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

Baris Gecer[1,2], Alexander 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
[2]FaceSoft.io
{b.gecer, alexandros.lattas17, s.ploumpis, j.deng16, a.papaioannou11, stylianos.moschoglou15, s.zafeiriou}@imperial.ac.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 pre-print is compressed due to size limitations, full resolution results and the accompanying video can be found at the project page: https://github.com/barisgecer/TBGAN.

1 Introduction

Generating 3D faces with high-quality texture, shape and normals is of paramount importance in computer graphics, movie post-production, computer games etc. Other applications of such approaches include generating synthetic training data for face recognition [1] and modeling the face manifold for 3D face reconstruction [2]. 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 [3] 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 capitalising on the recent developments on Generative Adversarial Networks (GANs).

The early face models such as [4] represents 3D face by disentangled PCA models of geometry, expression [5] and colored texture, called 3D morphable models (3DMM). 3DMMs and its variants were the most popular method for modelling shape and texture separately. However, the linear nature of PCA is often unable to capture high frequency signals properly and, 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 [6]. It has been demonstrated that when a GAN is trained on facial images, it can generate images that have main realistic characteristics. In particular, the recently introduced GANs [7], [8], [9] can generate photo-realistic high-resolution faces. Nevertheless, because they are trained on 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, [10] and [1] utilize style transfer GANs to generate photorealistic images of 3DMM-sampled novel identities. [11] directly generates high quality 3D facial textures by GANs and [2] replaces 3D Morphable Models (3DMMs) with GAN models for 3D texture reconstruction while shape is still maintained by statistical models. On the other hand, [12] model 3D shape by GANs in a parametric UV map and [13] 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-

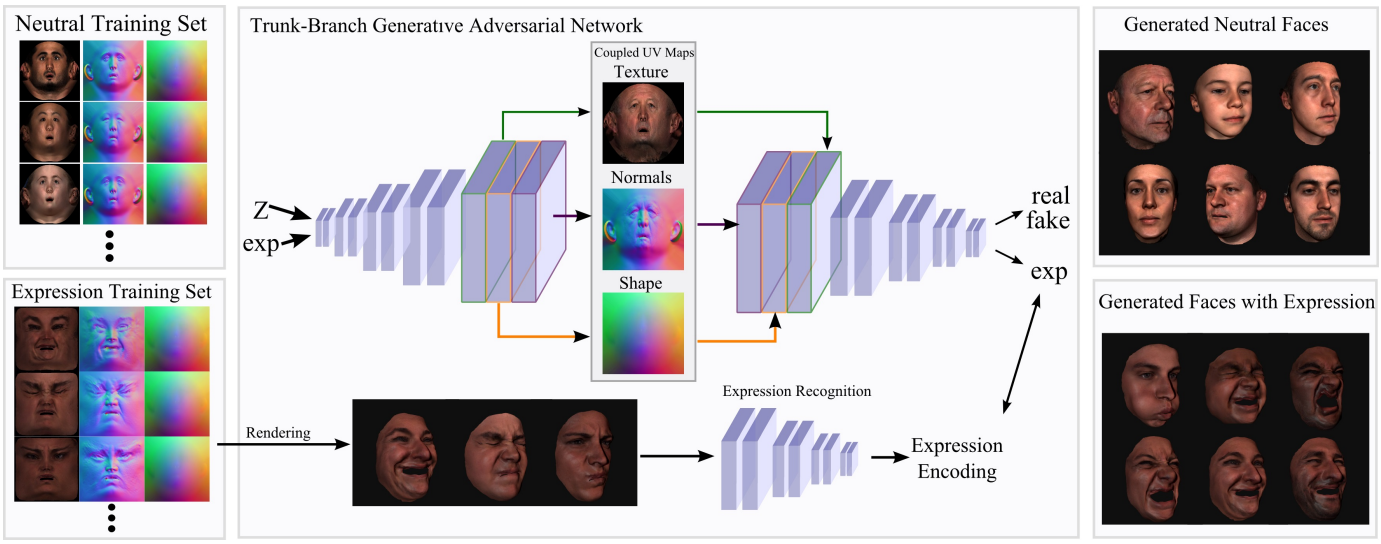


Fig. 1: We propose a novel generative adversarial network that can synthesize high-quality texture, shape, and normals jointly for realistic and coherent 3D faces. Moreover, we demonstrate how we can condition the generation on the expression and create faces with various facial expressions.

crafting novel identities limits their usage on applications that requires a large number of identities. The proposed approach tackles this down with reasonable photorealism with 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 dataset becomes available.

The contributions of this paper can be summarized as following:

- We propose to model and synthesize coherent 3D faces by jointly training a novel Trunk-branch based GAN architecture for shape, texture and normals modalities. TB-GAN is designed to maintain correlation while tolerating domain-specific differences of the three modalities and can be easily extended to other modalities and domains.
- In the domain of identity-generic face modelling, 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. modelling personalized expression) and also the first attempt to model expression in texture and normals space.

