces can be seen as a domain adaptation problem. A straightforward adaptation approach is to align the distributions at the feature level by simply adding a loss to measure the mismatch either through second-order moments [52] or with adversarial losses [56,57,16].

Recently, pixel level domain adaptation became popular due to practical breakthroughs on Kullback-Leibler divergence [21,20,39], namely GANs which optimize a generative and discriminative network through a mini-max game. It has been applied to a wide range of problems including fashion clothing [31], person specific avatar creation [60], text-to-image synthesis [67], face frontalization [65], and retinal image synthesis [12].

Pixel domain adaptation can be done in a supervised manner simply by conditioning the discriminator network [26] or directly the output of the generator [9] with the expected output when there is enough paired data from both domains. Please note collecting a large number of paired training examples is expensive, and often requires expert knowledge. [40] proposes a text-to-image synthesis GAN with a matching aware discriminator. They optimize their discriminator for image-text matching beside requiring realism with the information provided by additional mismatched text-image pairs.

For the cases where paired data is not available, many approaches adapt unsupervised learning such as imposing pixel-level consistency between input and output of the generator network [6,48], an encoder architecture that is shared by both domains[7] and adaptive instance normalization [24]. An interesting approach is to have two way translation between domains with two distinct generator and discriminator networks. They constrain the two mappings to be inverses of each other with either ResNet [71] or encoder-decoder network [33] as the generator.

*Synthetic Training Data Generation.* The usage of synthetic data as additional training data is shown to be helpful even if they are graphically rendered images in many applications such as 3D face reconstruction [42], gaze estimation [69,61], human pose, shape and motion estimation [58]. Despite the availability of almost infinite number of synthetic images, those approaches are limited due to the domain difference from that of in-the-wild images.

Many existing works utilize adversarial domain adaptation to translate images into photorealistic domain so that they are more useful as training data. [70] generates many unlabeled samples to improve person re-identification in a semi-supervised fashion. RenderGAN [51] proposes a sophisticated approach to refine graphically rendered synthetic images of tagged bees to be used as training data for a bee tag decoding application. WaterGAN [32] synthesizes realistic underwater images by modeling camera parameters and environment effects explicitly to be used as training data for a color correction task. Some studies deform existing images by a 3D model to augment diverse datasets [36] without adversarial learning.

One of the recent works, simGAN [48], generates realistic synthetic data to improve eye gaze and hand pose estimation. It optimizes the pixel level correspondence between input and output of the generator network to preserve the content of the synthetic image. This is in fact a limited solution since the pixel-consistency loss encourages the generated images to be similar to synthetic input images and it partially contradicts

adversarial realism loss. Instead, we employ an inverse translation network similar to cycleGAN [71] with an additional pair-wise supervision to preserve the initial condition without hurting realism. This network also behaves as a discriminator to a straight mapping network trained with real paired data to avoid possible biased translation.

*Identity Preservation.* To preserve the identity/category of the synthesized images, some of the recent works such as [10,55] keep categorical/identity information in discriminator network as an additional task. Some of the others propose to employ a separate classification network which is usually pre-trained [35,65]. In both these cases, the categories/identities are known beforehand and are fixed in number. Thus it is trivial to include such supervision in a GAN framework by training the classifier with real data. However such setup is not feasible in our case as images of new identities to-be-generated are not available to pre-train a classification network.

To address the limitation of existing methods of retaining identity/category information of synthesized images, we employ a combination of different set-based supervision approaches for unknown identities to be distinct in the pre-trained embedding space. We keep track of moving averages of same-id features by the momentum-like centroid update rule of center loss [59] and penalize distant same-id samples and close different-id samples by a simplified variant of the magnet loss[43] without its sophisticated sampling process and with only a single cluster per identity (see Section 3.3 for further discussions).

## 3 Adversarial Identity Generation

In this Section, we describe in details the proposed method. Fig. 1 shows a schematic diagram of our method. Specifically, the synthetic image set $x \in \mathcal{S}$ is formed by a graphical engine for the randomly sampled of 3DMM with its identity, pose and lighting parameters $\alpha$. The generated images they are translated into a more photorealistic domain $G(x)$ through the network, $G$, and mapped back to its synthetic domain ($G'(G(x))$) through the network, $G'$, to retain $x$. The adversarial synthetic and real domain translation of $G$ and $G'$ networks are supervised by the discriminator networks $D_R$ and $D_S$, with an additional adversarial game between $G$ and $G'$ as a generator and a discriminator respectively. During training, the identities generated by 3DMM are preserved with a set-based loss on a pre-trained embedding network $C$. In the following sub-sections, we further describe these components *i.e.* domain adaptation, real-synthetic pair discriminator, and identity preservation.

