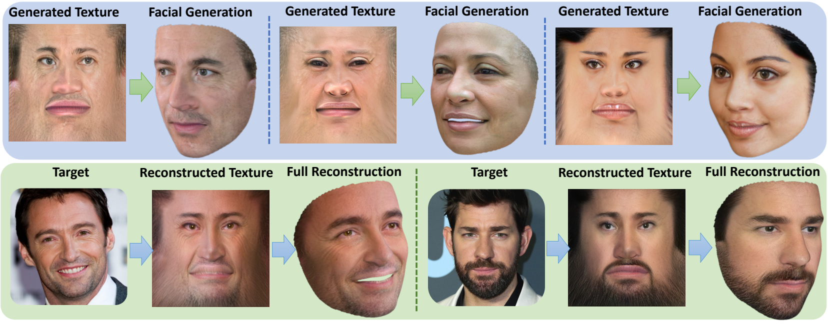
Unsupervised High-Fidelity Facial Texture Generation and Reconstruction

Ron Slossberg[1*], Ibrahim Jubran[2*], and Ron Kimmel[1]
ronslos@gmail.com, ibrahim.jub@gmail.com, ron@cs.technion.ac.il

[1] Technion Institute, Haifa, Israel
[2] University of Haifa, Haifa, Israel



Abstract. Many methods have been proposed over the years to tackle the task of facial 3D geometry and texture recovery from a single image. Such methods often fail to provide high-fidelity texture without relying on 3D facial scans during training. In contrast, the complementary task of 3D facial generation has not received as much attention. As opposed to the 2D texture domain, where GANs have proven to produce highly realistic facial images, the more challenging 3D domain has not yet caught up to the same levels of realism and diversity.

In this paper, we propose a novel unified pipeline for both tasks, generation of texture with coupled geometry, and reconstruction of high-fidelity texture. Our texture model is learned, in an unsupervised fashion, from natural images as opposed to scanned textures. To our knowledge, this is the first such unified framework independent of scanned textures.

Our novel training pipeline incorporates a pre-trained 2D facial generator coupled with a deep feature manipulation methodology. By applying our two-step geometry fitting process, we seamlessly integrate our modeled textures into synthetically generated background images forming a realistic composition of our textured model with background, hair, teeth, and body. This enables us to apply transfer learning from the 2D image domain, thus leveraging the high-quality results obtained in this domain.

We provide a comprehensive study on several recent methods comparing our model in generation and reconstruction tasks. As the extensive qualitative, as well as quantitative analysis, demonstrate, we achieve state-of-the-art results for both tasks.

[*] These authors contributed equally to this work

1 Introduction

Generation of 3D facial geometry and full texture, as well as their reconstruction from a single 2D image, are highly challenging and important tasks at the intersection of computer vision, graphics, and machine learning. These tasks arise within endless applications ranging from virtual reality to facial editing.

Our main motivation is that while 2D generation methods have been successful, it is difficult to carefully control attributes such as expression, pose and lighting within such image generators. At the other end, achieving similar results in the 3D domain is difficult due to lack of data and the requirement for generating corresponding geometries for each texture map. Our goal is to enable such control while maintaining the convenience of training on 2D images. In addition, we propose to construct a joint pipeline for both tasks for the sake of resource conservation as well as model standardization for applications where both generated and reconstructed faces are used.

At the heart of such generation and reconstruction methods lies a hidden common assumption that natural facial geometries and textures reside on a low-dimensional manifold. Following this assumption, the above tasks can be carried out within this simpler representation space, instead of the original high-dimensional space. The recovery of this manifold is termed *facial modeling* and the mathematical bridge between the high and low dimensional representations is termed a *facial model.* Many different types of facial models have been proposed over the years, including linear, non-linear, deep learning-based, hybrid, implicit, and dense landmark regression models, to name a few. While most models are geared towards reconstruction tasks, only a few models are successful at synthetically generating realistic samples due to the added complication of sampling the facial manifold. In addition, regardless of the models used, the facial generation process must account for the inter-dependency between geometry and texture, thus, producing compatible geometry-texture pairs in order to achieve realistic 3D facial generations. Ideally, when performed correctly, the newly generated faces will reside on the combined geometry-texture manifold.

In previous efforts, training a generative model for facial geometry and texture depended either on (i) 3D facial scans, via supervised learning, that yielded high-quality results, or (ii) on 2D facial images only, via unsupervised or semi-supervised learning, which produced lower-quality results; see overview in Section 2. Here we combine the best of both worlds, and provide an unsupervised training pipeline, independent of a dataset of 3D facial scans, producing state-of-the-art facial generation results. In addition to performing 3D facial texture generation, which is our main contribution, the proposed model can also be utilized for the task of full-texture recovery from a single 2D image for which we also demonstrate results on par with fully supervised methods. The proposed high-resolution model is achieved by incorporating a linear as well as a direct regression facial model, a pre-trained 2D generative model, a deep feature manipulation component, and a differentiable rendering layer, all integrated as building blocks for a novel unsupervised training pipeline.

2 Background and Related Efforts

Next, we review related efforts. Techniques incorporated in the proposed pipeline are described in detail. Table 1 summarizes the fundamental differences between our work and relevant prior works.

