Normalized Avatar Synthesis Using StyleGAN and Perceptual Refinement

Huiwen Luo Koki Nagano Han-Wei Kung Qingguo Xu
Zejian Wang Lingyu Wei Liwen Hu Hao Li
Pinscreen

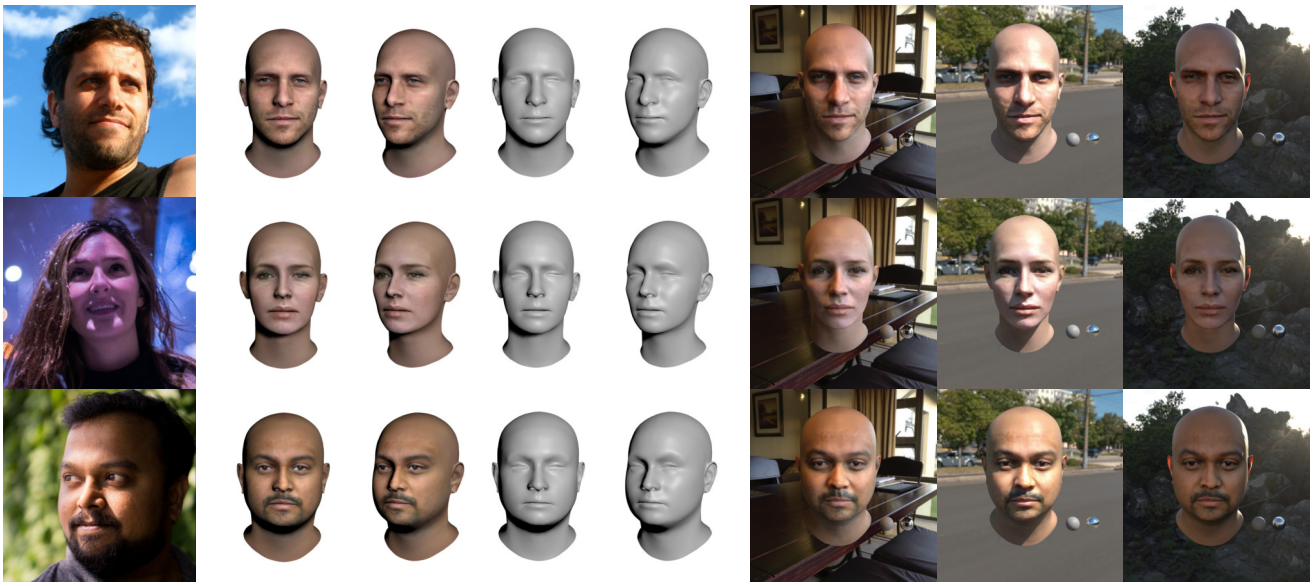

Figure 1: Given a single photograph, we can reconstruct a high-quality textured 3D face with neutral expression and normalized lighting condition. Our approach can handle extremely challenging cases and our generated avatars are animation friendly and suitable for complex relighting in virtual environments.

Abstract

We introduce a highly robust GAN-based framework for digitizing a normalized 3D avatar of a person from a single unconstrained photo. While the input image can be of a smiling person or taken in extreme lighting conditions, our method can reliably produce a high-quality textured model of a person's face in neutral expression and skin textures under diffuse lighting condition. Cutting-edge 3D face reconstruction methods use non-linear morphable face models combined with GAN-based decoders to capture the likeness and details of a person but fail to produce neutral head models with unshaded albedo textures which is critical for creating relightable and animation-friendly avatars for integration in virtual environments. The key challenges for existing methods to work is the lack of training and ground truth data containing normalized 3D faces. We propose a two-stage approach to address this problem. First, we adopt a highly robust normalized 3D face generator by embedding a non-linear morphable face model into a StyleGAN2 network. This allows us to generate detailed but normalized facial assets. This inference is then followed by a perceptual refinement step that uses the generated assets as regularization to cope with the limited available training samples of normalized faces. We further introduce a Normalized Face Dataset, which consists of a combination photogrammetry scans, carefully selected photographs, and generated fake people with neutral expressions in diffuse lighting conditions. While our prepared dataset contains two orders of magnitude less subjects than cutting edge GAN-based 3D facial reconstruction methods, we show that it is possible to produce high-quality normalized face models for very challenging unconstrained input images, and demonstrate superior performance to the current state-of-the-art.

Hao Li is affiliated with Pinscreen and UC Berkeley; Koki Nagano is currently at NVIDIA. This work was fully conducted at Pinscreen.

1. Introduction

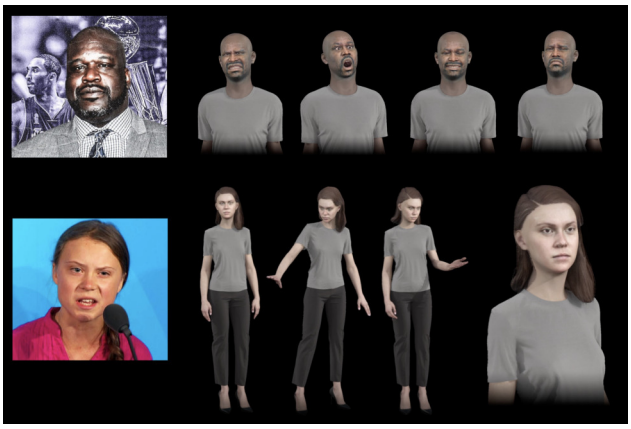

Figure 2: Automated digitization of normalized 3D avatars from a single photo.

