# Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks

Baris Gecer[1,2(✉)], Alexandros Lattas[1,2], Stylianos Ploumpis[1,2], Jiankang Deng[1,2], Athanasios Papaioannou[1,2], Stylianos Moschoglou[1,2], and Stefanos Zafeiriou[1,2]

[1] Imperial College, London, UK
{b.gecer,a.lattas,s.ploumpis,j.deng16,a.papaioannou11, stylianos.moschoglou15,s.zafeiriou}@imperial.ac.uk, https://ibug.doc.ic.ac.uk/
[2] FaceSoft.io, London, UK

**Abstract.** Generating realistic 3D faces is of high importance for computer graphics and computer vision applications. Generally, research on 3D face generation revolves around linear statistical models of the facial surface. Nevertheless, these models cannot represent faithfully either the facial texture or the normals of the face, which are very crucial for photo-realistic face synthesis. Recently, it was demonstrated that Generative Adversarial Networks (GANs) can be used for generating high-quality textures of faces. Nevertheless, the generation process either omits the geometry and normals, or independent processes are used to produce 3D shape information. In this paper, we present the first methodology that generates high-quality texture, shape, and normals jointly, which can be used for photo-realistic synthesis. To do so, we propose a novel GAN that can generate data from different modalities while exploiting their correlations. Furthermore, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions. The qualitative results shown in this paper are compressed due to size limitations, full-resolution results and the accompanying video can be found in the supplementary documents. The code and models are available at the project page: https://github.com/barisgecer/TBGAN.

**Keywords:** Synthetic 3D Face · Face generation · Generative Adversarial Networks · 3D morphable models · Facial expression generation

## 1 Introduction

Generating 3D faces with high-quality texture, shape, and normals is of paramount importance in computer graphics, movie post-production, computer

**Electronic supplementary material** The online version of this chapter (https://doi.org/10.1007/978-3-030-58526-6_25) contains supplementary material, which is available to authorized users.


A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12374, pp. 415–433, 2020.
https://doi.org/10.1007/978-3-030-58526-6_25

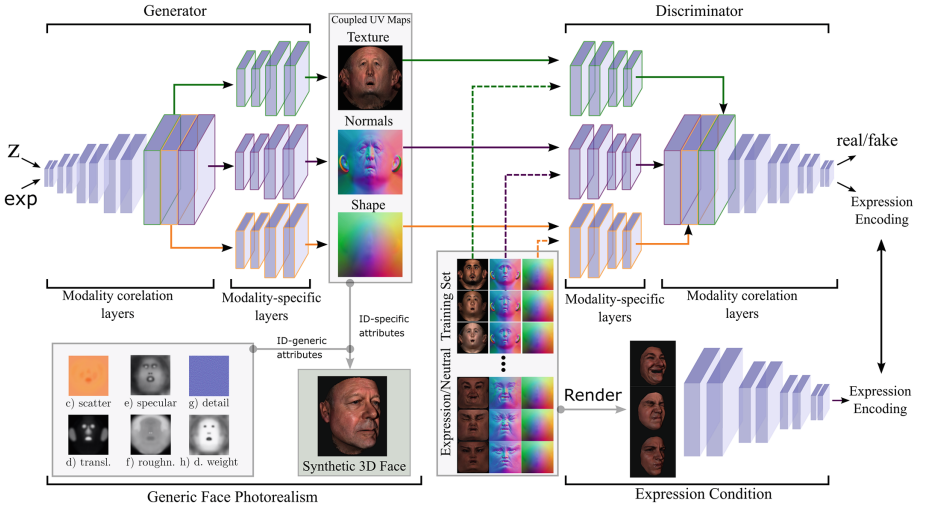



**Fig. 1.** We propose a novel GAN that can synthesize high-quality texture, shape, and normals jointly for realistic and coherent 3D faces of novel identities. The separation of branch networks allows the specialization of the characteristicofeach one of the modalities while the trunk network maintains the local correspondences among them. Moreover, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions. We annotate the training dataset automatically by an expression recognition network to couple those expression encodings to the texture, shape, and normals UV maps.

games, etc. Other applications of such approaches include generating synthetic training data for face recognition [23] and modeling the face manifold for 3D face reconstruction [24]. Currently, 3D face generation in computer games and movies is performed by expensive capturing systems or by professional technical artists. The current state-of-the-art methods generate faces, which can be suitable for applications such as caricature avatar creation in mobile devices [29] but do not generate high-quality shape and normals that can be used for photo-realistic face synthesis. In this paper, we propose the first methodology for high-quality face generation that can be used for photo-realistic face synthesis (i.e., joint generation of texture, shape, and normals) by capitalizing on the recent developments on Generative Adversarial Networks (GANs).

The early face models, such as [6], represent 3D face by disentangled PCA models of geometry, expression [13], and colored texture, called 3D morphable models (3DMM). 3DMMs and its variants were the most popular method for modeling shape and texture separately. However, the linear nature of PCA is often unable to capture high-frequency signals properly, thus the quality of generation and reconstruction by PCA is sub-optimal.

GANs is a recently introduced family of techniques that train samplers of high-dimensional distributions [25]. It has been demonstrated that when a GAN

is trained on facial images, it can generate images that have realistic characteristics. In particular, the recently introduced GANs [11,32,33] can generate photo-realistic high-resolution faces. Nevertheless, because they are trained on partially-aligned 2D images, they cannot properly model the manifold of faces and thus (a) inevitably create many unrealistic instances and (b) it is not clear how they can be used to generate photo-realistic 3D faces.