The 3D Morhpable Model (3DMM) [1] is arguably the most commonly used model both when generating or reconstructing facial geometries and textures; see [9]. The 3DMM model is obtained by semantically aligning facial scans to a template model comprised of n vertices and performing PCA [16] on the geometry, texture and expression vectors. The obtained k principal components for shape and expression $\mathbf{U}_s, \mathbf{U}_e \in \mathbb{R}^{3n \times k}$ and mean shape $\mathbf{M} \in \mathbb{R}^{3n}$ comprise the 3DMM geometry model. Given a set of shape and expression parameters $(\mathbf{p}_s, \mathbf{p}_e \in \mathbb{R}^k)$, the facial geometry is constructed as $\mathbf{S} = \mathbf{M} + \mathbf{U}_s \cdot \mathbf{p}_s + \mathbf{U}_e \cdot \mathbf{p}_e$. Texture modeling and formation are produced per-vertex in a similar manner. Many improvements were suggested, for example, [3,22,2], who improve the data acquisition and registration processes. However, due to their linear nature, such models usually produce unrealistic samples [35].

3DMM fitting. Given a 2D face image and the 3DMM geometry and expression basis, the goal of *3DMM fitting* is to recover the 3DMM geometry and expression coefficients as well as a 3D rigid transformation. Numerous approaches have been suggested for tackling this problem, ranging from optimization-based methods [1,12], to one-shot deep learning pipelines originated in [31,45], and followed up in [39,14,7,15] to name a few. In this work, we utilize the model introduced in [7], due to its high precision in estimating model parameters, as well as the available code and pre-trained model; see Section 3.

Non-linear and hybrid model fitting. Recent efforts have built upon classical 3DMMs, proposing both hybrid [32,38,33,35,36,12,4,34] and non-linear models [42,44]. These deep network-based methods may also incorporate linear components. Some models are presented only in the context of monocular geometry and texture recovery while others are also utilized for synthesis.

Dense landmark regression. In [20], a regression network is trained to predict a dense collection of landmarks directly on a given facial 2D image. These landmarks represent the projected vertex locations of a 3D canonical facial model. This method achieves better alignment relative to the target image and is not constrained to the limitations of the linear model. However, the landmark based facial representation presents low-detail geometry due to the limited number of recovered landmarks. We therefore propose a two-step fitting scheme combining both landmark regression as well as 3DMM geometry reconstruction in order to gain the benefits of both regimes; see details in Section 3.3.

Realistic 2D face generation. In a long line of efforts culminating in [17], various models have been proposed for the task of 2D face image generation. Such models are capable of generating highly realistic 2D facial images as well as project real 2D facial images onto the model's latent manifold. As we aim to mitigate the need for 3D scans of facial textures, we heavily rely on well-established 2D facial generative models as the basis for the proposed pipeline. Throughout the proposed pipeline, we utilize the framework and pre-trained

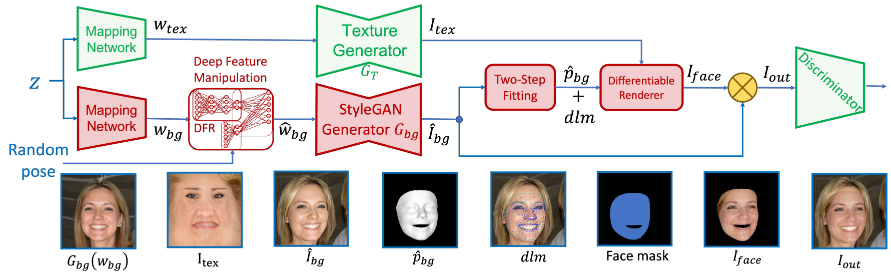


Fig. 1. Our training pipeline. A vector $\mathbf{z} \in \mathbb{R}^{512}$ of Gaussian random noise is plugged into two identical mapping networks [19], producing two latent vectors $\mathbf{w}_{tex}, \mathbf{w}_{bg} \in \mathbb{R}^{18\times512}$, respectively. The facial image obtained by feeding $\mathbf{w}_{bg}$ into *StyleGAN* is illustrated on the lower left. The vector $\mathbf{w}_{bg}$ is plugged into a deep feature manipulation network [37] called *StyleRig* to obtain the (manipulated) latent vector $\hat{\mathbf{w}}_{bg} \in \mathbb{R}^{18\times512}$ which encodes the same facial information as $\mathbf{w}_{bg}$ (*e.g.* facial expression, identity, lighting, etc.) but with a modified facial orientation. We then feed $\mathbf{w}_{tex}$ and $\hat{\mathbf{w}}_{bg}$ into our texture and pre-trained *StyleGAN* generators G_T and G_{bg}, outputting a texture image I_{tex} and a 2D facial image $\hat{I}_{bg}$ respectively. Then, we apply a two-step fitting approach to recover 3DMM parameters ($\hat{\mathbf{p}}_{bg}$) [7] as well as a dense landmarks mask (dlm) [20] that best fit $\hat{I}_{bg}$; see Section 3.3. We then use $\hat{\mathbf{p}}_{bg}$ to render the texture I_{tex} into a 2D facial image I_{face}, and perform a masking operation according to the face mask extracted from the dlm. The masked facial (foreground) image I_{face} and the (background) image $\hat{I}_{bg}$ are then composed together to form I_{out}. Finally, I_{out} is fed into a pre-trained discriminator which is further trained. Trainable and frozen models are depicted in green and red respectively.

model weights provided by the seminal papers of Karras *et al*. [18,17], which are regarded the golden standard for this task; see Section 3.

Manipulating facial properties via deep feature mapping. Most synthesis methods described above, specifically [18], learn to map an input random noise vector, through some latent representation, into a realistic 2D facial image. Following this popular approach, a variety of papers have emerged which learn to manipulate this latent vector to change some desired facial properties in the output 2D image. Such manipulation can be either statistics-based [5] or, more often, learning based [37]; see [40] with references therein.