The creation of high-fidelity virtual avatars have been mostly reserved to professional production studios and typically involves sophisticated equipment and controlled capture environments. Automated 3D face digitization methods that are based on unconstrained images such as selfies or downloaded internet pictures are gaining popularity for a wide range of consumer applications, such as immersive telepresence, video games, or social media apps based on personalized avatars.

Cutting-edge single-view avatar digitization solutions are based on non-linear 3D morphable face models (3DMM) generated from GANs [66, 65, 28, 45], outperforming traditional linear models [10] which often lack facial details and likeness of the subject. To successfully train these networks, hundreds of thousands of subjects in various lighting conditions, poses, and expressions are needed. While highly detailed 3D face models can be recovered, the generated textures have the lighting of the environment baked in, and expressions are often difficult to neutralize making these methods unsuitable for applications that require relighting or facial animation. In particular, inconsistent textured models are obtained when images are taken under different lighting conditions.

Collecting the same volume of 3D face data with neutral expressions and controlled lighting condition is intractable. Hence, we introduce a GAN-based facial digitization framework that can generate a high-quality textured 3D face model with neutral expression and normalized lighting using only thousands of real world subjects. Our approach consists of dividing the problem into two stages. The first stage uses a non-linear morphable face model embedded into a StyleGAN2 [40] network to robustly generate detailed and clean assets of a normalized face. The likeness of the person is then transferred from the input photograph using a perceptual refinement stage based on iterative optimization using a differentiable renderer. StyleGAN2 has proven to be highly expressive in generating and representing real world images using an inversion step to convert image to latent vector [3, 60, 4, 33] and we are adopting the same two step GAN-inversion approach to learn facial geometry and texture jointly. To enable 3D neutral face inference from an input image, we connect the image with the embedding space of our non-linear 3DMM using an identity regression network based on identity features from FaceNet [58]. To train a sufficiently effective generator, we introduce a new *Normalized Face Dataset* which consists of a combination of high-fidelity photogrammetry scans, frontal and neutral portraits in diffuse lighting conditions, as well as fake subjects generated using a pre-trained StyleGAN2 network with FFHQ dataset [39].

Despite our data augmentation effort, we show that our two-stage approach is still necessary to handle the large variation of possible facial appearances, expressions and lighting conditions. We demonstrate the robustness of our digitization framework on a wide range of extremely challenging examples, and provide extensive evaluations and comparisons with current state-of-the-art methods. Our method outperforms existing techniques in terms of digitizing textured 3D face models with neutral expressions and diffuse lighting conditions. Our normalized 3D avatars can be converted into parametric models with complete bodies and hair, and the solution is suitable for animation, relighting, and integration with game engines as shown in Fig. 2. We summarize our key contributions as follows:

- We propose the first StyleGAN2-based approach for digitizing a 3D face model with neutral expressions and diffusely lit textures from an unconstrained image.
- We present a two-stage digitization framework which consists of a robust normalized face model inference stage followed by a perception-based iterative face refinement step.
- We introduce a new data generation approach and dataset based on a combination of photogrammetry scans, photographs of expression and lighting normalized subjects, and generated fake subjects.
- Our method outperforms existing single-view 3D face reconstruction techniques for generating normalized faces, and we also show that our digitization approach works using limited subjects for training.

2. Related Works

While a wide range of avatar digitization solutions exist for professional production, they mostly rely on sophisticated 3d scanning equipment (e.g., multi-view stereo, photometric stereo, depth sensors etc.) and controlled capture settings [8, 30, 25]. We focus our discussion on monocular

3D face reconstruction methods as they provide the most accessible and flexible way of creating avatars for end-users, where only a selfie or downloaded internet photo is needed.

3D Morphable Face Models. Linear 3D Morphable Models (3DMM) have been introduced by Blanz and Vetter [10] two decades ago, and have been established as the de-facto standard for 3D face reconstruction from unconstrained input images. The linear parametric face model encodes shape and textures using principal component analysis (PCA) built from 200 laser scans. Various extensions of this work include the use of larger numbers of high-fidelity 3D face scans [12, 11], web images [41], as well as facial expressions often based on PCA or Facial Action Coding Systems(FACS)-based blendshapes [9, 68, 16].

The low dimensionality and effectiveness of 3DMMs make them suitable for robust 3D face modeling as well as facial performance capture in monocular settings. To reconstruct a textured 3D face model from a photograph, conventional methods iteratively optimize for shape, texture, and lighting condition by minimizing energy terms based on constraints such as facial landmarks, pixel colors [10, 57, 26, 61, 15, 37, 64, 27, 17, 48], or depth information if available such as for the case of RGB-D sensors [70, 69, 13, 46, 35, 50, 36].