Recently, GANs have been applied for generating facial texture for various applications. In particular, [54] and [23] utilize style transfer GANs to generate photorealistic images of 3DMM-sampled novel identities. [57] directly generates high-quality 3D facial textures by GANs and [24] replaces 3D Morphable Models (3DMMs) with GAN models for 3D texture reconstruction while the shape is still maintained by statistical models. [35] propose to generate 4K diffuse and specular albedo and normals from a texture map by an image-to-image GAN. On the other hand, [44] model 3D shape by GANs in a parametric UV map and [53] utilize mesh convolutions with variational autoencoders to model shape in its original structure. Although one can model 3D faces with such shape and texture GAN approaches, these studies omit the correlation between shape, normals, and texture which is very important for photorealism in identity space. The significance of such correlation is most visible with inconsistent facial attributes such as age, gender, and ethnicity (i.e. old-aged texture on a baby-face geometry).

In order to address these gaps, we propose a novel multi-branch GAN architecture that preserves the correlation between different 3D modalities (such as texture, shape, normals, and expression). After converting all modalities into UV space and concatenate over channels, we train a GAN that generates all modalities in a meaningful local and global correspondence. In order to prevent incompatibility issues due to the intensity distribution of different modalities, we propose a trunk-branch architecture that can synthesize photorealistic 3D faces with coupled texture and geometry. Further, we condition this GAN by expression labels to generate faces in any desired expression.

From a computer graphics point of view, a photorealistic face rendering requires a number of elements to be tailored, i.e. shape, normals and albedo maps, some of which should or can be specific to a particular identity. However, the cost of hand-crafting novel identities limits their usage on large-scale applications. The proposed approach tackles this down with reasonable photorealism with a massively generalized identity space. Although the results in this paper are limited to aforementioned modalities by the dataset at hand, the proposed method allows adding more identity-specific modalities (i.e. cavity, gloss, scatter) once such a dataset becomes available.

The contributions of this paper can be summarized as follows:

– We propose to model and synthesize coherent 3D faces by jointly training a novel Trunk-branch based GAN (TBGAN) architecture for shape, texture, and normals modalities. TBGAN is designed to maintain correlation while tolerating domain-specific differences of these three modalities and can be easily extended to other modalities and domains.

– In the domain of identity-generic face modeling, we believe this is the first study that utilizes normals as an additional source of information.
– We propose the first methodology for face generation that correlates expression and identity geometries (i.e. modeling personalized expression) and also the first attempt to model expression in texture and normals space.

## 2 Related Work

### 2.1 3D Face Modeling

There is an underlying assumption that human faces lie on a manifold with respect to the appearance and geometry. As a result, one can model the geometry and appearance of the human face analytically based upon the identity and expression space of all individuals. Two of the first attempts in the history of face modeling were [1], which proposes part-based 3D face reconstruction from frontal and profile images, and [48], which represents expression action units by a set of muscle fibers.

Twenty years ago methods that generated 3D faces revolved around parametric generative models that are driven by a small number of anthropometric statistics (e.g., sparse face measurements in a population) which act as constraints [18]. The seminal work of 3D morphable models (3DMMs) [6] demonstrated for the first time that is possible to learn a linear statistical model from a population of 3D faces [12,46]. 3DMMs are often constructed by using a Principal Component Analysis (PCA) based on a dataset of registered 3D scans of hundreds [47] or thousands [7] subjects. Similarly, facial expressions are also modeled by applying PCA [2,10,38,62], or are manually defined using linear blendshapes [9,36,58]. 3DMMs, despite their advantages, are bounded by the capacity of linear space that under-represents the high-frequency information and often result in overly-smoothed geometry and texture models. [14] and [59] attempt to address this issue by using local displacement maps. Furthermore, the 3DMM line of research assumes that texture and shape are uncorrelated, hence they can only be produced by separate models (i.e., separate PCA models for texture and shape). Early attempts in correlated shape and texture have been made in Active Appearance Models (AAMs) by computing joint PCA models of sparse shape and texture [16]. Nevertheless, due to the inherent limitations of PCA to model high-frequency texture, it is rarely used to correlate shape and texture for 3D face generation.

Recent progress in generative models [25,34] is being utilized in 3D face modeling to tackle this issue. [44] trained a GAN that models face geometry based on UV representations for neutral faces, and likewise, [53] modeled identity and expression geometry by variational autoencoders with mesh convolutions. [24] proposed a GAN-based texture modeling for 3D face reconstruction while modeling geometry by PCA and [57] trained a GAN to synthesize facial textures. To the best of our knowledge, these methodologies totally omit the correlation between geometry and texture and moreover, they ignore identity-specific expression modeling by decoupling them into separate models. In order to address this

issue, we propose a trunk-branch GAN that is trained jointly for texture, shape, normals, and expression in order to leverage non-linear generative networks for capturing the correlation between these modalities.

### 2.2 Photorealistic Face Synthesis

Although most of the aforementioned 3D face models can synthesize 2D face images, there are also some dedicated 2D face generation studies. [42] combines non-parametric local and parametric global models to generate various set of face images. Recent family of GAN approaches [11,32,33,52] offers the state-of-the-art high quality random face generation without constraints.

Some other GAN-based studies allow to condition synthetic faces by rendered 3DMM images [23], by landmarks [5] or by another face image [4] (i.e. by disentangling identity and certain facial attributes). Similarly, facial expression is also conditionally synthesized by an audio input [31], by action unit codes [51], by predefined 3D geometry [65] or by expression of an another face image [37].

In this work, we jointly synthesize the aforementioned modalities for coherent photorealistic face synthesis by leveraging high-frequency generation by GANs. Unlike many of its 2D and 3D alternatives, the resulting generator models provide absolute control over disentangled identity, pose, expression and illumination spaces. Unlike many other GAN works that are struggling due to misalignments among the training data, our entire latent space correspond to realistic 3D faces as the data representation is naturally aligned on UV space.