As discussed in detail in Section 3, our pipeline makes use of a 2D facial image generator to compensate for the lack of 3D facial scans. However, it is infeasible to compensate for 3D geometry and full facial texture using only *uncontrolled randomly generated* 2D facial images. We thus utilize a method providing control over the pose of generated images and show that the *controlled* 2D images indeed suffice for full-texture learning. To this end, we utilize the method of [37] for deep feature manipulation; see Section 3.

Generation. While most prior efforts have focused on 3D reconstruction from a 2D image, few methods have been proposed for the generation of random but realistic facial models. In [25], a GAN-based approach was also proposed for improving facial recognition models via synthetic augmentation; however, their pipeline focuses more on controlling model parameters intending to supplement the training data. This, as opposed to the realistic generation of completely random faces, leads to a less desirable outcome in terms of realism and resolution. Hence, the results are not visually pleasing; see Fig. 4. In [35,34,12], 3DMMs combined with generative models were used for either generation or reconstruction of realistic textures. However, these methods rely on proprietary high-quality facial scans during training, obtained by specialized facial scanners. This makes these results difficult to reproduce. Moreover, such scanned data is far less diverse than abundant facial 2D images in common datasets.

Reconstruction. Many methods have been previously suggested for 3D face reconstruction from a given 2D image. In [31,32,33,8], a mapping from 2D images to a 3D geometric representation is learned based on synthetic data. In [12,13], real facial textures were utilized, to obtain, in a supervised manner, a realistic reconstruction. However, acquiring such textures requires laborious and expensive 3D scanning, hence, impractical to scale to large numbers. In this paper, we provide an unsupervised alternative that requires only the freely available geometric models, and does not directly require 3D scans, and achieves either comparable or higher quality reconstructions. A pipeline for completion of a facial texture containing large holes was suggested in [6], however, they also rely on scanned textures as training data. A one-shot learning approach was proposed in [11] which applies an iterative and slow optimization process to complete a facial texture.

Two additional methods that do not rely on 3D scanned data were proposed by Lin *et al.* and Kim *et al.* [21,23]. In Kim *et al.* [21] an unsupervised model for partial texture completion is trained by combining a global and a local patch discriminator to the full rendering as well as the uv mapped texture. The uv maps in this work are of dimension 512×512, a quarter of the resolution presented by our maps. Lin *et al.* [23] takes a different approach to the completion task by utilizing Graph Convolutional Networks. However, by not basing their pipelines on a 2D image generator which can produce controlled 2D images (*e.g. StyleGAN* combined with a model as [37]), their method is not intended for the task of generation of expressive 3D models and does not account for the coupling of geometry and texture. See detailed comparison with [21,23] in Section 4.

Method	Unsupervised Training	High-Fidelity Output	Supports Generation	Supports Reconstruction
Deng *et al.* [7]	✓			✓
Lin *et al.* [23]	✓	✓		✓
Kim *et al.* [21]	✓	✓		✓
Deng *et al.* [6]		✓		✓
Gecer *et al.* [12]		✓		✓
Shamai *et al.* [35,34]		✓	✓	
Marriott *et al.* [25]	✓		✓	
Ours	✓	✓	✓	✓

Table 1. Comparison to prior art.

2.1 Our contribution

The main contributions of our method are the following:

(i) We provide the first unsupervised high-fidelity generation pipeline capable of producing realistic textures coupled with corresponding geometries; See Table 1 and Section 3. This is achieved by a novel training pipeline which successfully decouples intrinsic texture features related to the person's identity, from extrinsic properties such as pose and Lambertian illumination.

(ii) In addition to texture generation, we utilize the very same model for the task of texture recovery from a single image, successfully reconstructing frontal as well as peripheral facial details; see Section 3.2.

(iii) We present state-of-the-art results in both model generation as well as full texture recovery. We support this claim via both qualitative as well as quantitative results and comparisons; see Section 4. We prepared an additional demonstration video with further results; see supplementary material.

(iv) Our results are fully reproducible as only freely available datasets and models are required during training and inference. In addition, we provide all our trained model weights for both generation and reconstruction tasks [26].

3 Unsupervised Learning of Facial Textures and Geometries

In this section, we detail our unsupervised pipeline for generation and reconstruction of full facial textures and coupled geometries. While the proposed pipeline generates and recovers both texture as well as corresponding geometry, we rely on existing methods and models for geometric recovery and generation, and focus our attention mainly on high-quality texture modeling. We are guided by the notion that the main effect on the perception of model realism stems from high-resolution texture rather than highly detailed geometry. This idea was also noted *e.g.* by [35]. Nevertheless, recovery of highly detailed geometry is still an important research topic with many successful efforts such as [32,33,41,4].

An overview of the suggested training and inference pipelines is depicted in Figs. 1 and 2 respectively. The proposed approach to unsupervised learning of facial textures utilizes an adversarial loss to train a texture generator, G_T, while harnessing a pre-trained 2D facial image generator, G_{bg}, in the following fashion. We start by generating, via G_{bg}, a 2D facial image which we term a *background image*. We then fit a corresponding geometry to the background image using a two-step geometry recovery process utilizing [7,20]; see Section 3.3. We proceed to generate a facial texture I_{tex} via our trainable texture generator G_T. The generated synthetic texture is stored as a 2D image coupled with a canonical UV parametrization relating between image locations to the vertices of the 3D facial model. We base our UV unwraping on Floater [10]. The model fitted to the background image enables the seamless mapping of the synthetic texture image I_{tex} into the background image as depicted in Fig. 1.