While robust face reconstruction is possible, linear face models combined with gradient optimization-based optimization are ineffective in handling the wide variation of facial appearances and challenging input photographs. For instance, detailed facial hair and wrinkles are hard to generate and the likeness of the original subject is typically lost after the reconstruction. Deep learning-based inference techniques [71, 28, 21, 29, 63, 67, 22, 7, 63] were later introduced and have demonstrated significantly more robust facial digitization capabilities but they are still ineffective in capturing facial geometric and appearance detail due to the linearity and low dimensionality of the face model. Several post-processing techniques exist and use inferred linear face models to generate high-fidelity facial assets such as albedo, normal, and specular maps for relightable avatar rendering [43, 18, 72]. AvatarMe [43] for instance uses GANFIT [28] to generate a linear 3DMM model as input to their post processing framework. Our proposed method can be used as alternative input to AvatarMe, and we compare it to GANFIT later in Section 4.

More recently, non-linear 3DMMs have been introduced. Instead of representing facial shapes and appearances as a linear combination of basis vectors, these models are formulated implicitly as decoders using neural networks where the 3D faces are generated directly from latent vectors. Some of these methods use fully connected layers or 2D convolutions in image space [66, 6, 24, 65, 47], while others use decoders in the mesh domain to represent local geometries [51, 55, 76, 19, 5, 45, 49]. With the help of differentiable renderers [63, 29, 56], several methods [66, 65, 45] have demonstrated high-fidelity 3D face reconstructions using non-linear morphable face models using fully unsupervised or weakly supervised learning, which is possible using massive amounts of images in the wild. While the reconstructed faces are highly detailed and accurate w.r.t. the original input image, the generated assets are not suitable for relightable avatars nor animation friendly, since lighting conditions of the environment and expressions are baked into the output. Our work focuses on producing normalized 3D avatars with unshaded albedo textures and neutral expressions. Due to the limited availability of training data with normalized faces and the wide variation of facial appearances and capture conditions, the problem is significantly more challenging and ill-posed.

Generative Adversarial Network. We adopt StyleGAN2 [40] to encode our non-linear morphable 3D face model. Among all generative models in deep learning, Generative Adversarial Networks (GANs) [31] have achieved a great success in producing realistic 2D natural images, nearly indistinguishable from real world images. After a series of advancements, state-of-the-art GANs like PGGAN [38], BigGAN [14] and StyleGAN/StyleGAN2 [39, 40] have proven to be also effective in generating high resolution images and the ability to handle an extremely wide range of variations. In this work, we mainly focus on adopting StyleGAN2 [40] to jointly learn facial geometry and texture, since its intermediate latent representation has been proven effective to best reconstruct a plausible target image with clean assets [3, 60, 4, 33].

Facial Image Normalization. To address the problem of unwanted lighting and expressions during facial digitization, several methods have been introduced to normalize unconstrained portraits. Cole et al. [20] introduced a deep learning-based image synthesis framework based on FaceNet's latent code [58], allowing one to generate a frontal face with neutral expression and normalized lighting from an input photograph. More recently, Nagano et al [53] improved the method to generate higher resolution facial assets for the purpose generating high-fidelity avatars. In particular, their method breaks down the inference problem into multiple steps, solving explicitly for perspective undistortion, lighting normalization, followed by pose frontalization and expression neutralization. While the successful normalized portraits were demonstrated, their method rely on transferring details from the input subject to the generated output. Furthermore, both methods rely on the linear 3DMMs for expression neutralization and thus cannot capture detailed appearance variations. Neutralizing expression from nonlinear 3DMM, however, is not straightforward

since the feature space of identity and expression are often entangled. Our new normalization framework with GAN-based reconstruction fills in this gap.

3. Normalized 3D Avatar Digitization

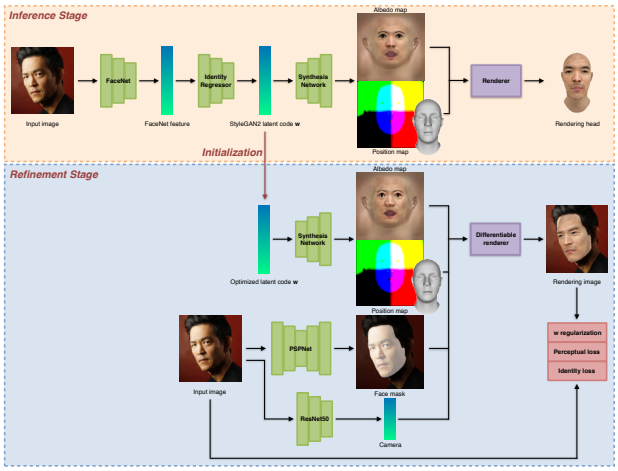

Figure 3: Two-stage facial digitization framework. The avatar is firstly predicted in the inference stage, and then improved to match the input image in the refinement stage.

An overview of our two-stage facial digitization framework is illustrated in Fig. 3. At the inference stage, our system uses a pre-trained face recognition network FaceNet [58] to extract a person-specific facial embedding feature given an unconstrained input image. This identity feature is then mapped to the latent vector $\mathbf{w} \in \mathcal{W}+$ in the latent space of our *Synthesis Network* using an *Identity Regressor*. The synthesis network decodes $\mathbf{w}$ to an expression neutral face geometry and a normalized albedo texture. For the refinement, the latent vector $\mathbf{w}$ produced by the inference is then optimized iteratively using a differentiable renderer by minimizing the perceptual difference between the input image and the rendered one via gradient descent.