### 3.1 Unsupervised Domain Adaptation

Given a 3D morphable model (3DMM) [3], we synthesize face images of new identities sampled from its Principal Components Analysis (PCA) coefficients' space with random variation of expression, lighting and pose. Similar to [71], a synthetic input image ($x \in \mathcal{S}$) is mapped to a photorealistic domain by a residual network ($G : S \rightarrow \hat{R}$) and mapped back to the synthetic domain by a 3DMM fitting network ($G' : \hat{R} \rightarrow \hat{S}$) to

complete the forward cycle only[3]. To preserve cycle consistency, the resulting image $G'(G(x))$ is encouraged to be the same as input $x$ by a pixel level $L_1$ loss:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x \in \mathcal{S}} \|G'(G(x)) - x\|_1 \tag{1}$$

In order to encourage the resulting images $G(x)$ and $G'(G(x))$ to have a similar distribution as real and synthetic domains respectively, those refiner networks are supervised by discriminator networks $D_R$ and $D_S$ with images of the respective domains. The discriminator networks are formed as auto-encoders as in the boundary equilibrium GAN (BEGAN) architecture [2] in which the generator and discriminator networks are trained by the following adversarial training formulation:

$$\mathcal{L}_G = \mathbb{E}_{x \in \mathcal{S}} \|G(x) - D_R(G(x))\|_1 \tag{2}$$

$$\mathcal{L}_{G'} = \mathbb{E}_{x \in \mathcal{S}} \|G'(G(x)) - D_S(G'(G(x)))\|_1 \tag{3}$$

$$\mathcal{L}_{D_R} = \mathbb{E}_{x \in \mathcal{S}, y \in \mathcal{R}} \|y - D_R(y)\|_1 - k_t^{D_R} \mathcal{L}_G \tag{4}$$

$$\mathcal{L}_{D_S} = \mathbb{E}_{x \in \mathcal{S}} \|x - D_S(x)\|_1 - k_t^{D_S} \mathcal{L}_{G'} \tag{5}$$

where for each training step $t$ and the generator network ($G$ for $k_t^{D_R}$, $G'$ for $k_t^{D_S}$) we update the balancing term with $k_t^D = k_{t-1}^D + 0.001(0.5\mathcal{L}_D - \mathcal{L}_G)$. As suggested by [2], this term helps to maintain a balance between the interests of the generator and discriminator and stabilize the training.

### 3.2 Adversarial Pair Matching

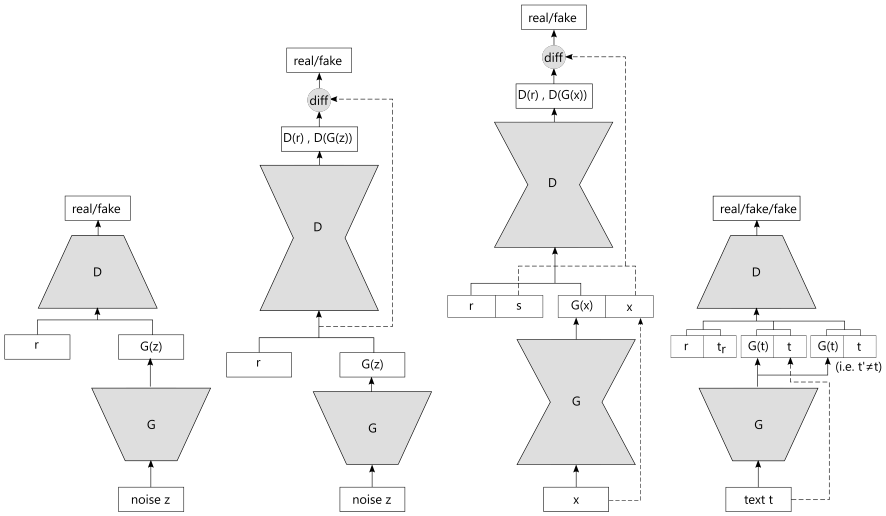
The cycle consistency loss ensures the bijective transitivity of functions $G$ and $G'$ which means generated image $G(x) \in \hat{R}$ should be transformed back to $x \in \hat{S}$. Convolutional networks are highly under-constrained and they are free to make any unintended changes as long as the cycle consistency is satisfied. Therefore, without an additional supervision, it is not guaranteed to achieve the correct mapping that preserves shape, texture, expression, pose and lighting attributes of the face image from domains $S$ to $\hat{R}$ and $\hat{R}$ to $\hat{S}$. This problem is often addressed by introducing pixel-level penalization between input and output of the networks [71,48] which is sub-optimal for domain adaptation as it encourages to stay in the same domain.