The texture generator is trained within a GAN framework for which a discriminator model is trained to differentiate between blended and real images

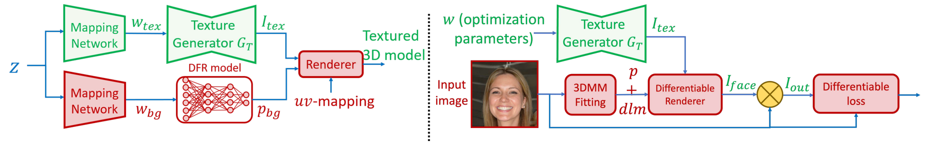


Fig. 2. Our inference pipelines. During inference, we drop some components related to the training pipeline (see Fig. 1). **(Left) Generation:** As before, a single latent vector $\mathbf{z}$ is used to generate $\mathbf{w}_{tex}$ and $\mathbf{w}_{bg}$ via two mapping networks. The latent vector $\mathbf{w}_{bg}$ is used to generate 3DMM geometry parameters $\mathbf{p}_{bg}$ via the trained DFR model while $\mathbf{w}_{tex}$ is introduced to the trained texture generator yielding the corresponding texture image I_{tex}. The parameters $\mathbf{p}_{bg}$ are used, along with our canonical UV parametrization, to generate the 3DMM geometry which we render using I_{tex} as the mesh texture. **(Right) Reconstruction:** A given input image I is first plugged into our two-step fitting model producing its 3DMM parameters $\mathbf{p}$ and a dense landmark mask dlm. A latent vector $\mathbf{w}$ containing our optimization parameters is then inserted into our trained texture generator producing a texture image I_{tex}. Using a differentiable renderer, $\mathbf{p}$ and I_{tex} are rendered into a 2D face image I_{face}, which is blended with I according to the dlm to produce our output I_{out}. Finally, a VGG loss similar to [17] is evaluated between I and I_{out}.

and thus continuously improves the generator quality. In order to generate high-resolution facial textures from all viewing angles, it is crucial to control various properties within the images generated by G_{bg}. For example, we require that each generated identity appears under a range of poses. We therefore utilize a deep feature manipulation component, as in [37], that encodes the desired properties within the input of G_{bg}. In addition, in order to disentangle between the albedo and shading components of the texture, we estimate the Lambertian lighting conditions in G_{bg} and apply them to our texture within the rendering process. Section 3.2 further elaborates on these components.

Learning from 3D facial scans. Prior efforts approached the task of training facial texture models by relying on difficult-to-obtain 3D scans. For example, in [35,12], high-resolution scans obtained by a 3DMD scanner are geometrically aligned and mapped to a canonical 2D domain. The mapped textures are used as training data for a GAN which is tasked to generate new and realistic ones. This methodology suffers from two main drawbacks: (i) The 3D scans are not easily obtained or freely distributed, thus posing a significant barrier in reproducing such models. (ii) High-quality 3D scanners are expensive and cumbersome, limiting the ability to collect data. Hence, even when available, such datasets are comprised of at most a few thousand subjects, which can not encompass the huge variety of human faces. We mitigate the above issues by eliminating the dependency on scans and replacing them with widely available 2D facial images and freely distributed geometric models, thus, providing a more accessible method and producing a more diverse texture model.

Replacing 3D facial scans with 2D facial images. Replacing 3D facial scans with prevalent 2D facial images is commonly achieved by utilizing a differentiable rendering layer. The rendering of 3D textured models into 2D images enables the incorporation of 2D image-related architectures and losses. This process also requires a 3D mesh, usually represented by a pair (V, Tri) of vertex coordinates $V \in \mathbb{R}^{N_v \times 3}$ and triangulation $Tri \in \mathbb{R}^{N_f \times 3}$, as well as a *uv* parametrization $\phi : \{1, \cdots, N_v\} \to [0, 1] \times [0, 1]$ that maps every vertex to coordinates on the canonical plane. To obtain the desired facial rendering I_{face}, the vertex coordinates are first projected onto the 2D camera plane and the final pixel colors are determined by a rasterization process mapping the facial texture onto the projected mesh according to the predetermined UV parametrization.

Using this methodology, we can transform our training losses from the 3D to the 2D domain. We can thus utilize the vast corpus of prior art regarding 2D images, including pre-trained models as well as large, high resolution, and freely available datasets; see Section 3.1.

Having established the above, the question remains how to obtain synthetic facial renderings which are indistinguishable from real facial images, considering that the rendered images lack hair, ears, inside of the mouth, background, etc. Possible solutions include segmenting-out the background in the real image, or adding a synthetic background to the rendered (synthetic) facial image. The former can be achieved via image segmentation or 3D model fitting, both of which produce sub-optimal results that are easily distinguishable from the synthetic image, due to artifacts at the face boundary. We therefore choose the latter option and propose to generate an additional 2D facial image I_{bg}, *e.g.* using *StyleGAN*, and utilize I_{bg} as the background to our (foreground) rendered image I_{face}. This is achieved by first fitting a geometric model to I_{bg} (see Section 3.3), which serves as the 3D mesh required for rendering I_{tex} into a 2D image I_{face}, as previously detailed. This process embeds our synthetic facial texture image I_{tex} into I_{bg}, enforcing the facial texture to be generated in a way that realistically blends with the surrounding parts in I_{bg}, like hair and ears; see Fig. 1.