### 2.3 Boosting Face Recognition by Synthetic Training Data

There have been also some works to synthesize face images to be used as synthetic training data for face recognition methods either by directly using GAN-generated images [61] or by controlling pose-space with a conditional-GAN [30,56,60]. [41] propose many augmentation techniques, such as rotation, expression, and shape, based on 3DMMs. Other GAN-based approaches that capitalize 3D facial priors include [66], which rotates faces by fitting 3DMM and preserves photorealism by translation GANs and [64], which frontalize face images by a GAN and 3DMM regression network. [19] complete missing parts of UV texture representations of 2D images after 3DMM fitting by a translation GAN. [23] first synthesizes face images of novel identities by sampling from 3DMM and then removes the photorealistic domain gap by an image-to-image translation GAN.

All of these studies show the significance of photorealistic and identity-generic face synthesization for the next generation of facial recognition algorithms. Although this study focuses more on the graphical aspect of face synthesization, we show that synthetic images can also improve face recognition performance.

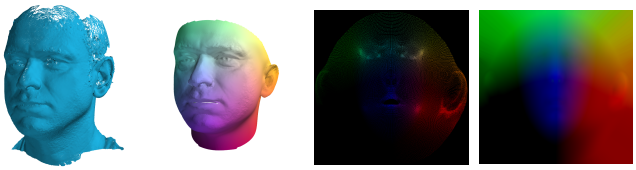



**Fig. 2.** UV extraction process. In (a) we present a raw mesh, in (b) the registered mesh using the Large Scale Face Model (LSFM) template [8], in (c) the unwrapped 3D mesh in the 2D UV space, and (d) the interpolated 2D UV map. Interpolation is carried out using the barycentric coordinates of each pixel in the registered 3D mesh.

## 3 Approach

### 3.1 UV Maps for Shape, Texture and Normals

In order to feed the shape, the texture, and the normals of the facial meshes into a deep network we need to reparameterize them into an image-like tensor format to apply 2D-convolutions[1]. We begin by describing all the raw 3D facial scans with the same topology and number of vertices (dense correspondence). This is achieved by morphing non-rigidly a template mesh to each one of the raw scans. We employ a standard non-rigid iterative closest point algorithm as described in [3,17] and we deform our chosen template so that it captures correctly the facial surface of the raw scans. As a template mesh, we choose the mean face of the LSFM model proposed in [8], which consists approximately of $54K$ vertices that are sufficient enough to depict non-linear, high facial details.

After reparameterizing all the meshes into the LSFM [8] topology, we cylindrically unwrap the mean face of the LSFM [8] to create a UV representation for that specific mesh topology. In the literature, a UV map is commonly utilized for storing only the RGB texture values. Apart from storing the texture values of the 3D meshes, we utilize the UV space to store the 3D coordinates of each vertex $(x, y, z)$ and the normal orientation $(n_x, n_y, n_z)$. Before storing the 3D coordinates into the UV space, all meshes are aligned in the 3D spaces by performing General Procrustes Analysis (GPA) [26] and are normalized to be in the scale of $[1, -1]$. Moreover, we store each 3D coordinate and normals in the UV space given the respective UV pixel coordinate. Finally, we perform a barycentric interpolation based on the barycentric coordinates of each pixel on the registered mesh to fill out the missing areas in order to produce a dense illustration of the UV map. In Fig. 2, we illustrate a raw 3D scan, the registered 3D scan on the LSFM [8] template, the sparse UV map of 3D coordinates and finally the interpolated one.

[1] Another line of research is mesh convolutional networks [15,39,53] which cannot preserve high-frequency details of the texture and normals at the current state-of-the-art.

### 3.2 Trunk-Branch GAN to Generate Coupled Texture, Shape and Normals

In order to train a model that handles multiple modalities, we propose a novel trunk-branch GAN architecture to generate entangled modalities of the 3D face such as texture, shape, and normals as UV maps. For this task, we exploit the MeIn3D dataset [8] which consists of approximately 10,000 neutral 3D facial scans with wide diversity in age, gender, and ethnicity.

Given a generator network $\mathcal{G}^L$ with a total of $L$ convolutional upsampling layers and gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as input, the activation at the end of layer $d$ (i.e., $\mathcal{G}^d(\mathbf{z})$) is split into three branch networks $\mathcal{G}_T^{L-d}$, $\mathcal{G}_N^{L-d}$, $\mathcal{G}_S^{L-d}$ each of which consists of $L-d$ upsampling convolutional layers that generate texture, normals and shape UV maps respectively. The discriminator $\mathcal{D}^L$ starts with the branch networks $\mathcal{D}_T^{L-d}$, $\mathcal{D}_N^{L-d}$, $\mathcal{D}_S^{L-d}$ whose activations are concatenated before fed into trunk network $\mathcal{D}^d$. The output of $\mathcal{D}^L$ is regression of real/fake score.

Although the proposed approach is compatible with most of the GAN architectures and loss functions, in our experiments, we base TBGAN on progressive growing GAN architecture [32] train it by WGAN-GP loss [27] as following:

$$\mathcal{L}_{\mathcal{G}^L} = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[-\mathcal{D}^L\left(\mathcal{G}^L(\mathbf{z})\right)\right] \tag{1}$$

$$\mathcal{L}_{\mathcal{D}^L} = \mathbb{E}_{x\sim p_{\text{data}},\ \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathcal{D}^L\left(\mathcal{G}^L(\mathbf{z})\right) - \mathcal{D}^L(x) + \lambda * GP(x, \mathcal{G}^L(\mathbf{z}))\right] \tag{2}$$

where gradient penalty calculated by $GP(x,\hat{x}) = \left(\|\nabla\mathcal{D}^L\left(\alpha\hat{x} + (1-\alpha)x\right)\right.$ $\left.\|_2 - 1\right)^2$ and $\alpha$ denotes uniform random numbers between 0 and 1. $\lambda$ is a balancing factor which is typically $\lambda = 10$. An overview of this trunk-branch architecture is illustrated in Fig. 1

### 3.3 Expression Augmentation by Conditional GAN

Further, we modify our GAN in order to generate 3D faces with expression by conditioning it with expression annotations ($\mathbf{p_e}$). Similar to the MeIn3D dataset, we have captured approximately 35,000 facial scans of around 5,000 distinct identities during a special exhibition in the Science Museum, London. All subjects were recorded in various guided expressions with a 3dMD face capturing apparatus. All of the subjects were asked to provide meta-data regarding their age, gender, and ethnicity. The database consists of 46% male, 54% female, 85% White, 7% Asian, 4% Mixed Heritage, 3% Black, and 1% other.