To overcome this issue, we propose an additional pair-wise adversarial loss that assigns the $G'$ network an additional role as a pair-wise discriminator to supervise the $G$ network. Given a set of paired synthetic and real images $(\mathcal{P}_\mathcal{S}, \mathcal{P}_\mathcal{R})$, the discriminator loss is computed by BEGAN as follows:

$$\mathcal{L}_{D_P} = \mathbb{E}_{s \in \mathcal{P}_\mathcal{S}, r \in \mathcal{P}_\mathcal{R}} \|s - G'(r)\|_1 - k_t^{D_P} \mathcal{L}_{cyc} \tag{6}$$

While the $G'$ network is itself a generator network ($G' : \hat{R} \rightarrow \hat{S}$) with a separate discriminator ($D_S$), we use it as a third pair-matching discriminator to supervise $G$ by means of a distribution of paired correspondence of real and synthetic images. Thus

[3] We empirically found that removing the backward cycle-loss improves performance when the task is to map from artificial images to real as also shown in Tab.4 of [71]




(a) DC-GAN[39] (b) BEGAN [2] (c) Ours (d) GAN-CLS [40]

Fig. 2: Comparison of our pair matching method to the related work. (a) In the traditional GAN approach, the discriminator module aligns the distribution of real and synthetic images by means of a classification network. (b) BEGAN[2] and many others showed that the alignment of reconstruction error distributions offers a more stable training. (c) We propose to utilize this autoencoder approach to align the distribution of pairs to encourage generated images to be transformed to the realistic domain with a game between real and synthetic pairs. (d) An alternative to our method is to introduce wrongly labeled images to the discriminator to teach pair-wise matching as proposed by [40] for text to images synthesis.

while cycle-loss optimizes for the biject correspondences, we expect the resulting pairs of $(x \in S, G(x) \in \hat{R})$ to have the similar correlation distribution as paired training data $(s \in \mathcal{P}_{\mathcal{S}}, r \in \mathcal{P}_{\mathcal{R}})$. Fig 2 shows its relation to the previous related art and comparison to an alternative which is a matching aware discriminator with paired inputs for text to image synthesis, as suggested by [40]. Please note that how BEGAN autoencoder architecture is utilized to align the distribution of pairs of synthetic and real images with synthetic and generated images.

Alternatively, one could pretrain the $G'$ network as a 3DMM fitting network as in [54,49,53,11] . However, we trained it from scratch to balance the adversarial zero-sum game between the generator ($G$) and the pair-wise discriminator ($G'$). Otherwise the gradient would vanish as there would be no success in fooling the discriminator. Moreover, those networks provide only fitted 3DMM parameters which then would need to be rendered into 3DMM images by a differentiable tensor operation.

Fig. 3: Quality of 9 images of 3 identities (per row) during the training. Red plot shows the proposed identity preservation loss over the iterations. Note the changes of fine-details on the faces which is the main motivation of set-based identity preservation.

### 3.3 Identity Preservation

Although identity information is provided by the 3DMM in shape and texture parameters, it may be lost to some extent by virtue of a non-linear transformation. Some studies [65,55] address this issue by employing identity labels of known subjects as additional supervision either with a pre-trained classification network or within the discriminator network. However, we intend to generate images of new identities sampled from the 3DMM parameter space and their photorealistic images simply do not exist yet. Furthermore, training a new softmax layer and the rest of the framework simultaneously becomes a chicken-egg problem and results in failed training.

In order to preserve identity in the changing image space, we propose to adapt a set-based approach over a pre-trained face embedding network. We import the idea of pulling same-id samples as well as pushing close samples from different identities in the embedding space such that same-id images are gathered and distinct from other identities regardless of the quality of the images during the training. At the embedding layer of a pre-trained network $C$, the generator network ($G$) is supervised by a combination of the center [59] and pushing loss [19] (which is also a simplified version of the Magnet loss [43] formulation) defined for a given mini-batch (M) as:

$$\mathcal{L}_C = \mathbb{E}_{x\in\mathcal{S}, i_x\in\mathbb{N}^+} \sum_{x}^{M} -log \frac{\exp(\frac{1}{2\sigma^2}\|C(G(x)) - c_{i_x}\|_2^2 - \eta)}{\sum_{j\neq i_x}\exp(\frac{1}{2\sigma^2}\|C(G(x)) - c_j\|_2^2)} \tag{7}$$

where $i_x$ stands for the identity label of $x$ provided by 3DMM sampling and $c_j$ stands for the mean embedding of identity $j$. The margin term, $\eta$, is set to 1 and the variance is computed by $\sigma = \frac{\sum_x^M \|C(G(x)) - c_{i_x}\|_2^2}{M-1}$.