3.1 Transfer learning

The process described above results in a 2D facial image, enabling the use of standard 2D image losses. As common in generative models, we use an adversarial loss to discriminate between real and fake images. Fortunately, many such pre-trained GANs are available for the task of 2D facial image generation [17]. We base our mapping network, texture generator, and discriminator, on the architecture proposed in *StyleGAN2* [19]. As facial textures are closely related to 2D facial images, we initialize the above models with the pre-trained *StyleGAN2* weights. This transfer learning approach has dramatically reduced our pipeline training time and improves texture quality, as was also reported by [17].

3.2 Pose and illumination invariant textures

As detailed above, the proposed unsupervised approach relies on rendering 2D images from the generated textures. However, this approach alone, has two inherent problem: **(i)** Since every input vector z corresponds to a facial image in a specific known pose, the generator can leverage this correlation and generate high-resolution details only in the visible regions with no penalty on occluded regions within the rendered 2D image. We propose to mitigate this issue by introducing random facial rotations during training via *deep feature manipulation*, as detailed below. **(ii)** Without properly addressing scene illumination, the generator will incorporate the lighting effects into the generated textures; see Fig. 5. We thus aim to decouple the albedo from the illumination effects, enabling post-relighting of the texture. To this end, we relight the models during training using the lighting parameters recovered by [7], forcing the generator to produce textures without baked-in lighting effects. Here, we assume a simplified Lambertian lighting model and do not consider reflective effects.

Deep feature pose manipulation. In order to overcome the orientation-decoupling problem, we manipulate the latent vector w_{bg}, related to the background image I_{bg} enforcing the generation of faces in a variety of orientations. This successfully decouples the pose from the input vector z, thus encouraging our texture generator to produce full high resolution texture from all viewing angles. We adopt the deep feature manipulation methodology proposed in [37].

The manipulation model, termed *StyleRig*, is comprised of two parts. A Differentiable Face Reconstruction Network, or DFR model, which takes as input the latent vector $\mathbf{w}$ and produces estimated 3DMM parameters $\mathbf{p} = DFR(\mathbf{w})$ which include $(\mathbf{p_s}, \mathbf{p_e}, \mathbf{p_t}, \gamma, \mathbf{R}, \mathbf{t})$, shape, expression, texture and lighting, rotation and translation parameters respectively. We train our model utilizing the highly versatile 3DMM model generated by [3].

A second network termed *StyleRig* takes as input a latent vector $\mathbf{w}$ and a set of parameters $\mathbf{p}$ and outputs a modified latent parameter vector $\hat{\mathbf{w}}$, where ideally the image $I = G_{bg}(\hat{\mathbf{w}})$ portrays the face $G_{StyleGan}(\mathbf{w})$ produced by $\mathbf{w}$ but modified to fit the parameters $\mathbf{p}$. In order to produce a rotated version of I_{bg} we first modify the rotation parameters of $\mathbf{p}_{bg} := DFR(\mathbf{w}_{bg})$ to derive $\hat{\mathbf{p}}_{bg}$ and then apply $\hat{\mathbf{w}}_{bg} = \mathit{StyleRig}(\hat{\mathbf{p}}_{bg}, \mathbf{w}_{bg})$. The image $\hat{I}_{bg} = G_{bg}(\hat{\mathbf{w}}_{bg})$ contains a rotated version of the same person as in I_{bg}. We then generate a texture image using the latent vector $\mathbf{w}_{tex}$, regardless of the rotation angles which were modified in $\mathbf{w}_{bg}$. This yields the desired pose-invariance within the texture generator; see Fig. 1

The same DFR model used above will later also be utilized during inference in order to recover corresponding geometries for our generated textures; see Section 3.3 and Fig. 2. This allows us to efficiently generate corresponding geometries directly from latent vectors without requiring the trained *StyleGAN* generator during inference.

Training for re-illumination. To generate textures without Lambertian illumination effects, we first estimate the background scene lighting and relight the texture during training. Assuming a simplified Lambertian reflectance model, we estimate the parameters $\gamma \in \mathbb{R}^{3\times 9}$ from I_{bg}, as coefficients of 9 Spherical Harmonics (SH) basis functions [28,29] for R,G and B illumination bands, and relight the rendered image I_{face} under the recovered illumination. The coefficients γ with the computed vertex normals $\{\mathbf{n}_i\}$ and SH functions Φ produce the per-vertex lighting value $\mathbf{C}(\mathbf{n}_i|\gamma)_l = \sum_{b=1}^{9} \gamma_{l,b}\Phi_b(\mathbf{n}_i)$. We perform two rendering passes, one for the illumination component and another for the albedo. The final illuminated rendering is obtained by pixelwise multiplication of the two rendering results.

$$I_{face} = \mathcal{R}(\mathbf{S}(p_s, p_e), G_T(w_{tex})) \cdot \mathcal{R}(\mathbf{S}(p_s, p_e), \mathbf{C}(\mathbf{n}_i|\gamma)),$$

where $\mathcal{R}(G, T)$ signifies the rendering operator applied to a geometry G and a texture T, $\mathbf{p}_s, \mathbf{p}_e$ are respectively shape and expression parameters recovered from I_{bg}, and $\mathbf{w}_{tex}$ is the input latent vector for the texture generator.

This process results in the texture generator producing textures with no baked-in Lambertian lighting effects, so that the re-illuminated texture via γ would match the lighting present in I_{bg} and seem realistic to the discriminator.