2 Related Work

2.1 3D face modelling

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 [14], which proposes part-based 3D face reconstruction from frontal and profile images, and [15], 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 [16]. The seminal work of 3D morphable models (3DMMs) [4] demonstrated for the first time that is possible to learn a linear statistical model from a population of 3D faces [17], [18]. 3DMMs are often constructed by using a Principal Component Analysis (PCA) based on a dataset of registered 3D scans of hundreds [19] or thousands [20] subjects. Similarly, facial expressions are also modeled by applying PCA [21], [22], [23], [24], or are manually defined using linear blendshapes [25], [26], [27]. 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. 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 [28]. Nevertheless, due to the inherent limitations of PCA to model high-frequency texture PCA is rarely used to correlate shape and texture for 3D face generation.

Recent progress in generative models [6], [29] is using 3D face modelling to tackle this issue. [12] trained a GAN that models face geometry based on UV representations for neutral faces and likewise, [13] modelled identity and expression geometry by variational autoencoders with mesh convolutions. [2] proposed a GAN-based texture modelling for 3D face reconstruction while modelling geometry by PCA and [11] trained a GAN to synthesize facial textures. To the best of our knowledge, these methodologies totally omits the correlation between geometry and texture and moreover they ignore identity-specific expression modelling 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 are safe to synthesize 2D face images, there are also some worth-mentioning

2D face generation studies. [30] combine non-parametric local and parametric global models to generate various set of face images. Recent family of GAN approaches [7], [8], [9], [31] 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 [1], by landmarks [32] or by another face image [33] (i.e. by disentangling identity and certain facial attributes). Similarly, facial expression is also conditionally synthesized by an audio input [34], by action unit codes [35], by predefined 3D geometry [36] or by expression of an another face image [37].

In this work, we jointly synthesize 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 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 [38] or by controlling pose-space with a conditional-GAN [39], [40], [41]. [42] propose many augmentation techniques, such as rotating, changing expression and shape, based on 3DMMs. Other GAN-based approaches that capitalize 3D facial priors include [43], which rotates faces by fitting 3DMM and preserves photorealism by translation GANs and [44] which frontalize face images by an end-to-end translation framework that consist of 3DMM regression network and adversarial supervision. [45] complete missing parts of UV texture representations of 2D images after 3DMM fitting by a translation GAN. [1] first synthesize face images of novel identities by sampling from 3DMM and then remove photorealistic domain gap by a 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 focus more on graphical aspect of face synthesization, we show that the synthetic images can also improve face recognition performance.

2.4 Person-specific face models

There have been a number of studies that propose to model appearance and geometry of only a single or a few identities with an excellent quality. [46] utilize appearance models for the identities whose large number of images captured in a controlled environment. Similarly, [47] propose deep appearance models by variational autoencoders. Even though it produces very high rendering quality with various expressions, this method also requires to capture $\sim$20 million images of subjects in a controlled environment. Although [48] reduce this number to a few in-the-wild images, the quality of identity geometry and appearance is limited with the number and quality of the provided images. Nevertheless, all these studies model up to a few individuals at the same time either by interpolating different captures or training a person-specific generative networks. Our method differs from these methods by its capability to generalize on identity space. It reduces the dimensionality of a 3D mesh with $\sim$50,000 nodes into a vector of 512 in latent space that generalize very well with different identities and expression.

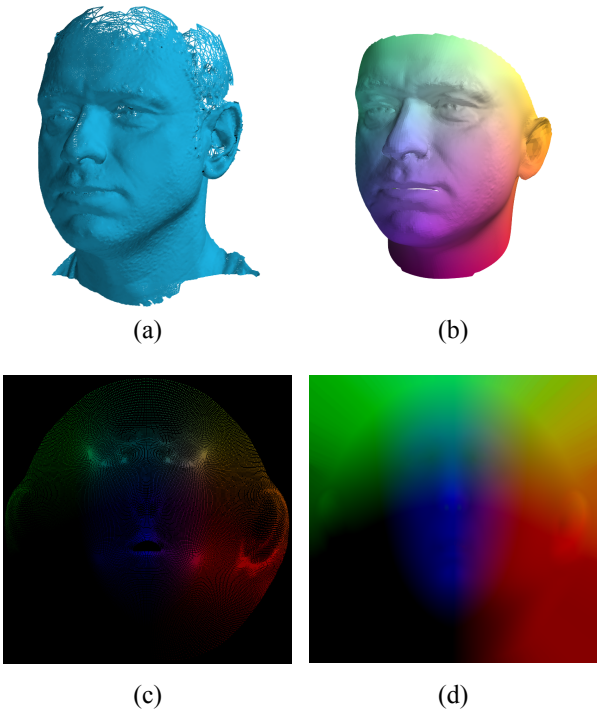


Fig. 2: UV extraction process. In (a) we present a raw mesh, in (b) the registered mesh using the Large Scale Face Model (LSFM) template [49], in (c) the unwrapped 3D mesh in the 2D UV space, and finally in (d) the interpolated 2D UV map. Interpolation is carried out using the barycentric coordinates 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 in order 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 [52], [53] 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 [49], 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 [49] topology, we cylindrically unwrap the mean face of the LSFM [49] to create a UV representation for that specific mesh topology. In literature, a UV map is commonly utilized for storing only the

1. Another line of research is to use convolutions directly on the 3D mesh structure. Nevertheless, when these lines were written the state-of-the-art mesh convolutional networks, e.g. [13], [50], [51], were not able to preserve high-frequency details of the texture and normals.

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) [54] 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 [49] template, the sparse UV map of 3D coordinates and finally the interpolated one.