3.1. Robust GAN-Based Facial Inference

Our synthesis network G generates the geometry as well as the texture in UV space. Each pixel in the UV map represents the 3D position and the RGB albedo color of the corresponding vertex using a 6-channel tuple $(\mathsf{r}, \mathsf{g}, \mathsf{b}, \mathsf{x}, \mathsf{y}, \mathsf{z})$. The synthesis network is first trained using a GAN to ensure robust and high quality mapping from any normal distributed latent vector $\mathcal{Z} \sim \mathcal{N}(\mu, \sigma)$. Then, the identity regression network R is trained by freezing G to ensure accurate mapping from the identity feature of an input image. Further details of each network are described below.

We train our synthesis network to embed a nonlinear 3D Morphable Model into its latent space, in order to model the cross correlation between the 3D neutral face geometry

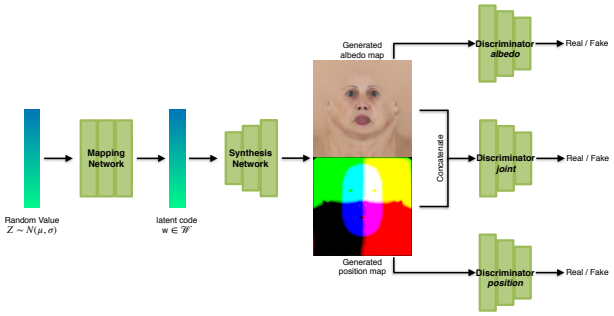

Figure 4: GAN-based geometry and texture synthesis.

and the neutral albedo texture, as well as to generate high fidelity and diverse 3D neutral faces from a latent vector. Inspired by [47], we adopt the StyleGAN2 [40] architecture to train a morphable face model using 3D geometry and albedo texture as shown in Fig. 4. Rather than predicting vertex positions directly, we infer vertex position offsets relative to the mean face mesh to improve numerical stability. To jointly learn geometry and texture, we project the geometry representation of classical linear 3DMMs $S \in \mathbb{R}^{3\times N}$, which consists of a set of $N = 13557$ vertices on the face surface, onto a UV space using cylindrical parameterization. The vertex map is then rasterized to a 3-channel position map with 256×256 pixels. Furthermore, we train 3 discriminators jointly, including 2 individual ones for albedo and vertex position as well as a joint discriminator taking both maps as input. The individual discriminators ensure the quality and sharpness of each generated map, while the joint discriminator can learn and preserve their correlated distribution. This GAN is trained solely from the provided ground truth 3D geometries and albedo textures without any knowledge of the identity features.

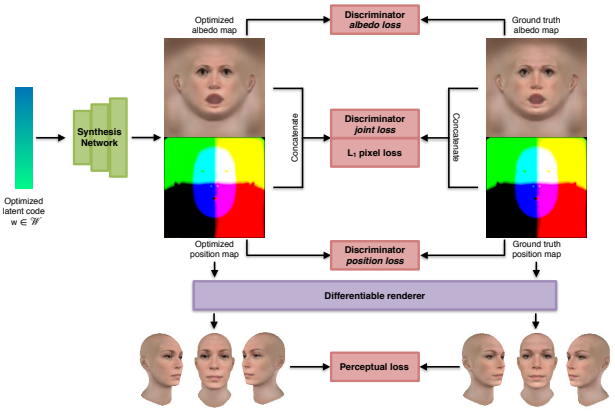

Figure 5: Our GAN-inversion searches a corresponding $\mathbf{w}$, which can reconstruct the target geometry and texture.

After obtaining G, we retrieve the corresponding input latent code via our code inversion algorithm. Inspired by [3,

77], we choose the disentangled and extended latent space $\mathcal{W}+ := \mathbb{R}^{14\times512}$ of StyleGAN2 as the inversion space to achieve better reconstruction accuracy. As shown in Fig. 5, we adopt an optimization approach to find the embedding of a target pair of position and albedo map with the following loss function:

$$L_{inv} = L_{pix} + \lambda_1 L_{\text{LPIPS}} + \lambda_2 L_{adv} \quad (1)$$

where L_{pix} is the L_1 pixel error of the synthesized position and texture maps, L_{LPIPS} is the LPIPS distance [74] as a perceptual loss, and L_{adv} is the adversarial loss favoring realistic reconstruction results using the three discriminators trained with G. Note that while LPIPS outperforms other perceptual metrics in practice [74], it is trained with real images and measuring the perceptual loss directly on our UV maps would lead to unstable results. Therefore, we use a differentiable renderer [56] to render the geometry and texture maps from three fixed camera viewpoints and compute the perceptual loss based on these renderings. Finally, the identity regressor R can be trained using the solved latent codes of the synthesis network and their corresponding identity features from the input images.

3.2. Unsupervised Dataset Expansion

Figure 6: Examples of synthetic faces from our *Normalized Face Dataset*.