3.3 Recovering corresponding geometry via two-step fitting

In order to facilitate the realistic incorporation of foreground rendering and background image we propose a two step geometry fitting approach. We observed that, in general, geometry reconstruction methods tend to exhibit a trade-off between geometry realism and precise image alignment. For example, while 3DMM based fitting methods produce high-resolution facial meshes, the mesh alignment relative to the target image is imperfect. In contrast, landmark regression based methodologies are precisely aligned to the target facial image but produce a very sparse geometry reconstruction based on the landmark arrangement. To achieve a realistic blending between the foreground rendering I_{face} and synthetic background image I_{bg}, both accuracy as well as high-resolution geometry are crucial.

To this end, we propose a two step fitting scheme comprising of the 3DMM recovery proposed by [7] followed by a dense landmark regression model [20], gaining the benefits of both. This is achieved by extracting the geometry parameters $\hat{\mathbf{p}}_{bg}$ from the former while utilizing the boundary mask *dlm* from the latter. While [7] provides a good high-resolution fitting which can be realistically rendered, we use the dense landmark mask (*dlm*) from [20] in order to perform foreground blending with high precision. By adopting the two-step approach we harness the strengths from both fitting techniques enabling us to perform accurate and realistic blending. In Section 4.3 we provide an ablation study comparing the naive one-step 3DMM-only approach to our proposed two-step fitting approach. Indeed we observe that the two-step approach helps mitigate unwanted misalignment artifacts, especially in the mouth region.

3.4 Unsupervised training

The proposed pipeline above generates full facial textures along with corresponding geometries, and, using a differentiable renderer, synthesizes a 2D facial image. Adhearing to the GAN framework, the composed facial image along with a real 2D facial image are fed into a discriminator network tasked to differentiate between the real and the fake samples. Such 2D real images are widespread and can be taken from any dataset of facial images, for example, [18].

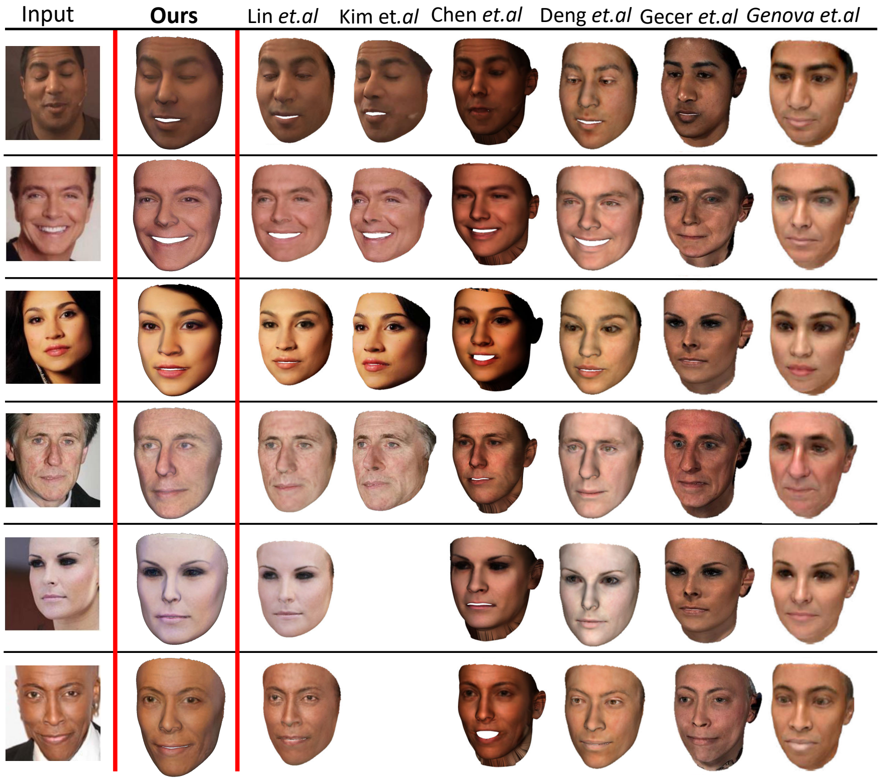


Fig. 3. Qualitative Reconstruction Comparison Results: We present texture reconstruction results on the MOFA test-set [39] compared to previous methods by [23,21,4,7,12,14], respectively. This figure is best viewed when zoomed in.

4 Experimental results

We compare the proposed approach to several state-of-the-art 3D generation and texture reconstruction methods. We provide quantitative and qualitative evidence that our model performance is on par with and often outperforms previous methods, both supervised and unsupervised by scanned textures (see Table 1), in terms of texture reconstruction quality, realism, and details. Our supplementary material contains additional reconstruction results for extreme side views, as well as a demonstration video presenting more viewing angles for our output results. Our code and pre-trained models are available [3].

Implementation details. We implemented our pipeline in Python using Pytorch [27] and Pytorch3D [30], and trained it on 4 RTX 3090 GPUs on the FFHQ dataset [18] consisting of 70k facial images. We initialized our models

[3] Link for our open-source code on Github: https://github.com/ronslos/Unsupervised-High-Fidelity-Facial-Texture-Generation-and-Reconstruction