3.2 Generative Adversarial Networks

Recent advances in generative models have achieved impressive performance in synthesizing diverse set of photorealistic images [6], [29], [31]. Particularly, variants of GANs perform quite well on various generative applications including style/domain transfer [55], super-resolution [56], pose/label guided image generation [39], image inpainting/editing [57] etc. for face [7], [8], body [58] and natural [9] images. In order to achieve variation and photorealism, GANs are being carefully trained by a zero-sum game loss function between two competitor networks: Generator and Discriminator. That is to say, while the generator is being trained to generate samples similar to the training data, the discriminator is trained to separate artificial and real training images from the training set. Both networks improve its performance by benefiting action of one another over the training. At the end, generator network can synthesize photorealistic samples that are aligned with the distribution of the training set.

3.3 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 architecture to generate entangled modalities of 3D face such as texture, shape and normals as UV maps. For this task we exploit the MeIn3D dataset [49] 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 $\mathbf{z} \sim \mathcal{N}(0, 1)$ as input, the activation at the end of layer d (i.e., $\mathcal{G}^d(z)$) is split into three branch networks $\mathcal{G}_T^{L-d}$, $\mathcal{G}_N^{L-d}$, $\mathcal{G}_S^{L-d}$ each of which consist 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 discriminator $\mathcal{D}^d$. The output of $\mathcal{D}^L$ is typically regression of real/fake score (i.e. 1 indicating realism).

Although the proposed approach is competible with most of the GAN architectures and loss functions, in our experiments, we use progressive growing GAN architecture [7] trained by WGAN-GP Wasserstein loss function [59] as following:

$$\mathcal{L}_{\mathcal{G}^L} = -\mathcal{D}^L\left(\mathcal{G}^L(\mathbf{z})\right) \tag{1}$$

$$GP = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x}),\ \boldsymbol{z}\sim p_z(\boldsymbol{z})}\left(\|\nabla\mathcal{D}^L\left(a\mathcal{G}^L(\mathbf{z}) + (1-a)x\right)\right\|_2 - 1)^2 \tag{2}$$

$$\mathcal{L}_{\mathcal{D}^L} = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x}),\ \boldsymbol{z}\sim p_z(\boldsymbol{z})} - \mathcal{D}^L(x) + \mathcal{D}^L\left(\mathcal{G}^L(\mathbf{z})\right) + \lambda * GP \tag{3}$$

where a denotes uniform random numbers between 0 and 1. λ is a balancing factor which is typically $\lambda = 10$. An overview of this trunk-branch architecture is illustrated in Fig. 3

3.4 Expression Augmentation by Conditional GAN

Further, we modify our GAN in order to generate 3D faces with expression by conditioning it with expression annotations. Similarly 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 are used as label vector during the training of our trunk-branch conditional GAN. This training scheme is illustrated in Fig. 4.

Unlike previous expression models which omits the effect of 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.

3.5 Photorealistic Rendering with Generated UV maps

For the final rendering to appear photorealistic, we use the generated identity-specific mesh, texture and normals, in combination with generic modalities the reflectance properties and employ a commercial rendering application. We use *Marmoset Toolbag* [60], which performs real-time forward rendering that is highly parameterisable and allows the control of a wide-variety of reflectance models and modalities, such as subsurface scattering, specular reflection and high-frequency normals.

In order to extract the 3D representation from the UV domain we employ the inverse procedure explained in section 3.1 based on the UV pixel coordinates of each vertex of the 3D mesh. Fig. 6 shows the rendering results, under a single light source, when using the generated geometry (Fig. 6(a)) and the generated texture (Fig. 6(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. 6(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