In order to avoid the cost and potential inconsistency of manual annotation, we render those scans and automatically annotate them by an expression recognition network. The resulting expression encodings $((*, \mathbf{p_e}) \sim p_{\text{ data}})$ are used as label vector during the training of our trunk-branch conditional GAN. This training scheme is illustrated in Fig. 1. $\mathbf{p_e}$ is basically a vector of 7 for universal expressions (neutral, happy, angry etc.), randomly drawn from our dataset. During the training, Eq. 1 and 2 are updated to condition expression encodings by AC-GAN [45] as following:

$$\mathcal{L}_{\mathcal{G}^L} \mathrel{+}= \mathbb{E}_{(*,\mathbf{p_e})\sim p_{\text{data}},\ \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\sum_e \mathbf{p_e}\log(\mathcal{D}_e^L(\mathcal{G}^L(\mathbf{z},\mathbf{p_e})))\right] \tag{3}$$

$$\mathcal{L}_{\mathcal{D}^L} \mathrel{+}= \mathbb{E}_{(x,\mathbf{p_e})\sim p_{\text{data}},\ \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\sum_e \mathbf{p_e}\log(\mathcal{D}_e^L(x)) + \mathbf{p_e}\log(\mathcal{D}_e^L(\mathcal{G}^L(\mathbf{z},\mathbf{p_e})))\right] \tag{4}$$

which performs softmax cross entropy between expression prediction of the discriminator ($\mathcal{D}_e^L(x)$) and the random expression vector input ($\mathbf{p_e}$) for real ($x$) and generated samples ($\mathcal{G}^L(\mathbf{z},\mathbf{p_e})$).

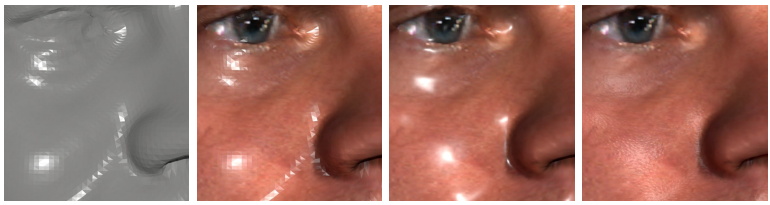
Unlike previous expression models that omit the effect of the expression on textures, the resulting generator is capable of generating coupled texture, shape, and normals map of a face with controlled expression. Similarly, our generator respects the identity-expression correlation thanks to correlated supervision provided by the training data. This is in contrast to the traditional statistical expression models which decouples expression and identity models into two separate entities.

### 3.4 Photorealistic Rendering with Generated UV Maps

For the renderings to appear photorealistic, we use the generated identity-specific mesh, texture, and normals, in combination with the generic reflectance properties, and employ a commercial rendering application: *Marmoset Toolbag* [40].

In order to extract the 3D representation from the UV domain we employ the inverse procedure explained in Sect. 3.1 based on the UV pixel coordinates of each vertex of the 3D mesh. Figure 3 shows the rendering results, under a single light source, when using the generated geometry (Fig. 3(a)) and the generated texture (Fig. 3(b)). Here the specular reflection is calculated on the per-face normals of the mesh and exhibits steep changes between on the face's edges. By interpolating the generated normals on each face (Fig. 3(c)), we are able to smooth the specular highlights and correct any high-frequency noise on the geometry of the mesh. However, these results do not correctly model the human skin and resemble a metallic surface. In reality, the human skin is rough and as a body tissue, it both reflects and absorbs light, thus exhibiting specular reflection, diffuse reflection, and subsurface scattering.

Although we can add such modalities as additional branches with the availability of such data, we find that rendering can be still improved by adding some identity-generic maps. Using our training data, we create maps that define certain reflectance properties per-pixel, which will match the features of the average generated identity, as shown in bottom-left of Fig. 1. *Scattering* (c) defines the intensity of subsurface scattering of the skin. *Translucency* (d) defines the amount of light, that travels inside the skin and gets emitted in different directions. *Specular albedo* (e) gives the intensity of the specular highlights, which differ between hair-covered areas, the eyes, and the teeth. *Roughness* (f) describes the scattering of specular highlights and controls the glossiness of the skin.



(a) shape (b) Shp+tex (c) Shp+tex+nor (d) Final

**Fig. 3.** Zoom-in on rendering results with (a) only the shape, (b) adding the albedo texture, (c) adding the generated normals, and (d) using identity-generic detail normal, specular albedo, roughness, scatter and translucency maps.

A *detail normal map* (g) is also tilled and added on the generated normal maps, to mimic the skin pores and a *detail weight map* (h) controls the appearance of the detail normals, so that they do not appear on the eyes, lips, and hair. The final result (Fig. 3(d)) properly models the skin surface and reflection, by adding plausible high-frequency specularity and subsurface scattering, both weighted by the area of the face where they appear.