While datasets exist for frontal human face images in neutral expression [52, 23, 32, 42], the amount of such data is still limited and the lighting conditions often vary between datasets. Instead of manually collecting more images from the Internet for expanding our training data, we propose an automatic approach to produce frontal neutral portraits based on the pre-trained StyleGAN2 network trained with FFHQ dataset. Similar to a recent technique for semantic face editing [60], we train a neural network to predict identity attributes α of an input image in latent space. We used images collected from internet as input and estimate each α and apply it to $\mathbf{w}_{mean}$. $\mathbf{w}_{mean}$ is a fixed value in latent space, which could generate a mean and frontalized face. We then use a latent editing vector β to neutralize the expressions. The final latent value $\mathbf{w}' = \mathbf{w}_{mean} + \alpha + \beta$ produces a frontalized and neutralized face by feeding into StyleGAN2. Some examples are shown in Fig. 6. We further emphasize that all images in our *Normalized Face Dataset* are frontal and have neutral expressions. Also, these images have well conditioned diffuse scene illuminations, which are preferred for conventional gradient descent-based 3D face reconstruction methods.

For each synthesized image, we apply light normalization [53] and 3D face fitting based on Face2Face [64] to generate a 3D face geometry and then project the light normalized image for the albedo texture. Instead of using the linear 3DMM completely, which results in coarse and smooth geometry, we first run our inference pipeline to generate the 3D geometry and take it as the initialization for the Face2Face optimization. After optimization, the resulting geometry is in fact the non-linear geometry predicted from our inference pipeline plus a linear combination of blendshape basis optimized by Face2Face, thus preserving its non-linear expressiveness. Also note that the frontal poses of the input images facilitate our direct projections onto UV space to reconstruct high-fidelity texture maps.

The complete training procedure works as follows: we first collect a high quality *Scan Dataset* with 431 subjects with accurate photogrammetry scans, with 63 subjects from 3D Scan Store [1] and 368 subjects from Triplegangers [2]. The synthesis network G_0 is then trained from such scan data, and is then temporarily frozen for latent code inversion and the training of identity regressor R_0. These bootstrapping networks (R_0, G_0) trained on the small *Scan Dataset* are applied onto our *Normalized Face Dataset* to infer the geometry and texture, which are then optimized and/or corrected by the Face2Face algorithm. Next, the improved geometry and texture are added back into the training of (R_0, G_0) to obtain the fine-tuned networks (R_1, G_1) with improved accuracy and robustness.

Our final *Normalized Face Dataset* consists of 5601 subjects, with 368 subjects from Triplegangers, 597 from Chicago Face Dataset (CFD) [52], 230 from the compound facial expressions (CFE) dataset [23], 153 from The CMU Multi-PIE Face Dataset [32], 67 from Radboud Faces Database (RaFD) [42], and the remaining 4186 generated by our method. We use most of the frontal and neutral face images that are available to increase diversity, but still rely on the large volume of synthetic data for the training.

3.3. Perceptual Refinement

While the inference pipeline described in Sec. 3.1 with training data from Sec. 3.2 can reliably infer the normalized texture and geometry from an unconstrained image, a second stage with perceptual refinement can help determine a neighbor of the predicted latent code in the embedding space that matches the input image better. The work from Shi et al. [62] shows that an embedding space learned for face recognition is often noisy and ambiguous due to the nature of fully unconstrained input data. While FaceNet

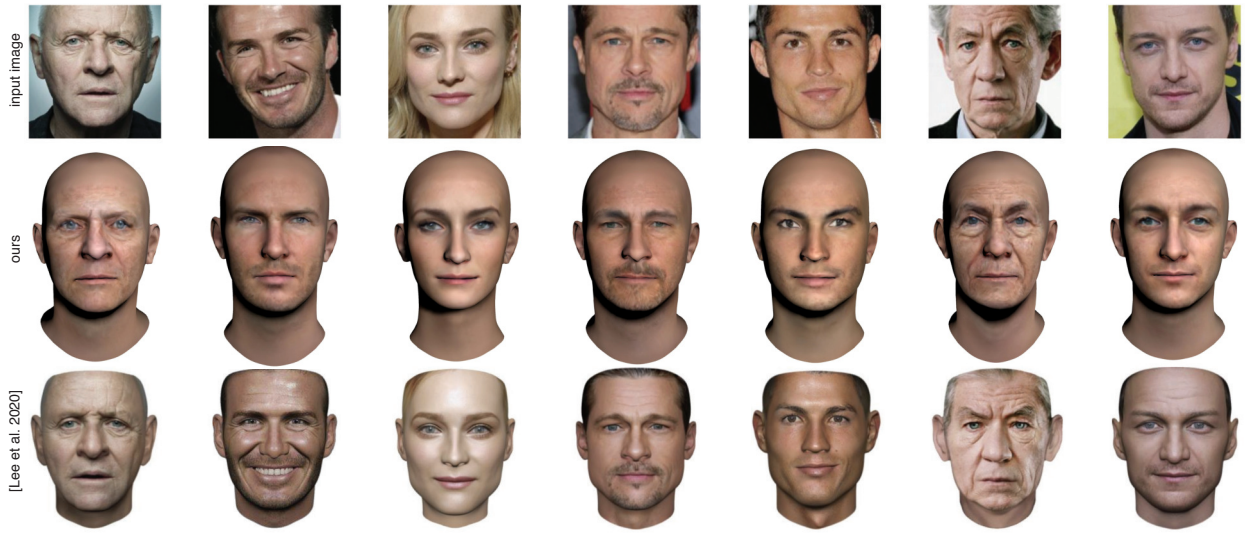

Figure 7: Qualitative comparison with other state-of-the-art 3D face reconstruction method. The first row shows the input images and the second row shows our results, and the third row are the reconstructed 3D faces obtained by [45].

predicts the most likely latent code, the variance (or *uncertainty* in Shi et al.'s work) could be large. A small perturbation of the latent code may not affect the identity feature training at all. On the other hand, such a small error in the identity code may cause greater inconsistency in our inference pipeline after passing R and G.