While the quality of images is improved during the training, *i.e.* better photo-realism, their projection on the embedding space is shifting. In order to adapt to those changes, we update identity centroids ($c_j$) with a momentum of $\beta = 0.95$ when new images of id $j$ become available. Following [59], for a given $x$, the moving average of an identity centroid is calculated by $c_j^{t+1} = c_j^t - \beta\delta(i_x = j)(c_j^t - C(G(x)))$ where $\delta(condition) = 1$, if the condition is satisfied and $\delta(condition) = 0$ if not. Centroids ($c_j$) are initialized with zero and after few iterations, they converge to embedding centers and then continue updating to adapt to the changes caused by the simultaneous training of $G$. Fig. 3 shows the quality of 9 images of 3 identities over training iterations. Please note the difference of the images after convergence with the images at the

beginning of the training, produced by the Softmax layer which fails to supervise the forthcoming images in later iterations.

**Full Objective**

Overall, the framework is optimized by the following updates simultaneously:

$$\theta_G = \arg\min_{\theta_G} \mathcal{L}_G + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_C\mathcal{L}_C \tag{8}$$

$$\theta_{G'} = \arg\min_{\theta_{G'}} \mathcal{L}_{G'} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{D_P}\mathcal{L}_{D_P} \tag{9}$$

$$\theta_{D_R}, \theta_{D_S} = \arg\min_{\theta_{D_R}, \theta_{D_S}} \mathcal{L}_{D_R} + \mathcal{L}_{D_S} \tag{10}$$

where $\lambda$ parameters balance the contribution of different modules. The selection of those parameters is discussed in the next section.

## 4 Implementation Details

*Network Architecture:* For the generator networks ($G$ and $G'$), we use a shallow ResNet architecture as in [27] which supplies smooth transition without changing the global structure because of its limited capacity, having only 3 residual blocks. In order to benefit from 3DMM images fully, we also add skip connections to the network $G$. We also add dropout layers after each block in the forward pass with a 0.9 keep rate to introduce some noise that could be caused by uncontrolled environmental changes.

We construct the discriminator networks ($D_R$ and $D_S$) as autoencoders trained by boundary equilibrium adversarial learning with Wasserstein distance as proposed by [2]. The classification network $C$, is a shallow FaceNet architecture [47], more specifically we use NN4 network with an input size of $96 \times 96$ where we randomly crop, rotate and flip generated images $G(x)$ which are of size $108 \times 108$.

*Data:* Our framework needs a large amount of real and synthetic face images. For real face images, we use CASIA-Web Face Dataset [64] that consists of ∼500K face images of ∼10K individuals.

The proposed method trains the $G'$ network as a discriminator ($D_P$) with a small number of paired real and synthetic images. For that, we use a combination of 300W-3D [46,45,4] and AFLW2000-3D datasets as our paired training set [72] which consist of 5K real images with their corresponding 3DMM parameter annotations. We render synthetic images by those parameters and pair them with the matching real images. This dataset is relatively small, compared to the ones used by fully supervised transformation GANs (i.e. Amazon Handbag dataset used by [26] contains 137K bag images).

We randomly sample face images of new identities as our synthetic data set using Large Scale Face Model (LSFM) [5] for shape, Basel Face Model [25] for texture and Face Warehouse model [8] for expression. While the shape and texture parameters of new identities are sampled from the Gaussian distribution of the original model, expression, lighting and pose parameters are sampled with the same Gaussian distribution as that of synthetic samples of 300-3D [46,45,4] and AFLW2000-3D [72]. All images are aligned by MTCNN [68] and centre cropped to the size of $108 \times 108$ pixels.