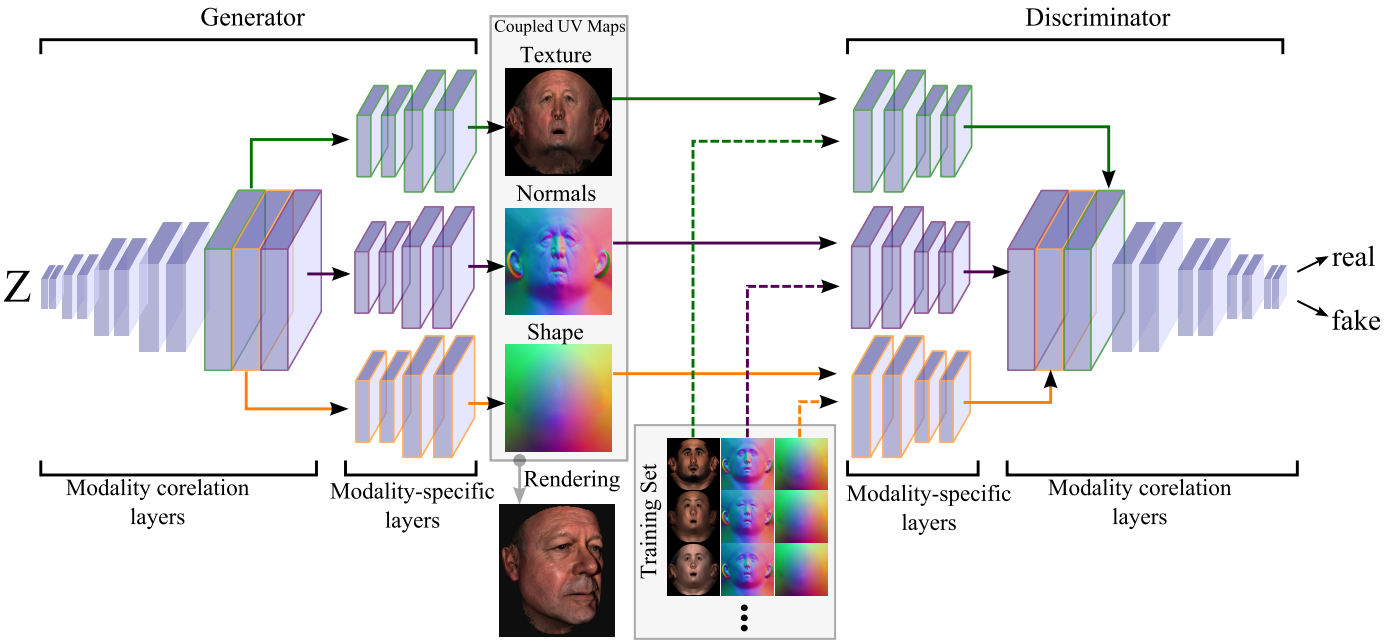


Fig. 3: An overview of the proposed trunk-branch network to generate multiple modalities to render a more photo-realistic face images. The network is designed to generate correlated texture, shape and normals UV maps of novel identities. The separation for different modalities allows branch-networks to specialize the characteristic of each one of the modalities while the trunk network maintain local correspondences among them.

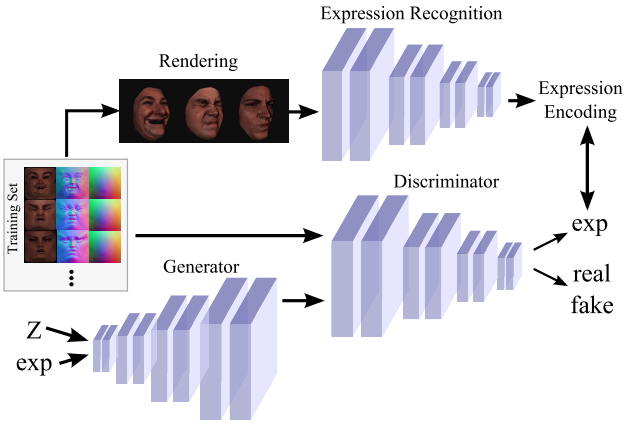


Fig. 4: Overview of expression-conditioned Trunk-Branch GAN. We annotate training dataset automatically by an expression recognition network and use output expression encodings as label. The generator network learns to couple those expression encodings to the texture, shape and normals UV maps. The generator and discriminator networks have the same architectures as Fig. 3 and abbreviated here.

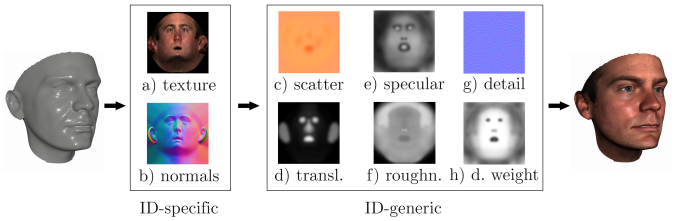


Fig. 5: Overview of the texture maps used for rendering. The generated id-specific texture (a) and normal maps (b) are mapped to the base geometry during the rendering process. Moreover, we enhance the rendered results by using identity-generic modalities, that describe in uv-space the scattering (c), translucency (d), specular intensity (e), roughness (f), detail normals (g) and their weights (h), and are derived from the training data.

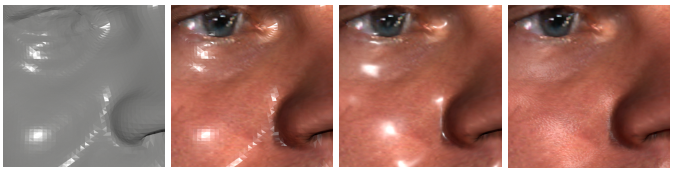
as a body tissue, it both reflects and absorbs light, thus exhibiting specular reflection, diffuse reflection and subsurface scattering.

Although we can add such modalities, to our multi-branch GAN with the availability of such data, we find that rendering can be still improved by adding some identity-generic maps for rendering. 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 (Fig. 5). *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 *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. 6(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



(a) shp (b) shp+tex (c) shp+tex+nor (d) photorealistic

Fig. 6: Zoom-in on rendering results when using (a) only the geometry, (b) adding the albedo texture, (c) adding the generated mesoscopic normals and (d) using identity-generic detail normal, specular albedo, roughness, scatter and translucency maps.

expressions. In our experiments, we set $L = 8$ and $d = 6$, meaning that there are 8 upsampling/downsampling layers in total, 6 of them in the trunk, 2 of them 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 1.25GB compared to the 6GB PCA model of the same resolution which contains $\%95$ of the total entropy.

In the following sections, we first visualize unwrapped UV representations of the generated modalities and their contributions to the final renderings on a number of generated faces. Next, we show the generalization ability of the identity and expression generators by means of number of attributes. 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 a number of face recognition experiments by using the generated face images as additional training data.