An "end-to-end" refinement step is introduced, to handle never seen before images while ensuring consistency between the final renderings using the predicted geometry and texture, and the input image. Fig. 3 shows the end-to-end architecture for this refinement step. We reuse the differentiable renderer to generate a 2D face image $\hat{I}$ from the estimated 3D face, and compute the perceptual distance with the input image I. To project the 3D face back to the head pose in image I, we train a regression network with ResNet50 [34] as backbone to estimate the camera $c = [t_x, t_y, t_z, r_x, r_y, r_z, f]^T$ from I, where $[t_x, t_y, t_z]^T$ and $[r_x, r_y, r_z]^T$ denote the camera translation and rotation and f is the focal length. The network is trained using the accurate camera data from the *Scan Dataset* and the estimated camera data from *Normalized Face Dataset*, computed by Face2Face. Furthermore, in order to blend the projected face only image with the background from the original image I, we train a PSPNet [75] with ResNet101 [34] as backbone using CelebAMask-HQ [44]. We then blend the rendered image $\hat{I}$ into the segmented face region from I to produce I_0. The final loss is simply represented as:

$$L_{refine} = L_w + \lambda_1 L_{\text{LPIPS}} + \lambda_2 L_{id} \quad , \qquad (2)$$

where L_w is a regularization term on $\mathbf{w}$, i.e., the Euclidean distance between the variable $\mathbf{w}$ and its initial prediction derived by R, enforcing the similarity between the modified latent and the initial prediction. L_{LPIPS} is the perceptual loss measured by LPIPS distance [74] between I_0 and I, which enables improved matching in terms of robustness and better preservation of semantically meaningful facial features compared to using pixel differences. L_{id} is the cosine similarity between the identity feature of $\hat{I}$ and I, to preserve consistent identity.

4. Results

We demonstrate the performance of our method in Fig. 1 and 7, and show how our method can handle extremely challenging unconstrained photographs with very harsh illuminations, extreme filtering, and arbitrary expressions. We can produce plausible textured face models where the likeness of the input subject is preserved and visibly recognizable. Compared to the state-of-the-art 3D face reconstruction method (see Fig. 7) based on non-linear 3DMMs, our method can neutralize expressions and produce an unshaded albedo texture suitable for rendering in arbitrary lighting conditions as demonstrated using various HDRI-based lighting environments. We also show in Fig. 2 how we can obtain a fully rigged 3D avatar from a single photo including body and hair, by adopting the hair digitization algorithm in [36] (see accompanying video for live demo).

Evaluations. Sec. 3.2 further improves the performance of G and R using more training data. Fig. 8 compares the default Face2Face optimization using a linear 3DMM with the improved ones using an initialization from R_0 and G_0.

With such synthetic training data, Fig. 9 shows improved expressiveness of G_1 than G_0. Several artifacts from G_0

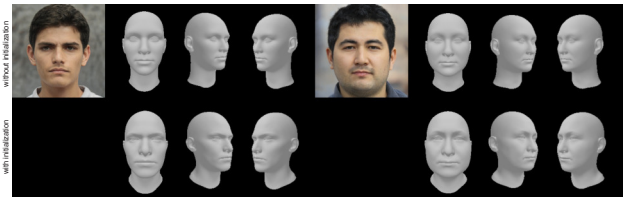

Figure 8: Face2Face optimization results. The first row is the original implementation [64]. The second row is our proposed improvement with nonlinear initialization.

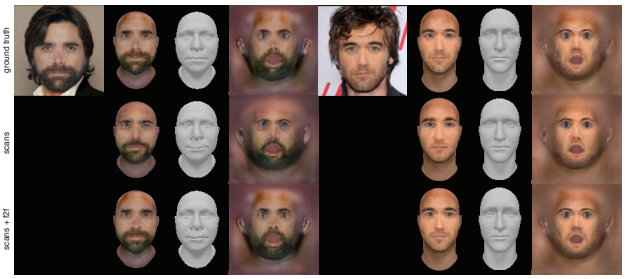

Figure 9: Expressiveness of the synthesis network trained with different datasets. From top to bottom: The ground truth; The GAN-inversion results based on G_0 trained with *Scan Dataset* only; The same process based on G_1, trained with *Normalized Face Dataset*.