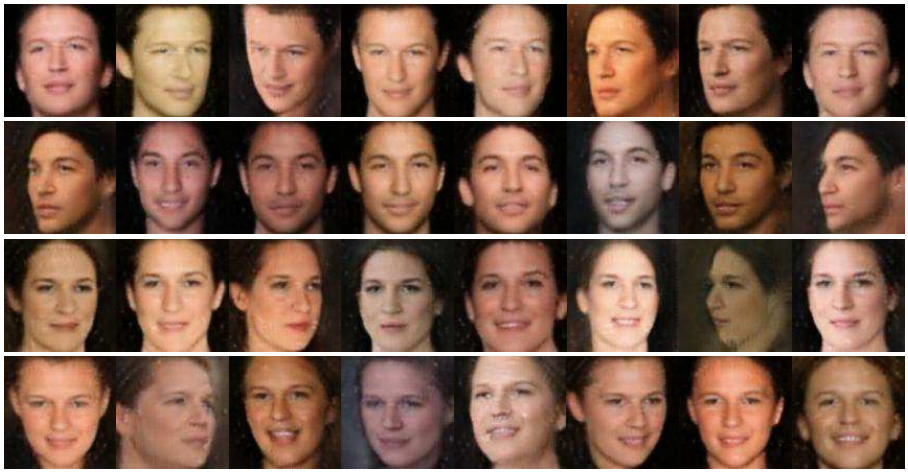


Fig. 4: Random samples from GANFaces dataset. Each row belongs to the same identity. Notice the variation in pose, expression and lighting.

*Training Details:* We train all the components of our framework together from scratch except for the classification network $C$ which is pre-trained by using a subset of Oxford VGG Face Dataset [37]. The whole framework takes about 70 hours to converge on a Nvidia GTX 1080TI GPU in 248K iterations with a batch size of 16. We start with a learning rate of $8 \times 10^{-5}$ with ADAM solver [29] and halve it after 128Kth, 192Kth, 224Kth, 240Kth, 244Kth, 246Kth and 247Kth iterations.

As shown in Eqn. 8, 9, $\lambda$ is a balancing factor which controls the contribution of each optimization. We set $\lambda_{cyc} = 0.5$, $\lambda_{D_P} = 0.5$, $\lambda_C = 0.001$ to achieve a balance between realism, cycle-consistency, identity preservation and the supervision by the paired data. We also add identity loss ($\mathcal{L}_{id} = \|x - G(x)\|$) as suggested by [71] to regularize the training with a balancing term $\lambda_{id} = 0.1$. During the training, we keep track of moving averages of the network parameters to generate images.

## 5 Results and Discussions

In this section, we show the qualitative and quantitative results obtained with the proposed framework. We also discuss and show the contribution of each module (i.e. $\mathcal{L}_{cyc}$, $D_P$, $C$) with an ablation study in the supplementary materials. For the experiments, we generate 500,000 and 5,000,000 images of 10,000 different identities with variations on expression, lighting and poses. We name this synthetic dataset **GANFaces**[4] (*i.e.* GANFaces-500K, GANFaces-5M).

[4] The dataset, training code, pre-trained models and face recognition experiments can be viewed at https://github.com/barisgecer/facegan

(a) 3DMM inputs (b) simGAN [[6] (c) CycleGAN [[71] (d) Reconstruction Err.

Fig. 5: Comparison to (a) input 3DMM images, (b) results with simGAN settings [6],(c) results with cycleGAN settings [71] and (d) results with additional reconstruction loss. Figures correspond to left half of the Fig. 4 and each row belongs to the same identity.

### 5.1 Qualitative Evaluation

Please see Fig. 4 for random samples from the dataset. Fig. 5 compares our results (left half of the Fig. 4) with the 3DMM inputs, the results with simGAN [6] and cycleGAN [71] settings, and our setup with the addition of the reconstruction loss of the paired data within the $G$ network. We observe good correspondence when we compare first 4 columns of Fig.4 to Fig. 5(a) in terms of identity, pose, expression and lighting. Compared to ours (Fig. 4), [6] suffers from the loss of identity-specific facial features (Fig. 5(b)) while [71] generates images visually less pleasant (Fig. 5(c)). An additional reconstruction loss used in our framework to train the $G$ network with the paired data produces the results in Fig. 5(d). We achieved less clear images by this step probably because of the severity of the influence of the direct reconstruction loss on the adversarial balance. The superiority of the proposed framework is also confirmed by the quantitative experiments shown in Table. 1.