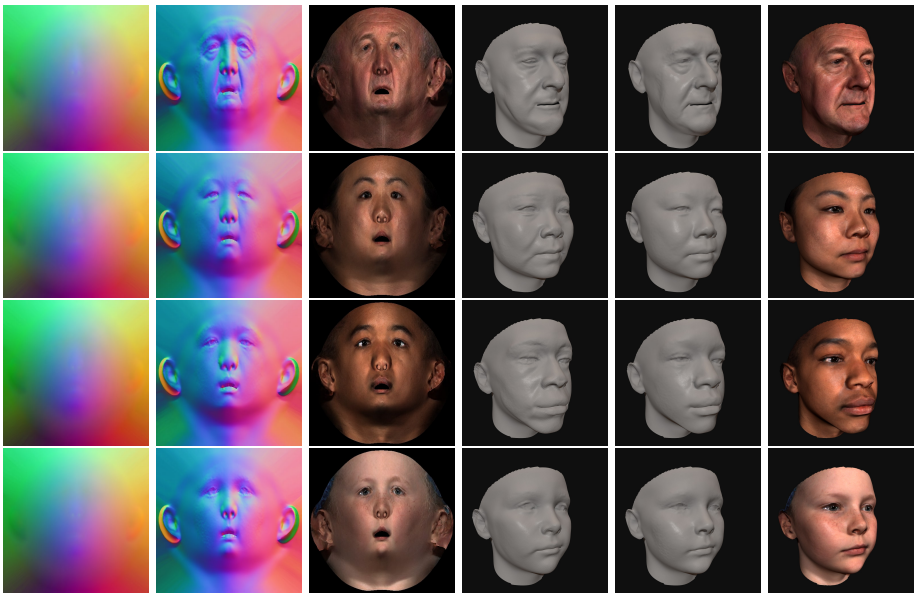
## 4 Results

In this section, we give qualitative and quantitative results of our method for generating 3D faces with novel identities and various expressions. In our experiments, there are total $L = 8$ up- and down-sampling layers where $d = 6$ of them in the trunk and 2 layers in each branch. These choices are empirically validated to ensure sufficient correlation among modalities without incompatibility artifacts. Running time is a few milliseconds to generate UV images from a latent code on a high-end GPU. Transforming from UV image to mesh is just sampling with UV coordinates and can be considered free of cost. Renderings in this paper take a few seconds due to high resolution but this cost depends on the application. The memory needed for the generator network is 1.25 GB compared to the 6 GB PCA model of the same resolution and %95 of the total variance.

In the following sections, we first visualize generated UV maps and their contributions to the final renderings on several generated faces. Next, we show the generalization ability of the identity and expression generators on some facial characteristics. We also demonstrate its well-generalization latent space by interpolating between different identities. Additionally, we perform full-head completion to the interpolated faces. Finally, we perform face recognition experiments by using the generated face images as additional training data.

### 4.1 Qualitative Results

**Combining Coupled Modalities:** Fig. 4 presents the generated shape, normals, and texture maps by the proposed GAN and their additive contributions



(a) Shape (b) Normals (c) Texture (d) Shape (e) Shp+Nor (f) All

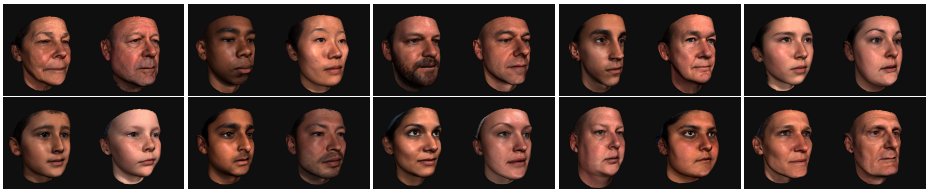
**Fig. 4.** Generated UV representations and their corresponding additive renderings. Please note the strong correlation between UV maps, high fidelity and photorealistic renderings. The figure is best viewed in zoom.

to the final renderings. As can be seen from local and global correspondences, the generated UV maps are highly correlated and coherent. Attributes like age, gender, race, etc. can be easily grasped from all of the UV maps and rendered images. Please also note that some of the minor artifacts of the generated geometry in Fig. 4(d) are compensated by the normals in Fig. 4(e).

**Diversity:** Our model is well-generalized with different age, gender, ethnicity groups and many facial attributes. Although Fig. 5 shows diversity in some of those categories, the reader is encouraged to see identity variation throughout the paper and the supplementary video.

**Expression:** We also show that our expression generator is capable of synthesizing quite a diverse set of expressions. Moreover, the expressions can be controlled by the input label as can be seen in Fig. 6. The reader is encouraged to see more expression generations in the supplementary video.

**Interpolation Between Identities:** As shown in the supplementary video and in Fig. 7, our model can easily interpolate between any generation in a visually continuous set of identities which is another indication that the model is free



(a) Age (b) Ethnicity (c) Gender (d) Weight (e) Roundness

**Fig. 5.** Variation of generated 3D faces by our model. Each block shows diversity in a different aspect. Readers are encouraged to zoom in on a digital version.

**Fig. 6.** (Top) generations of six universal expressions (i.e. each two columns respective the following expressions: Happiness, Sadness, Anger, Fear, Disgust, Surprise). (Middle) texture and (Bottom) normals maps are used to generate the corresponding 3D faces. Please note how expressions are represented and correlated in the texture and normals space.

from mode collapse. Interpolation is done by randomly generating two identities and generates faces by evenly spaced samples in latent space between the two.

**Full Head Completion:** We also extend our facial 3D meshes to full head representations by employing the framework proposed in [50]. We achieve this by regressing from a latent space that represents only the 3D face to the PCA latent space of the Universal Head Model (UHM) [49,50]. We begin by building a PCA model of the inner face based on the 10,000 neutral scans of the MeIn3D dataset. Similarly, we exploit the extended full head meshes of the same identities utilized by UHM model and project them to the UHM subspace to acquire the latent shape parameters of the entire head topology. Finally, we learn a regression matrix by solving a linear least-square optimization problem as proposed in [50], which maps the latent space of the face shape to the full head representation. Figure 7 demonstrates the extended head representations of our approach in conjunction with the synthesized crop faces.