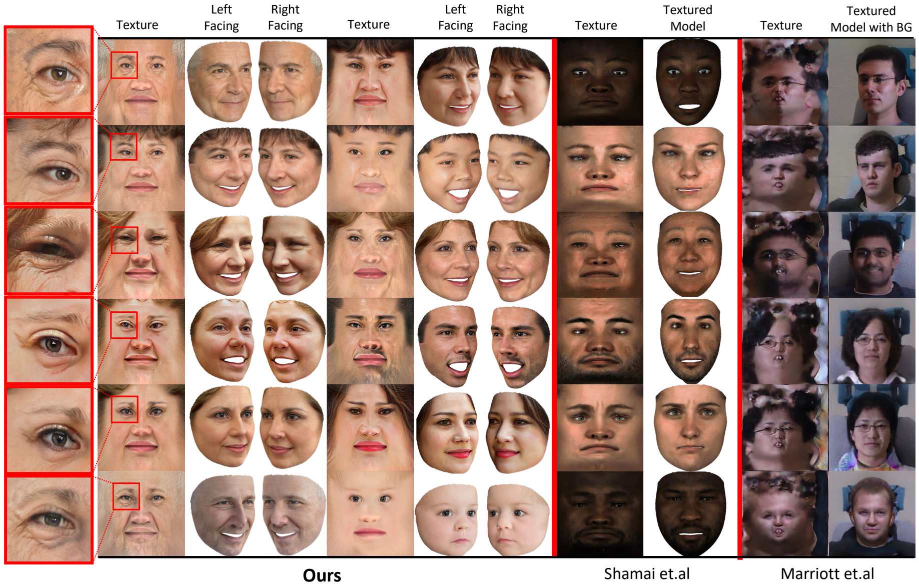


Fig. 4. Facial Synthesis: We visually compare our output textures and rendered textured geometries to: Shamai *et al.* [34] and Mariott *et al.* [25]. Our high resolution textures provide highly realistic faces spanning a wide variety of ages, ethnicity and appearance. The leftmost column provides a zoomed-in crop, highlighting the high resolution details. The proposed method presents finer details and realism as compared to both previous methods, even though [34] is supervised by scanned textures.

from the pre-trained weights of *StyleGAN* [19], using default parameters and losses, we train for 3 epochs; see Section 3.1. The 2D images as well as the generated textures are of size 1024 × 1024.

4.1 Face generation

We randomly generated textures and corresponding geometries via the proposed inference pipeline; see Section 3 and Fig. 2. We present the texture images with zoomed-in areas to highlight the high level of detail and realism. We compare our results to the supervised model of [34] and the unsupervised model from [25]; see Fig. 4. See supplementary material for additional results.

4.2 Facial texture reconstruction

Fig. 3 presents a qualitative comparison between our texture reconstruction pipeline from Fig. 2 to several state-of-the-art prior works [23,4,7,12,14]. The comparison demonstrates that our model can reproduce challenging textures *e.g.* difficult lighting conditions, makeup, and extreme expressions and compares favorably to previous approaches, including methods based on supervised training

Metric	[7]	[23]	[6]	**Ours**
L_1 distance ↓	0.052	0.034	/	**0.0244**
PSNR ↑	26.58	29.69	22.9∼26.5	**32.889**
SSIM ↑	0.826	0.894	0.887∼0.898	**0.972**
LightCNN [43] ↑	0.724	0.900	/	**0.96**

Table 2. Quantitative Evaluation: We evaluate reprojected reconstruction similarity on the CelebA[24] test-set, containing nearly 20k images.

from 3D scans. Note, that we utilize [7] for geometry recovery and thus focus our comparison on texture recovery only; see Section 3. Additional reconstruction results produced from high-resolution images are depicted in Fig. 1 and the supplementary material, which also presents reconstructions from side views. The results demonstrate that our model is capable of high-resolution texture recovery when presented with high-quality input images. We note that our proposed texture recovery method consists of a weight regularization term balancing between texture fidelity and realism. This is set manually to a constant desirable value throughout all our experiments. See supplementary for details alongside additional results obtained for varying regularization values.

4.3 Ablation study

In Fig. 5 we present an ablation study, where the full proposed model is shown to produce more realistic results compared to its variants with missing components. Additional ablation results are placed in the supplementary material. This suggests that each of our pipeline components is crucial for producing satisfactory output results. We show that: (i) model rotations during training are crucial for generating high details on the peripheral areas of the texture; see Section 3.2, (ii) the two-step fitting eliminates the unwanted teeth artifacts; see Section 3.3, and (iii) model illumination during training successfully disentangles albedo from Lambertian shading effects, producing models that can be realistically integrated into scenes with varying lighting conditions; see Section 3.2.

Training stability. The ablation study depicts three different training pipelines with parts of the original pipeline missing. However, the model still converges very similarly, maintains identities and only differs by the manner expected by the removal of each block. This demonstrated that our training is robust to modification of the pipeline.

4.4 Quantitative results

Table 2 presents a quantitative study for the task of texture reconstruction, using the CelebA [24] test-set. Our method achieves better scores in all tested metrics compared to previous state-of-the-art methods [7,23,6]. In contrast to [23], we do not omit problematic areas by semantically masking difficult regions.

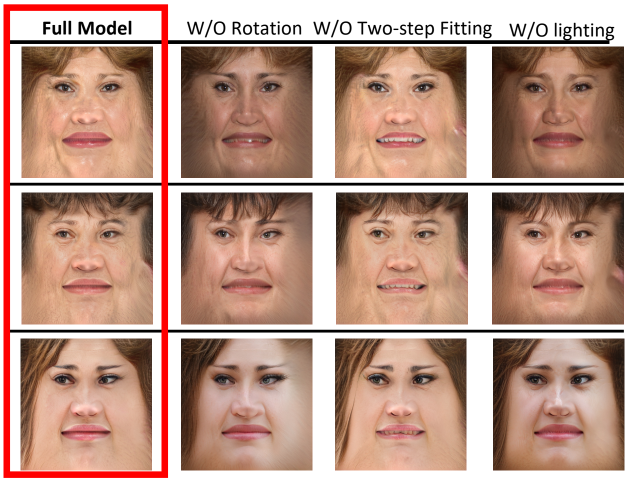


Fig. 5. Ablation Study. Left to right: (i) full model, (ii) without applying our deep feature manipulation component (see Section 3.2), (ii) without the two-step fitting, *i.e.* using only the 3DMM during the geometry fitting process, without the facial masking step (see Section 3.3), and (iii) without relighting the model (see Section 3.2). This leads to poor details in the texture periphery, unwanted teeth artifact, and baked-in Lambertian lighting effects, respectively; see Section 4.3.