4.1 Qualitative Results

4.1.1 Combining coupled modalities:

Fig. 7 presents the generated shape, normals and texture maps by the proposed GAN and their additive contributions to the final renderings. As can be seen 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 artefacts of the generated geometry in Fig. 7(d) are compensated by the normals in Fig. 7(e).

4.1.2 Diversity:

Our model is well-generalized with different age, gender, ethnicity groups and many facial attributes. Although Fig. 8 shows diversity in some of those categories, the reader is encouraged to see identity variation throughout the paper. From looking at the varition demostrated, we can safely claim that our models are free of global or local mode collapse of which many GAN studies are struggling. Please refer to the supplementary video to enjoy more diversity.

4.1.3 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. 9. The reader is encouraged to see more expression generations in the supplementary video.

4.1.4 Interpolation between identities:

As shown in the supplementary video and in Fig. 13, our model can easily interpolate between any generation in a visually continuous set of identities which is another indication that the model is free from mode collapse. Interpolation is done by randomly generating two identities and generates faces by evenly spaced samples in latent space between the two.

4.1.5 Full head completion

We also extend our facial 3D meshes to full head representations by employing the framework proposed in [61]. We achieve this by regressing from a latent space that represents only the 3D face to the PCA latent space of the Combined Face and head model (CFHM) [61]. 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 CFHM model and project them to the CFHM subspace in order to the 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 [61], which works as a mapping from the latent space of the face shape to the full head representation. In Figure 14 we demonstrate the extended head representations of our approach in conjunction with the faces in Figure 13.

4.1.6 Comparison to decoupled modalities and PCA

Results in Fig. 10 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, shape and normals is an important component for realistic face synthesis. Also generations by PCA models are missing photorealism and high-fidelity 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 diverge identities. That is, we use the generated faces to train a face recognition network. To do so, we employ the most recent state-of-the-art face recognition method, ArcFace [62], and show that the proposed shape and texture generation model can boost the performance of pose-invariant face recognition in the wild.

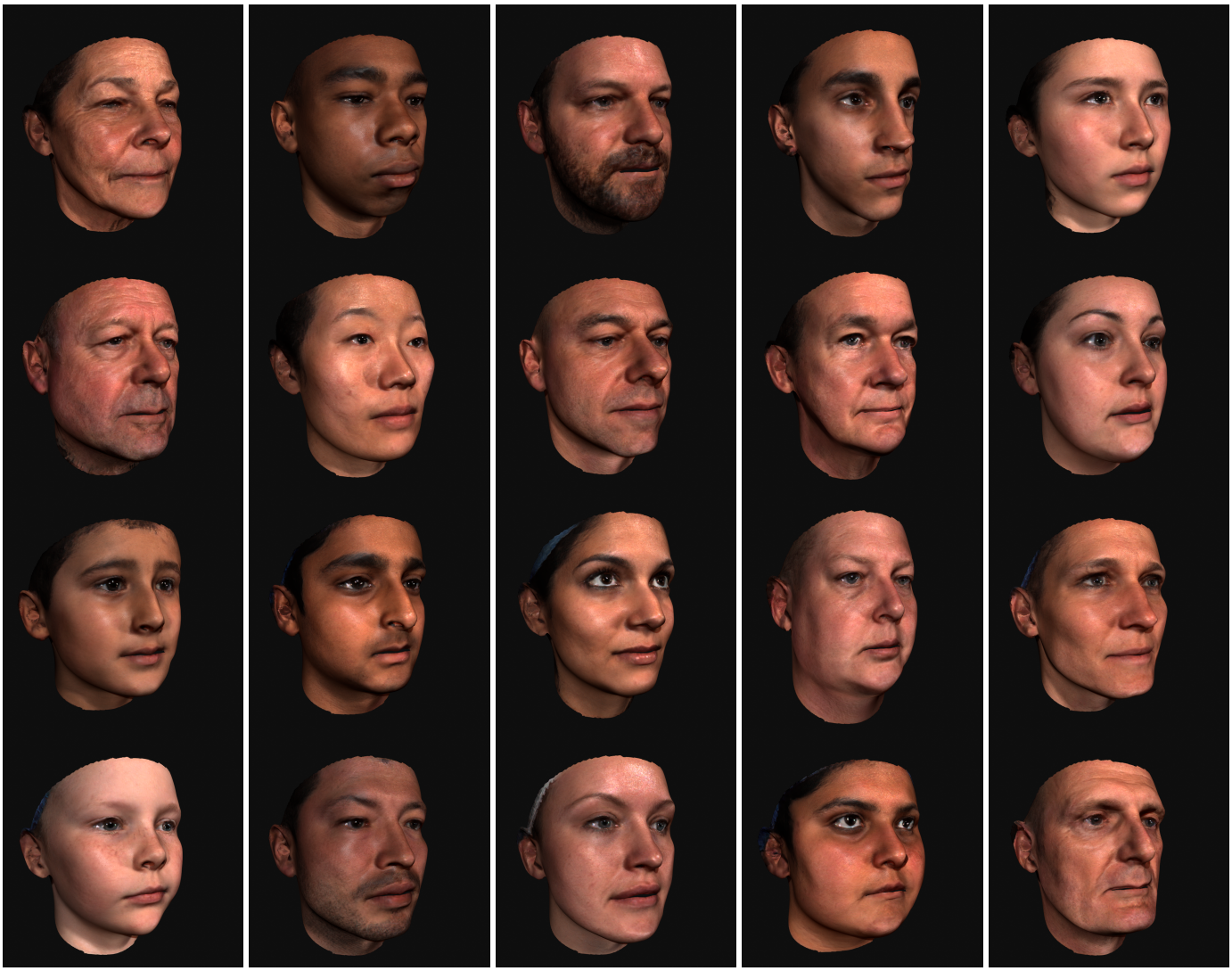
Training Data. We randomly synthesize 10K new identities from the proposed shape and texture generation model and render 50 images per identity with random camera and illumination parameters from the Gaussian distribution of the 300W-LP dataset [63]. For clearity, we call this dataset "Gen" in the rest of the text. Figure 11 illustrates 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 [64], which consists of 10,575 identities with 494,414 images.