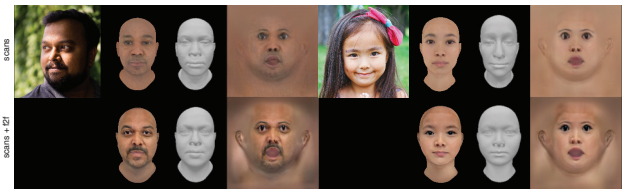

Figure 10: Quality of the regression network trained with different datasets. The first row shows the inference results by R_0m trained with *Scan Dataset*. The second row shows the results by R_1, trained with *Normalized Face Dataset*.

Figure 11: Qualitative comparison with different initialization schemes for iterative refinement. The mean initialization starts optimization from a mean latent vector of our training dataset. The inference initialization starts from the latent vector predicted by R.

around eyes and the lack of facial hair are fixed in G_1. In Fig. 10, R_1 also shows higher diversity of face shapes and superior accuracy compared to R_0 after training with the *Normalized Face Dataset*.

Fig. 11 demonstrates the effect of both the inference stage in Sec. 3.1 and the refinement stage. For each row of the experiment, the end-to-end iterative refinement can always improve the likeness and expressiveness of the 3D avatar. However, notice that the refinements from the mean latent vector would fail to produce a faithful result after 200 iterations, while the refinements from an accurate initial prior by R converges to a highly plausible face reconstruction.

Figure 12: Consistent reconstructions of the same person under different environments.

Since our proposed pipeline simply rely on the identity and perceptual features from I, the reconstructed 3D avatar is invariant to the factors FaceNet filters, such as occlusion, image resolution, lighting environment, and facial expression. Fig. 12 demonstrates how we can obtain consistent geometries from different lighting, viewpoints, and facial expressions. Further results of more challenging images, such as low resolution or largely occluded ones are provided in the supplemental material.

Comparisons. Fig. 7 compare our method with the most recent single view face reconstruction method [45]. Lee et al. [45] adopts a state-of-the-art nonlinear 3DMM on both geometry and texture. They also use a Graph Convolutional Neural Network to embed geometry and a Generative Adversarial Network to synthesize texture. However, they train two networks separately with different datasets, where facial shape and appearance are uncorrelated. More importantly, their results show that expressions and lighting are baked in, which makes their method unsuitable for relighting and facial animation purposes. More comparisons with other monocular face reconstruction methods [21, 28, 65] can be found in the supplemental material.

Fig. 13 shows our results compared to the deep face normalization method [53]. While some successful normalized results were demonstrated, their image-to-image translation architecture transfers details from the input subject to the generated output. If those details are deteriorated, then face normalization would fail.

Quantitative experiments on FaceScape [73] using high resolution 3D scans and corresponding images are shown in

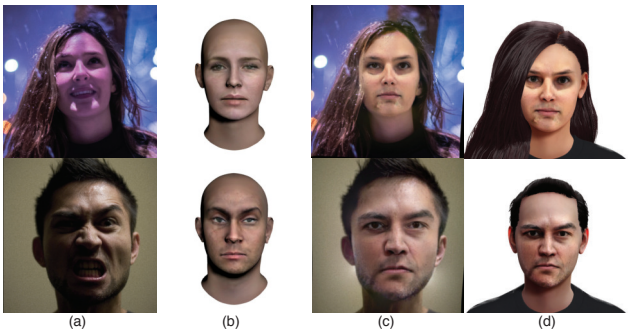

Figure 13: Qualitative comparison with state-of-the-art face normalization method [53]. From left to right, we show (a) input image; (b) our reconstructed result; (c) image-based face normalization result generated by Nagano et al. [53]; (d) Face2Face reconstruction result based on (c).

Tran et al. [65]	Deng et al. [21]	Ours
1.935mm	1.568mm	**1.557mm**

Table 1: Quantitative comparison of with other 3D face reconstruction methods.

Tran et al. [65]	Deng et al. [21]	Ours
0.304	0.392	**0.205**

Table 2: Quantitative comparison on texture.

Tables 1 and 2. For geometric accuracy, we randomly select 20 scans from FaceScape, and for each method, we compute the average point to mesh distance between the monocular reconstructed geometry and the ground truth scan. The proposed model has smaller reconstruction errors than other state-of-the-art ones. For texture evaluation, we augment the input images with lighting variations and compute the mean L1 pixel loss between generated textures from each method and the ground truth. Our method generates textures that are less sensitive to lighting conditions.

Implementation Details. All our networks are trained on a desktop machine with Intel i7-6800K CPU, 32GB RAM and one NVIDIA TITAN GTX (24GB RAM) GPU using PyTorch [54]. The StyleGAN2 network training takes 13 days with the *Normalized Face Dataset*. We use the PyTorch implementation [59] and remove the noise injection layer in the original implementation to remove the stochastic noise inputs and enable full control of the generated results from the latent vector. Our identity regression network is composed of four fully connected layers with Leaky ReLU activations, and the training takes 1 hour to converge with the same training data. At the testing stage, inference takes 0.13 s and refinement takes 45 s for 200 iterations.

Figure 14: Failure cases in our experiments. (a) shows a failure where the specularity at the chin is baked into the generated result; (b) shows that the robustness of the reconstruction result is affected by the exaggerated expression.