**Comparison to Decoupled Modalities and PCA:** Results in Fig. 8 reveal a set of advantages of such unified 3D face modeling over separate GAN and statistical models. Clearly, the figure shows that the correlation among texture,

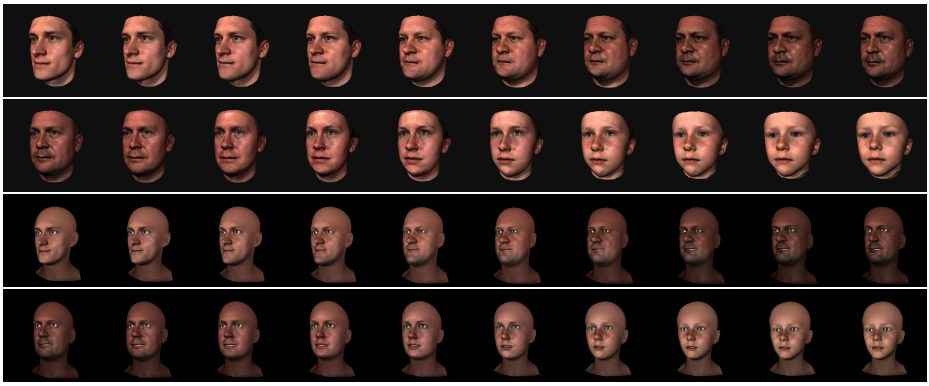


**Fig. 7.** Interpolation between pair of identities in the latent space. Smooth transition indicates generalization of our GAN model. The last two rows show complete full head representations respective to the first two rows.

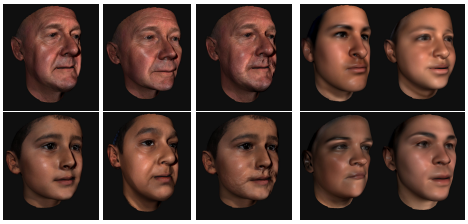



**Fig. 8.** Comparison with separate GAN models and PCA model. (a) Generation by our model. (b) Same texture with random shape and normals. (c) Same texture and shape with random normals (i.e. beard). (d) Generation by a PCA model constructed by the same training data and the same identity-generic rendering tools as explained in Sec. 3.4.

shape, and normals is an important component for realistic face synthesis. Also, generations by PCA models are missing photorealism and details significantly.

### 4.2 Pose-Invariant Face Recognition

In this section, we present an experiment that demonstrates that the proposed methodology can generate faces of different and diverse identities. That is, we use the generated faces to train one of the most recent state-of-the-art face recognition method, ArcFace [20], and show that the proposed shape and texture generation model can boost the performance of pose-invariant face recognition.
**Training Data:** We randomly synthesize 10 K new identities from the proposed model and render 50 images per identity with a random camera and illumination parameters from the Gaussian distribution of the 300W-LP dataset [23,67]. For clarity, we call this dataset "Gen" in the rest of the text. Figure 9 illustrates

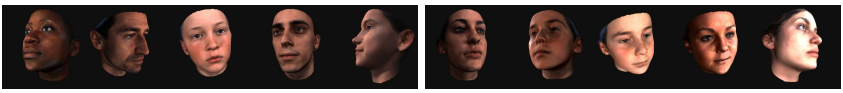


**Fig. 9.** Examples of generated data ("Gen") by the proposed method.

some examples of "Gen" dataset which show larger pose variations than the real-world collected data. We augment "Gen" with an in-the-wild training data, CASIA dataset [63], which consists of 10,575 identities with 494,414 images.

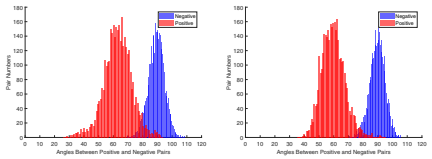
**Test Data:** For evaluation, we employ Celebrities in Frontal Profile (CFP) [55] and Age Database (AgeDB) [43]. **CFP** [55] consists of 500 subjects, each with 10 frontal and 4 profile images. The evaluation protocol includes frontal-frontal (FF) and frontal-profile (FP) face verification. In this paper, we focus on the most challenging subset, CFP-FP, to investigate the performance of pose-invariant face recognition. There are 3,500 same-person pairs and 3,500 different-person pairs in CFP-FP for the verification test. **AgeDB** [21,43] contains 12,240 images of 440 distinct subjects. The minimum and maximum ages are 3 and 101, respectively. The average age range for each subject is 49 years. There are four groups of test data with different year gaps (5 years, 10 years, 20 years and 30 years, respectively) [21]. In this paper, we only use the most challenging subset, AgeDB-30, to report the performance. There are 3,000 positive pairs and 3,000 negative pairs in AgeDB-30 for the verification test.

**Data Prepossessing:** We follow the baseline [20] to generate the normalized face crops ($112 \times 112$) by utilizing five facial points.

**Training and Testing Details:** For the embedding networks, we employ the widely used ResNet50 architecture [28]. After the last convolutional layer, we also use the BN-Dropout-FC-BN [20] structure to get the final 512-$D$ embedding feature. For the hyper-parameter setting and loss functions, we follow [20–22]. The overlapping identities between the CASIA data set and the test set are removed for strict evaluations, and we only use a single crop for all testing.

**Result Analysis:** In Table 1, we show the contribution of the generated data on pose-invariant face recognition. We take UV-GAN [19] as the baseline method, which attaches the completed UV texture map onto the fitted mesh and generates instances of arbitrary poses to increase pose variation during training and minimize pose discrepancy during testing. As we can see from Table 1, generated data significantly boost the verification performance on CFP-FP from 95.56% to 97.12%, decreasing the verification error by 51.2% compared to the result of UV-GAN [19]. On AgeDB-30, combining CASIA and generated data achieves similar performance compared to using single CASIA because we only include intra-variance from pose instead of age.