5 Discussion, Limitations, and Future Work

We introduced a novel unsupervised pipeline for generation as well as reconstruction of high resolution realistic facial textures. Our pipeline matches the geometry and texture via a single unified random input vector **z**, and combines common pre-existing building blocks, in a non-trivial manner, with new novel ideas, to achieve SOTA resuls. Those ideas include the incorporation of a background image during training, the decoupling of pose from texture by feature vector manipulation, the ability to generate coupled geometry and texture at inference directly from random features, and our two-step fitting approach.

Our experiments demonstrate that we surpass prior art in realism and quality, in both tasks, including models supervised by scanned facial textures.

Due to the presence of subjects wearing glasses within the FFHQ dataset used for training, in some cases, our output texture might contain glasses; See supplementary material. This can be mitigated in future work by using latent feature manipulation.In addition, while Lambertian lighting is disentangled within our pipeline, we do not tackle the challenging problem of specular disentanglement. Moreover, we noticed that during reconstruction, on occasion, the eye color is not consistent with the input image. In future work it is possible to explore regional weighting of the reconstruction loss in order to better control reconstruction trade-offs. Lastly, we did not utilize non-linear geometric representations as we note that high-resolution texture is the most crucial component in the quest for realistic facial generation.

References

1. 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. pp. 187–194 (1999) 3
2. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3d morphable models. International Journal of Computer Vision **126**(2), 233–254 (2018) 3
3. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5543–5552 (2016) 3, 9
4. Chen, A., Chen, Z., Zhang, G., Mitchell, K., Yu, J.: Photo-realistic facial details synthesis from single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9429–9439 (2019) 3, 6, 11, 12
5. Chen, Y.C., Lin, H., Shu, M., Li, R., Tao, X., Shen, X., Ye, Y., Jia, J.: Faceletbank for fast portrait manipulation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3541–3549 (2018) 4
6. 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 conference on computer vision and pattern recognition. pp. 7093–7102 (2018) 5, 13
7. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 3, 4, 5, 6, 9, 10, 11, 12, 13
8. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5908–5917 (2017) 5
9. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020) 3
10. Floater, M.S.: Parametrization and smooth approximation of surface triangulations. Computer aided geometric design **14**(3), 231–250 (1997) 6
11. Gecer, B., Deng, J., Zafeiriou, S.: Ostec: One-shot texture completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7628–7638 (2021) 5
12. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1155–1164 (2019) 3, 5, 7, 11, 12
13. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. arXiv preprint arXiv:2105.07474 (2021) 5
14. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8377–8386 (2018) 3, 11, 12
15. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3

16. Jolliffe, I.T.: Principal components in regression analysis. In: Principal component analysis, pp. 129–155. Springer (1986) 3
17. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020) 3, 4, 7, 8
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019) 4, 10, 11
19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020) 4, 8, 12
20. Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile gpus. In: Proceedings of CVPR Workshops (2019) 3, 4, 6, 10
21. Kim, J., Yang, J., Tong, X.: Learning high-fidelity face texture completion without complete face texture (2021) 5, 11
22. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813 3
23. Lin, J., Yuan, Y., Shao, T., Zhou, K.: Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5891–5900 (2020) 5, 11, 12, 13
24. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015) 13
25. Marriott, R.T., Romdhani, S., Chen, L.: A 3d gan for improved large-pose facial recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13445–13455 (2021) 4, 5, 12
26. Models, P.: The weights for all our pretrained models. (2021), the authors commit to publish upon acceptance of this paper or reviewer request. 6
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019) 11
28. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 497–500 (2001) 9
29. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 117–128 (2001) 9
30. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 (2020) 11
31. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 2016 fourth international conference on 3D vision (3DV). pp. 460–469. IEEE (2016) 3, 5
32. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1259–1268 (2017) 3, 5, 6

33. 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. pp. 1576–1585 (2017) 3, 5, 6
34. Shamai, G., Slossberg, R., Kimmel, R.: Synthesizing facial photometries and corresponding geometries using generative adversarial networks. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **15**(3s), 1–24 (2019) 3, 5, 12
35. Slossberg, R., Shamai, G., Kimmel, R.: High quality facial surface and texture synthesis via generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 3, 5, 6, 7
36. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zöllhofer, M., Theobalt, C.: Fml: Face model learning from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10812–10822 (2019) 3
37. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6142–6151 (2020) 4, 5, 7, 9
38. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3
39. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1274–1283 (2017) 3, 11
40. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion **64**, 131–148 (2020) 4
41. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.: Extreme 3d face reconstruction: Seeing through occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3935–3944 (2018) 6
42. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: In Proceeding of IEEE Computer Vision and Pattern Recognition. Salt Lake City, UT (June 2018) 3
43. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)
44. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12803–12813 (2021) 3
45. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016) 3