5. Discussion

We have demonstrated a StyleGAN2-based digitization approach using a non-linear 3DMM that can reliably generate high-quality normalized textured 3D face models from challenging unconstrained input photos. Despite the limited amount of available training data (only thousands of subjects), we have shown that our two-stage face inference method combined with a hybrid *Normalized Face Dataset* is effective in digitizing relightable and animation friendly avatars and can produce results of quality comparable to state-of-the-art techniques where generated faces are not normalized. Our experiments show that simply adopting existing methods using limited normalized facial training data is insufficient to capture the likeness and fine-scale details of the original subject, but a perceptual refinement stage is necessary to transfer person-specific facial characteristics from the input photo. Our experiments also show that perceptual loss enables more robust matching using deep features than only pixel loss, and is able to better preserve semantically meaningful facial features. Compared to state-of-the-art non-linear 3DMMs, our generated face models can produce lighting and expression normalized face models, which is a requirement for seamless integration of avatars in virtual environments. Furthermore, our experiments also indicate that our results are not only perceptually superior, but also quantitatively more accurate and robust than existing methods.

Limitations and Future Work. As shown in Fig. 14, the effectiveness of our method in generating faces with normalized expressions and lighting is limited by imperfect training data and challenging input photos. In particular, some expressions and specularities can still be found in the generated results. Furthermore, the fundamental problem of disentangling identity from expressions, or lighting conditions from skin tones is ill-posed. Nevertheless, we believe that such disentanglement can be improved using superior training data. In the future, we would like to explore how to increase the resolution and fidelity of the digitized assets and potentially combine our method with high-fidelity facial asset inference techniques such as [43, 18, 72].

References

[1] 3d scan store. https://www.3dscanstore.com/.

[2] Triplegangers. http://www.triplegangers.com.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[4] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. A decoupled 3d facial shape model by adversarial training. In *CVPR*, 2019.

[6] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *CVPR*, 2018.

[7] Anil Bas, Patrik Huber, William A. P. Smith, Muhammad Awais, and Josef Kittler. 3d morphable models as spatial transformer networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[8] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. *ACM Trans. Graph.*, 29(4), 2010.

[9] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22. Wiley Online Library, 2003.

[10] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[11] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models "in-the-wild". In *CVPR*, 2017.

[12] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016.

[13] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Trans. Graph.*, 32(4), 2013.

[14] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[15] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4), 2014.

[16] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 20(3), 2014.

[17] Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph.*, 35(4), 2016.

[18] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *CVPR*, 2019.

[19] Shiyang Cheng, Michael M. Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. Meshgan: Non-linear 3d morphable models of faces. *CoRR*, 2019.

[20] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3703–3712, 2017.

[21] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[22] Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[24] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[25] G. Fyffe, P. Graham, B. Tunwattanapong, A. Ghosh, and P. Debevec. Near-instant capture of high-resolution facial geometry and reflectance. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics*, EG '16, page 353–363, Goslar, DEU, 2016. Eurographics Association.

[26] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013)*, volume 32, November 2013.

[27] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Trans. Graph. (Presented at SIGGRAPH 2016)*, 35(3), 2016.

[28] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. volume 30, page 129. ACM, 2011.

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[32] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807 – 813, 2010. Best of Automatic Face and Gesture Recognition 2008.

[33] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding, 2020.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[35] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *IEEE CVPR*, 2015.

[36] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, Nov. 2017.

[37] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph.*, 34(4), 2015.

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[41] Ira Kemelmacher-Shlizerman. Internet-based morphable model. *ICCV*, 2013.

[42] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and Emotion*, 24(8):1377–1388, 2010.

[43] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction "in-the-wild". In *CVPR*, June 2020.

[44] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[45] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[46] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph. (Proceedings SIGGRAPH 2013)*, 32(4), 2013.

[47] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *CVPR*, 2020.

[48] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), Nov. 2017.

[49] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[50] Xiangkai Lin, Yajing Chen, Linchao Bao, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Xinwei Jiang, Jue Wang, Dong Yu, and Zhengyou Zhang. High-fidelity 3d digital human creation from rgb-d selfies. *arXiv preprint arXiv:2010.05562*, 2020.

[51] Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *CVPR*, 2018.

[52] Debbie Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47:1122–1135, 01 2015.

[53] Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. Deep face normalization. *ACM Trans. Graph.*, 2019.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[55] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3d faces using convolutional mesh autoencoders. In *ECCV*, 2018.

[56] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[57] Romdhani Sami and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[58] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[59] Kim Seonghyeon. stylegan2-pytorch. https://github.com/rosinality/stylegan2-pytorch.

[60] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[61] Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.*, 33(6), 2014.

[62] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6902–6911, 2019.

[63] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017.

[64] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[65] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.

[66] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

[67] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[68] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3), 2005.

[69] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. *ACM Trans. Graph. (Proceedings SIGGRAPH 2011)*, 30(4), July 2011.

[70] Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer animation (Proc. SCA'09)*, ETH Zurich, August 2009. Eurographics Association.

[71] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *CVPR*, 2019.

[72] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.*, 37(4), 2018.

[73] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[75] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[76] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, 2019.

[77] Jiapeng Zhu, Deli Zhao, Bo Zhang, and Bolei Zhou. Disentangled inference for gans with latently invertible autoencoder. *arXiv preprint arXiv:1906.08090*, 2019.