In Fig. 10, we show the angle distributions of all positive pairs and negative pairs from CFP-FP. By incorporating generation data, the overlap indistinguishable area between the positive histogram and the negative histogram is obviously decreased, which confirms that ArcFace can learn pose-invariant




(a) CASIA (b) CASIA+Gen

**Fig. 10.** Angle distributions of CFP-FP positive (red) and negative (blue) pairs in the 512-$D$ feature space

**Table 1.** Verification performance (%) of different models on CFP-FP and AgeDB-30.

| Methods | CFP-FP | AgeDB-30 |
|---|---|---|
| UVGAN [19] | 94.05 | 94.18 |
| Ours (CASIA) | 95.56 | 95.15 |
| Ours (CASIA+Gen) | **97.12** | **95.18** |

**Table 2.** The angles between face pairs from CFP-FP predicted by different models trained from the CASIA and combined data. The generated data can obviously enhance the pose-invariant feature embedding.

| Training Data |  |  |  |  |  |
|---|---|---|---|---|---|
| CASIA | 84.06° | 82.39° | 84.72° | 88.06° | 84.37° |
| CASIA+Gen | 57.60° | 63.12° | 66.10° | 59.72° | 60.25° |

feature embedding from the generated data. In Table 2, we select some verification pairs from CFP-FP and calculate the cosine distance (*angle*) between these pairs predicted by different models trained from the CASIA and combined data. Intuitively, the angles between these challenging pairs are significantly reduced when generated data are used for the model training.

## 5 Conclusion

We presented the first 3D face model for joint texture, shape, and normal generation based on Generative Adversarial Networks (GANs). The proposed GAN model implements a new architecture for exploiting the correlation between different modalities and can synthesize different facial expressions in accordance with the embeddings of an expression recognition network. We demonstrate that randomly synthesized images of our unified generator show strong relations between texture, shape, and normals and that rendering with normals provides excellent shading and overall visual quality. Finally, in order to demonstrate the generalization of our model, we have used a set of generated images to train a deep face recognition network.

**Acknowledgement.** Baris Gecer is supported by the Turkish Ministry of National Education, Stylianos Ploumpis by the EPSRC Project EP/N007743/1 (FACER2VM), and Stefanos Zafeiriou by EPSRC Fellowship DEFORM (EP/S010203/1).

## References


1. Akimoto, T., Suenaga, Y., Wallace, R.S.: Automatic creation of 3D facial models. IEEE Comput. Graphics Appl. **13**(5), 16–22 (1993). https://doi.org/10.1109/38.232096
2. Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3D face recognition with a morphable model. In: 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008, pp. 1–6. IEEE (2008). https://doi.org/10.1109/AFGR.2008.4813376
3. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid ICP algorithms for surface registration. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007). https://doi.org/10.1109/CVPR.2007.383165
4. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6713–6722 (2018). https://doi.org/10.1109/CVPR.2018.00702
5. Bazrafkan, S., Javidnia, H., Corcoran, P.: Face synthesis with landmark points from generative adversarial networks and inverse latent space mapping. arXiv preprint arXiv:1802.00390 (2018)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999). https://doi.org/10.1145/311535.311556
7. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3D morphable models. Int. J. Comput. Vision **126**(2–4), 233–254 (2018). https://doi.org/10.1007/s11263-017-1009-7
8. Booth, J., Roussos, A., Zafeiriou, S., Ponniahy, A., Dunaway, D.: A 3D morphable model learnt from 10,000 faces. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 5543–5552 (2016). https://doi.org/10.1109/CVPR.2016.598
9. Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Trans. Graph. **32**(4), 40 (2013). https://doi.org/10.1145/2461912.2461976
10. Breidt, M., Biilthoff, H.H., Curio, C.: Robust semantic analysis by synthesis of 3D facial motion. In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, pp. 713–719. IEEE (2011). https://doi.org/10.1109/FG.2011.5771336
11. Brock, A., Donahue, J., Simonyan, K.: Large scale GaN training for high fidelity natural image synthesis. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
12. Brunton, A., Salazar, A., Bolkart, T., Wuhrer, S.: Review of statistical shape spaces for 3D data with comparative analysis for human faces. Comput. Vis. Image Underst. **128**, 1–17 (2014). https://doi.org/10.1016/j.cviu.2014.05.005
13. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: FaceWarehouse: a 3D facial expression database for visual computing. IEEE Trans. Visual Comput. Graphics **20**(3), 413–425 (2014). https://doi.org/10.1109/TVCG.2013.249
14. Chen, A., Chen, Z., Zhang, G., Mitchell, K., Yu, J.: Photo-realistic facial details synthesis from single image. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2019-October, pp. 9428–9438, October 2019. https://doi.org/10.1109/ICCV.2019.00952

15. Cheng, S., Bronstein, M., Zhou, Y., Kotsia, I., Pantic, M., Zafeiriou, S.: MeshGAN: non-linear 3D morphable models of faces. arXiv preprint arXiv:1903.10384 (2019)
16. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0054760
17. De Smet, M., Van Gool, L.: Optimal regions for linear model-based 3D face reconstruction. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6494, pp. 276–289. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19318-7_22
18. DeCarlo, D., Metaxas, D., Stone, M.: An anthropometric face model using variational techniques. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1998, vol. 98, pp. 67–74 (1998). https://doi.org/10.1145/280814.280823
19. Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S.: UV-GAN: adversarial facial UV map completion for pose-invariant face recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7093–7102 (2018). https://doi.org/10.1109/CVPR.2018.00741
20. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2019, pp. 4685–4694 (2019). https://doi.org/10.1109/CVPR.2019.00482
21. Deng, J., Zhou, Y., Zafeiriou, S.: Marginal loss for deep face recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2017-July, pp. 2006–2014 (2017). https://doi.org/10.1109/CVPRW.2017.251
22. Gecer, B., Balntas, V., Kim, T.K.: Learning deep convolutional embeddings for face representation using joint sample- and set-based supervision. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 1665–1672, October 2017. https://doi.org/10.1109/ICCVW.2017.195
23. Gecer, B., Bhattarai, B., Kittler, J., Kim, T.-K.: Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 230–248. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_14
24. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: GANFIT: generative adversarial network fitting for high fidelity 3D face reconstruction. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1155–1164, June 2019. https://doi.org/10.1109/CVPR.2019.00125
25. Goodfellow, I.J., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 3, pp. 2672–2680 (2014). https://doi.org/10.3156/jsoft.29.5_177_2
26. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975). https://doi.org/10.1007/BF02291478
27. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein GANs. In: Advances in Neural Information Processing Systems, vol. 2017-December, pp. 5768–5778 (2017)