One of the main goals of this work is to generate face images guided by the attributes of synthetic input images *i.e.* shape, expression, lighting, and poses. We can see from Fig. 6 that our model is capable of generating photorealistic images preserving the attributes conditioned by the synthetic input images. In the figure, top row shows the variations of pose and expression on input synthetic faces and the left column shows the input synthetic faces of different identities. The rest are the images generated by our model, conditioned on the corresponding attributes from the top row and the left column. We can clearly see that the conditioning attributes are preserved on the images generated by our model. We can also observe that fine-grained attributes such as shapes of chin, nose and eyes are also retained in the images generated by our model. In the case of extreme poses, the quality of the image generated by our model becomes less sharp as the CASIA-WebFace dataset, which we used to learn the parameters of discriminator network $D_R$, lacks a sufficient number of examples with extreme poses.

### 5.2 The Added Realism and Identity Preservation

In order to show that synthetic images are effectively transformed to the realistic domain with preserving identities, we perform a face verification experiments on GANFaces dataset. We took pre-trained face-recognition CNN network, namely FaceNet

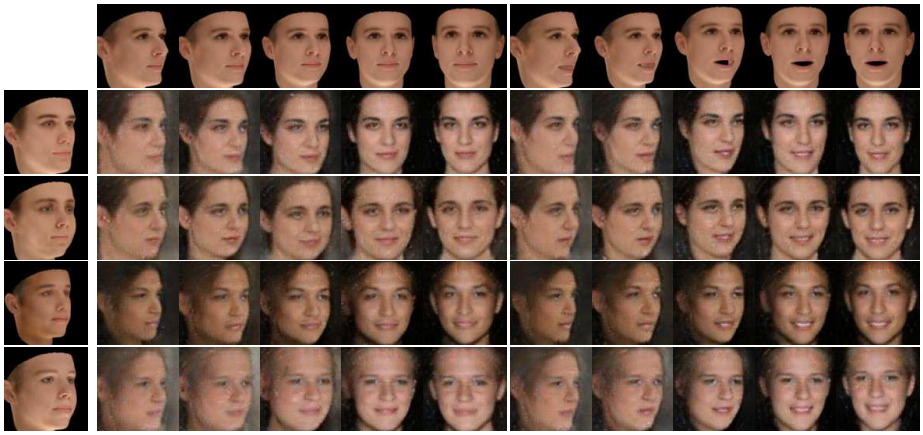


Fig. 6: Images generated by the proposed approach conditioned by identity variation in the vertical axis, normalized and mouth open expression in left and right blocks and pose variation in the horizontal axis. Images in this figure are not included in the training

NN4 architecture [47] trained on CASIA-WebFace [64] to compute the features of the face images. The verification performance of the network on LFW is $\%95.6$ accuracy and $\%95.5$ 1-EER which shows that the model is well optimized for in-the-wild face verification. We created 1000 similar (belonging to same identity) and 1000 dis-similar (belonging to different identities) face image pairs from GANFaces. Similarly, we also generated the same number of similar and dis-similar face image pairs from the VGG face dataset [37] and the synthetic 3DMM rendered faces dataset. Fig. 7 shows the histograms of euclidean distances between similar and dis-similar images measured in the embedding space for the three datasets. The addition of realism and preservation of identities of the GANFaces can be seen from the comparison of its distribution to the 3DMM synthetic dataset distribution. As the images become more realistic, they become better separable in the pre-trained embedding space. We also observe that the separation of positive and negative pairs of GANFaces is better than that of VGG faces pairs. The probable reason for VGG not achieving a better separation than GANFaces is noisy face labels as indicated in the original study [37].

### 5.3 Face Recognition with GANFaces dataset

We augmented GANFaces with real face dataset *i.e*. VGG Faces [37] and trained the VGG19 [50] network and tested its performance on two challenging datasets: LFW [23] and IJB-A [30]. We restrict ourselves from limited access to full access of real face dataset and train deep network on different combination of real and GANFaces. Following [36], we use a pre-trained VGGNet by [50] with 19 layers trained on the ImageNet dataset [44] and took these parameters as initial parameters. We train the network with different portions of the Oxford VGG Face dataset [37], augmented with the GANFaces dataset. We remove the last layer of the deep VGGNet and add two soft-max layers to