Test Data. For evaluation, we employ Celebrities in Frontal Profile (CFP) [65] and Age Database (AgeDB) [66]. **CFP** [65] 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

(a) Shape UV (b) Normals UV (c) Texture UV (d) Shape Rendered (e) Shape+Normals (f) Shp.+Nor.+Tex.

Fig. 7: 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.

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** [66], [67] 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) [67]. In this paper,



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

Fig. 8: Variation of generated 3D faces by our model. Each column show diverse images of a different aspect.

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 [62] to generate the normalized face crops (112×112) by utilizing five facial points.

Training and testing Details. For the embedding networks, we employ the widely used ResNet architecture (ResNet50) [68]. After the last convolutional layer, we also use the BN-Dropout-FC-BN [62] structure to get the final 512-D embedding feature. For the hyper-parameter setting, we follow [62] to set the feature scale s to 64 and choose the angular margin m at 0.5. We set the batch size to 512 and train models by MXNet [69] on four NVIDIA Tesla P40 (24GB) GPUs. On the CASIA dataset, the learning rate starts from 0.1 and is divided by 10 at 20K, 28K iterations. The training process is finished at 32K iterations. On the combined dataset (CASIA and generation data), we divide the learning rate at 30K and 42K iterations and finish training process at 60K iterations. We set momentum at 0.9 and weight decay at $5e-4$. During testing, we only keep the feature embedding network without the fully connected layer (160MB) and extract the 512-D features (8.9 ms/face) for each normalized face. Note that, overlap 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 influence of generated data on pose-invariant face recognition. We take UV-GAN [45] 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. For our experimental settings, we use ([training data, network structure, loss]) to facilitate understanding. 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 [45]. 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 Figure 12, 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

(a) Happy (b) Sadness (c) Anger (d) Fear (e) Disgust (f) Surprise

Fig. 9: First and forth rows shows generations of different expression categories. The other rows show texture and normals maps used to generate the corresponding 3D faces. Please note how expressions are represented in the texture and normals space.

(a) (b) (c) (d)

Fig. 10: Comparison with seperate 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.3.5.

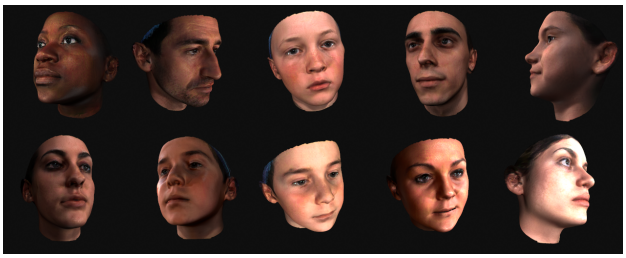


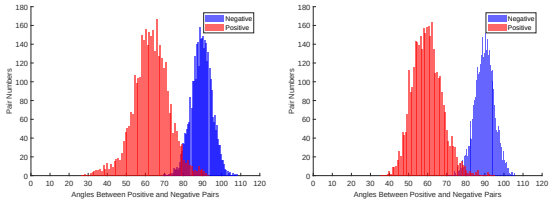
Fig. 11: Examples of generated data ("Gen") by the proposed method.

confirms that ArcFace can learn pose-invariant feature embedding

TABLE 1: Verification performance (%) of different models on LFW, CFP-FP and AgeDB-30.

Methods	CFP-FP	AgeDB-30
UVGAN [45]	94.05	94.18
CASIA, R50, ArcFace	95.56	95.15
CASIA+Gen, R50, ArcFace	**97.12**	**95.18**

from the generated data. In Table 2, we select some verification pairs from CFP-FP and calculate the angles 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.



(a) CASIA (b) CASIA+Gen

Fig. 12: Angle distributions of CFP-FP pairs in the 512-D feature space. Red area indicates positive pairs while blue indicates negative pairs. All angles between feature vectors are represented in degree.

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°

5 Conclusion

We presented the first neural 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. Furthermore, we propose a novel GAN model that can generate different expressions using as noise the embeddings of a facial expression recognition network. We demonstrate that randomly synthesized images of our unified generator shows strong relations between texture, shape and normals and that rendering with normals provides excellent shading and overall visual quality. Finally, in order to demonstrate that our methodology can manage to generate a diverse range of identities we have used a set of generated images to train a deep face recognition network.