28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
29. Hu, L., et al.: Avatar digitization from a single image for real-time rendering. ACM Trans. Graph. **36**(6), 195 (2017). https://doi.org/10.1145/3130800.3130887
30. Hu, Y., Wu, X., Yu, B., He, R., Sun, Z.: Pose-guided photorealistic face rotation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8398–8406 (2018). https://doi.org/10.1109/CVPR.2018.00876
31. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: synthesising talking faces from audio. Int. J. Comput. Vision **127**(11–12), 1767–1779 (2019). https://doi.org/10.1007/s11263-019-01150-y
32. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings (2018)
33. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, pp. 4396–4405 (2019). https://doi.org/10.1109/CVPR.2019.00453
34. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2014)
35. Lattas, A., et al.: AvatarMe: realistically renderable 3D facial reconstruction "in-the-wild". In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 760–769 (2020)
36. Li, H., Weise, T., Pauly, M.: Example-based facial rigging. ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010 **29**(4), 32 (2010). https://doi.org/10.1145/1778765.1778769
37. Li, K., Dai, Q., Wang, R., Liu, Y., Xu, F., Wang, J.: A data-driven approach for facial expression retargeting in video. IEEE Trans. Multimedia **16**, 299–310 (2014). https://doi.org/10.1109/TMM.2013.2293064
38. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. **36**(6), 194 (2017). https://doi.org/10.1145/3130800.3130813
39. Litany, O., Bronstein, A., Bronstein, M., Makadia, A.: Deformable shape completion with graph convolutional autoencoders. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1886–1895 (2018). https://doi.org/10.1109/CVPR.2018.00202
40. Marmoset LLC: Marmoset toolbag (2019)
41. Masi, I., Tran, A.T., Hassner, T., Leksut, J.T., Medioni, G.: Do we really need to collect millions of faces for effective face recognition? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 579–596. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_35
42. Mohammed, U., Prince, S.J., Kautz, J.: Visio-lization: generating novel facial images. ACM Trans. Graph. **28**(3), 57 (2009). https://doi.org/10.1145/1531326.1531363

43. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: AgeDB: the first manually collected, in-the-wild age database. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2017-July, pp. 1997–2005 (2017). https://doi.org/10.1109/CVPRW.2017.250
44. Moschoglou, S., Ploumpis, S., Nicolaou, M., Papaioannou, A., Zafeiriou, S.: 3DFaceGAN: adversarial nets for 3D face representation, generation, and translation. arXiv preprint arXiv:1905.00307 (2019)
45. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, vol. 70, pp. 2642–2651. ICML 2017, JMLR.org, August 2017
46. Patel, A., Smith, W.A.: 3D morphable face models revisited. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, vol. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1327–1334. IEEE (2009). https://doi.org/10.1109/CVPRW.2009.5206522
47. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301 (Sep 2009). https://doi.org/10.1109/AVSS.2009.58
48. Platt, S.M., Badler, N.I.: Animating facial expressions. In: Proceedings of the 8th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1981, vol. 15, pp. 245–252. ACM (1981). https://doi.org/10.1145/800224.806812
49. Ploumpis, S., et al.: Towards a complete 3D morphable model of the human head. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), 1–1 (2020). https://doi.org/10.1109/TPAMI.2020.2991150
50. Ploumpis, S., Wang, H., Pears, N., Smith, W.A., Zafeiriou, S.: Combining 3D morphable models: a large scale face-and-head model. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2019-June, pp. 10926–10935 (2019). https://doi.org/10.1109/CVPR.2019.01119
51. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: GANimation: anatomically-aware facial animation from a single image. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 835–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_50
52. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings (2016)
53. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 725–741. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_43
54. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. vol. 2017-October, pp. 1585–1594 (2017). https://doi.org/10.1109/ICCV.2017.175
55. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016 (2016). https://doi.org/10.1109/WACV.2016.7477558

56. Shen, Y., Luo, P., Yan, J., Wang, X., Tang, X.: FaceID-GAN: learning a symmetry three-player GAN for identity-preserving face synthesis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2018). https://doi.org/10.1109/CVPR.2018.00092
57. Slossberg, R., Shamai, G., Kimmel, R.: High quality facial surface and texture synthesis via generative adversarial networks. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11131, pp. 498–513. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11015-4_36
58. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 181–183 (2015). https://doi.org/10.1145/2816795.2818056
59. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.: Extreme 3D face reconstruction: seeing through occlusions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 3935–3944. IEEE, June 2018. https://doi.org/10.1109/CVPR.2018.00414
60. Tran, L., Yin, X., Liu, X.: Representation learning by rotating your faces. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 3007–3021 (2019). https://doi.org/10.1109/TPAMI.2018.2868350
61. Trigueros, D.S., Meng, L., Hartnett, M.: Generating photo-realistic training data to improve face recognition accuracy. arXiv preprint arXiv:1811.00112 (2018)
62. Yang, F., Metaxas, D., Wang, J., Shechtman, E., Bourdev, L.: Expression flow for 3D-aware face component transfer. ACM Trans. Graph. **30**(4), 1–10 (2011). https://doi.org/10.1145/2010324.1964955
63. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)
64. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 4010–4019 (2017). https://doi.org/10.1109/ICCV.2017.430
65. Zhang, Q., Liu, Z., Guo, B., Shum, H.: Geometry-driven photorealistic facial expression synthesis. In: Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2003, vol. 12, no. 1, pp. 48–60 (2003)
66. Zhao, J., et al.: Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In: Advances in Neural Information Processing Systems, vol. 2017-December, pp. 66–76 (2017)
67. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 146–155 (2016). https://doi.org/10.1109/CVPR.2016.23