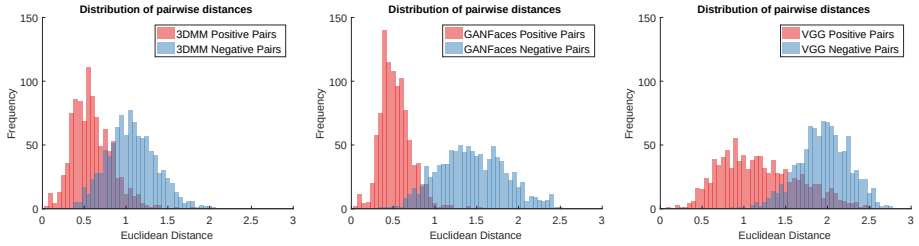



Fig. 7: Distances of 1000 positive and 1000 negative pairs from three different datasets (GANFaces, 3DMM synthetic images, Oxford VGG) embedded on a NN4 network that is trained with CASIA Face dataset

the previous layer, one for each of the datasets. The learning rate is set to 0.1 for the soft-max layers and 0.01 to the pre-trained layers with the ADAM optimizer. Also we halve the gradient coming from the GANFaces soft-max. We decrease the learning rate exponentially and train for 80,000 iterations where all of our models converge well without overfitting. For a given input size of $108 \times 108$, we randomly crop and flip $96 \times 96$ patches and the overall training takes around 9 hours on a GTX 1080TI GPU.

We train 6 models with $\%20$, $\%50$ and $\%100$ of the VGG Face dataset with and without the augmentation of GANFaces-500K. As seen in Fig. 8, we evaluate the models on LFW and IJB-A datasets and the benchmark scores are improved with the addition of this dataset even though the image resolution is low. The contribution of GANFaces-500K increases inversely proportional to the number of images included from the VGG dataset, which indicates more synthetic images might improve the results even further.

We compare our best model trained by full VGG dataset and GANFaces to the other state of the art methods in Table 1. Despite the lower resolution, GANFaces was able to improve our baseline to the numbers comparable to the state-of-the-art. Note that generative methods, such as [36,65], do generation (i.e. pose augmentation and

| Method | Real | Synth | Test time Synth | Image size | Acc. (%) | 100% - EER |
|---|---|---|---|---|---|---|
| FaceNet [47] | 200M | - | No | 220×220 | 98.87 | - |
| VGG Face [37] | 2.6M | - | No | 224×224 | 98.95 | 99.13 |
| Masi *et al*. [36] | 495K | 2.4M | Yes | 224×224 | 98.06 | 98.00 |
| Yin *et al*. [65] | 495K | 495K | Yes | 100×100 | 96.42 | - |
| VGG + Recons. Err. | 1.8M | 500K | No | 96×96 | 94.7 | 94.8 |
| VGG + simGAN [48] | 1.8M | 500K | No | 96×96 | 94.7 | 94.8 |
| VGG + cycleGAN [71] | 1.8M | 500K | No | 96×96 | 94.5 | 94.7 |
| VGG($\%100$) | 1.8M | - | No | 96×96 | 94.8 | 94.6 |
| VGG($\%100$) + GANFaces-500K | 1.8M | 500K | No | 96×96 | **94.9** | **95.1** |
| VGG($\%100$) + GANFaces-5M | 1.8M | 5M | No | 96×96 | **95.2** | **95.1** |

Table 1: Comparison with state-of-the-art studies on LFW performances

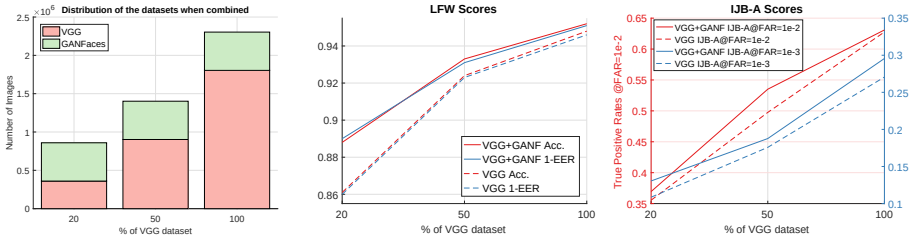



Fig. 8: Face recognition benchmark experiments. (Left) Number of images used from the two datasets in the experiments. The total number of images in the VGG data set is 1.8M since some images were removed from the URL (Middle) Performance on the LFW dataset with (solid) and without (dashed) the GANFaces-500K dataset. (Right) True Positive Rates on the IJB-A verification task with (solid) and without (dashed) the GANFaces-500K dataset.

normalization) in the test time whereas we use only given test images. Together with the benefit of low resolution, this makes our models more efficient at test time.

## 6 Conclusions

This paper proposes a novel end-to-end semi-supervised adversarial training framework to generate photorealistic faces of new identities with wide ranges of poses, expressions, and illuminations from 3DMM rendered faces. Our extensive qualitative and quantitative experiments show that the generated images are realistic and identity preserving.