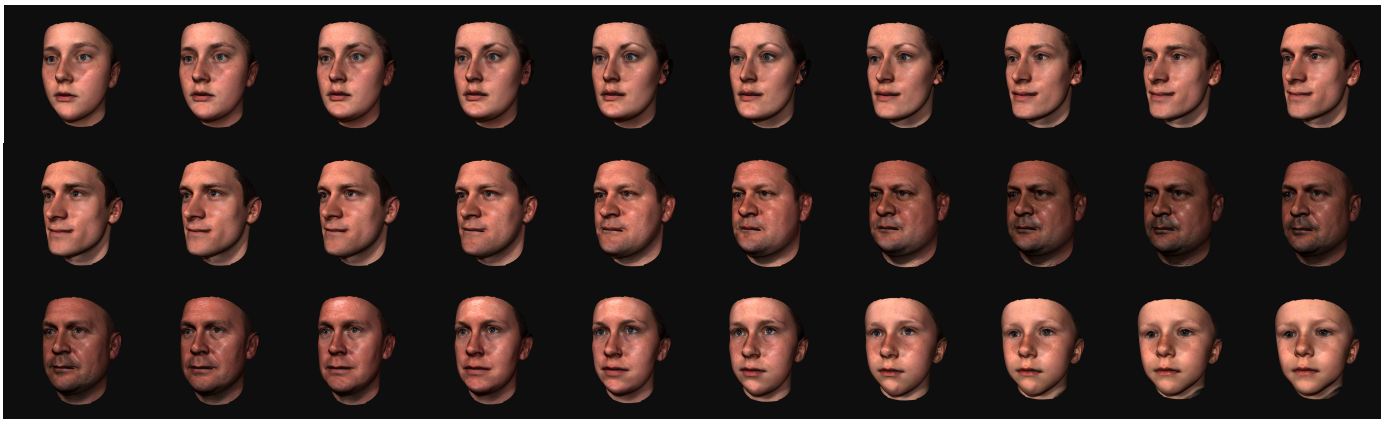


Fig. 13: The figure shows the interpolation between pair of identities. Smooth transition indicates generalization of our GAN model.

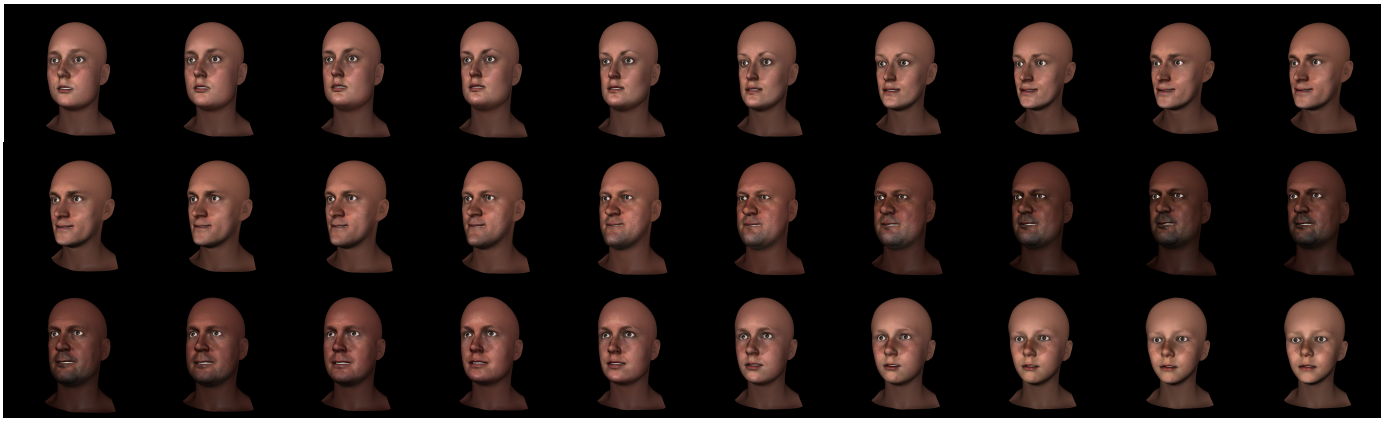


Fig. 14: Complete full head representations in association with the facial topology corresponding in Fig.13. Even in the full head topology our generation methodology ensures a smooth transition during interpolation.

ACKNOWLEDGMENTS

Baris Gecer is funded by the Turkish Ministry of National Education. Stefanos Zafeiriou acknowledges support by EPSRC Fellowship DEFORM (EP/S010203/1) and a Google Faculty Award.

REFERENCES

[1] B. Gecer, B. Bhattarai, J. Kittler, and T.-K. Kim, "Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model," *ECCV*, 2018. 1, 3

[2] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," *arXiv preprint arXiv:1902.05978*, 2019. 1, 2

[3] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li, "Avatar digitization from a single image for real-time rendering," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 195, 2017. 1

[4] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces." in *Siggraph*, vol. 99, no. 1999, 1999, pp. 187–194. 1, 2

[5] 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*, vol. 20, no. 3, pp. 413–425, 2013. 1

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 1, 2, 4

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb 1, 3, 4

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *arXiv preprint arXiv:1812.04948*, 2018. 1, 3, 4

[9] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018. 1, 3, 4

[10] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1576–1585. 1

[11] R. Slossberg, G. Shamai, and R. Kimmel, "High quality facial surface and texture synthesis via generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCVW)*, 2018, pp. 0–0. 1, 2