We generated a synthetic dataset of face images closer to a photorealistic domain and combined it with a real face image dataset to train a face recognition CNN and improved the performance in recognition and verification tasks. In the future, we plan to generate millions of high resolution images of thousands of new identities to boost the state-of-the-art face recognition.

The proposed framework helps to avoid some of the common GAN problems such as mode collapse and 3D coherency. It shows how the data generated by 3DMM or any other explicit model can be utilized to improve and control the behaviour of GANs.

## Acknowledgements

This work was supported by the EPSRC Programme Grant ‘FACER2VM’ (EP/N007743/1). We would like to thank Microsoft Research for their support with Microsoft Azure Research Award. Baris Gecer is funded by the Turkish Ministry of National Education. This study is morally motivated to improve face recognition to help prediction of genetic disorders visible on human face in earlier stages.

## References


1. A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa. The do's and don'ts for cnn-based face verification. *ICCVW*, 2017. 1
2. D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 2, 6, 7, 9
3. V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2, 5
4. J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models" in-the-wild". *CVPR*, 2017. 9, 10
5. J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 9
6. K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CVPR*, 2017. 4, 11
7. K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *NIPS*, 2016. 4
8. C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 9
9. Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *ICCV*, 2017. 4
10. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 5
11. F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *CVPR*, 2017. 7
12. P. Costa, A. Galdran, M. I. Meyer, M. D. Abràmoff, M. Niemeijer, A. M. Mendonça, and A. Campilho. Towards adversarial retinal image synthesis. *arXiv preprint arXiv:1701.08974*, 2017. 2, 4
13. C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016. 1
14. A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 1
15. V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 2
16. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 4
17. B. Geçer. *Detection and classification of breast cancer in whole slide histopathology images using deep convolutional networks*. PhD thesis, Bilkent University, 2016. 1
18. B. Gecer, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recognition*, 2018. 1
19. B. Gecer, V. Balntas, and T.-K. Kim. Learning deep convolutional embeddings for face representation using joint sample-and set-based supervision. In *ICCVW*, 2017. 3, 8
20. I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *NIPS*, 2016. 2, 4
21. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 4
22. K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1

23. G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 13
24. X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *ICCV*, 2017. 4
25. IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. 9
26. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3, 4, 9
27. J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016. 9
28. T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018. 2
29. D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 10
30. B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015. 13
31. C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. *ICCV*, 2017. 4
32. J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson. Watergan: unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation Letters*, 3(1):387–394, 2018. 4
33. M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 4
34. W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1
35. Y. Lu, Y.-W. Tai, and C.-K. Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017. 5
36. I. Masi, A. T. Trn, T. Hassner, J. T. Leksut, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016. 4, 13, 14
37. O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 1, 10, 12, 13
38. V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015. 2
39. A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2, 4, 7
40. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 4, 7
41. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1
42. E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *3D Vision (3DV)*, 2016. 1, 4
43. O. Rippel, M. Paluri, P. Dollar, and L. Bourdev. Metric learning with adaptive density discrimination. *arXiv preprint arXiv:1511.05939*, 2015. 5, 8
44. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 13
45. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 9, 10

46. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013. 9, 10
47. F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1, 9, 12, 13
48. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. *CVPR*, 2017. 2, 4, 6, 13
49. Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 7
50. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 13
51. L. Sixt, B. Wild, and T. Landgraf. Rendergan: Generating realistic labeled data. *arXiv preprint arXiv:1611.01331*, 2016. 2, 4
52. B. Sun and K. Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, 2015. 4
53. A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 7
54. A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017. 7
55. L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 3, 5, 8
56. E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 4
57. E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. *CVPR*, 2017. 4
58. G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. *CVPR*, 2017. 2, 4
59. Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016. 3, 5, 8
60. L. Wolf, Y. Taigman, and A. Polyak. Unsupervised creation of parameterized avatars. *ICCV*, 2017. 4
61. E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138. ACM, 2016. 2, 4
62. C. Xiong, L. Liu, X. Zhao, S. Yan, and T.-K. Kim. Convolutional fusion network for face verification in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):517–528, 2016. 1
63. C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 2015. 1
64. D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 9, 12
65. X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. *ICCV*, 2017. 3, 4, 5, 8, 13, 14
66. S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *CVPR*, 2017. 1
67. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 4
68. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 10

69. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015. 2, 4
70. Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *ICCV*, 2017. 4
71. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3, 4, 5, 6, 10, 11, 13
72. X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 9, 10