[12] S. Moschoglou, S. Ploumpis, M. Nicolaou, and S. Zafeiriou, "3dfacegan: Adversarial nets for 3d face representation, generation, and translation," *arXiv preprint arXiv:1905.00307*, 2019. 1, 2

[13] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3d faces using convolutional mesh autoencoders," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 704–720. 1, 2, 3

[14] T. Akimoto, Y. Suenaga, and R. S. Wallace, "Automatic creation of 3d facial models," *IEEE Computer Graphics and Applications*, vol. 13, no. 5, pp. 16–22, 1993. 2

[15] S. M. Platt and N. I. Badler, "Animating facial expressions," in *ACM SIGGRAPH computer graphics*, vol. 15, no. 3. ACM, 1981, pp. 245–252. 2

[16] D. DeCarlo, D. Metaxas, and M. Stone, "An anthropometric face model using variational techniques," in *SIGGRAPH*, vol. 98, 1998, pp. 67–74. 2

[17] A. Patel and W. A. Smith, "3d morphable face models revisited," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1327–1334. 2

[18] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer, "Review of statistical

shape spaces for 3d data with comparative analysis for human faces," *Computer Vision and Image Understanding*, vol. 128, pp. 1–17, 2014. 2

[19] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. Ieee, 2009, pp. 296–301. 2

[20] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 233–254, 2018. 2

[21] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas, "Expression flow for 3d-aware face component transfer," *ACM transactions on graphics (TOG)*, vol. 30, no. 4, p. 60, 2011. 2

[22] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 194, 2017. 2

[23] M. Breidt, H. H. Biilthoff, and C. Curio, "Robust semantic analysis by synthesis of 3d facial motion," in *Face and Gesture 2011*. IEEE, 2011, pp. 713–719. 2

[24] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3d face recognition with a morphable model," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2008, pp. 1–6. 2

[25] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," *Acm transactions on graphics (tog)*, vol. 29, no. 4, p. 32, 2010. 2

[26] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment." *ACM Trans. Graph.*, vol. 34, no. 6, pp. 183–1, 2015. 2

[27] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, p. 40, 2013. 2

[28] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001. 2

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 2, 4

[30] U. Mohammed, S. J. Prince, and J. Kautz, "Visio-lization: generating novel facial images," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, p. 57, 2009. 3

[31] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 3, 4

[32] S. Bazrafkan, H. Javidnia, and P. Corcoran, "Face synthesis with landmark points from generative adversarial networks and inverse latent space mapping," *arXiv preprint arXiv:1802.00390*, 2018. 3

[33] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6713–6722. 3

[34] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision*, pp. 1–13. 3

[35] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833. 3

[36] Q. Zhang, Z. Liu, G. Quo, D. Terzopoulos, and H.-Y. Shum, "Geometry-driven photorealistic facial expression synthesis," *IEEE Transactions on visualization and computer graphics*, vol. 12, no. 1, pp. 48–60, 2005. 3

[37] K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu, "A data-driven approach for facial expression synthesis in video," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 57–64. 3

[38] D. S. Trigueros, L. Meng, and M. Hartnett, "Generating photo-realistic training data to improve face recognition accuracy," *arXiv preprint arXiv:1811.00112*, 2018. 3

[39] L. Q. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE transactions on pattern analysis and machine intelligence*, 2018. 3, 4

[40] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8398–8406. 3

[41] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 821–830. 3

[42] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*. Springer, 2016, pp. 579–596. 3

[43] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Advances in Neural Information Processing Systems*, 2017, pp. 66–76. 3

[44] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3990–3999. 3

[45] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7093–7102. 3, 8, 9

[46] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin, "Making faces," in *ACM SIGGRAPH 2005 Courses*. ACM, 2005, p. 10. 3

[47] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 68, 2018. 3

[48] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, "Real-time facial animation with image-based dynamic avatars," *ACM Transactions on Graphics*, vol. 35, no. 4, 2016. 3

[49] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5543–5552. 3, 4

[50] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "Meshgan: Non-linear 3d morphable models of faces," *arXiv preprint arXiv:1903.10384*, 2019. 3

[51] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia, "Deformable shape completion with graph convolutional autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1886–1895. 3

[52] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007, pp. 1–8. 3

[53] M. De Smet and L. Van Gool, "Optimal regions for linear model-based 3d face reconstruction," in *Proceedings of the Asian Conference on Computer Vision*, 2010, pp. 276–289. 3

[54] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975. 4

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232. 4

[56] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690. 4

[57] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493. 4

[58] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems*, 2017, pp. 406–416. 4

[59] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777. 4

[60] Marmoset LLC, "Marmoset toolbag," 2019. [Online]. Available: http://www.marmoset.co/toolbag 4

[61] S. Ploumpis, H. Wang, N. Pears, W. A. Smith, and S. Zafeiriou, "Combining 3d morphable models: A large scale face-and-head model," *arXiv preprint arXiv:1903.03785*, 2019. 6

[62] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019. 6, 8

[63] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016. 6

[64] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv:1411.7923*, 2014. 6

[65] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *WACV*, 2016. 6

[66] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected in-the-wild age database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017. 6, 7

[67] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *CVPRW*, 2017. 7

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 8

[69] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv:1512.01274*, 2015. 8