Single-Shot Implicit Morphable Faces with Consistent Texture Parameterization

CONNOR Z. LIN*, Stanford University, USA and NVIDIA, USA

KOKI NAGANO, NVIDIA, USA

JAN KAUTZ, NVIDIA, USA

ERIC R. CHAN*, Stanford University, USA and NVIDIA, USA

UMAR IQBAL, NVIDIA, USA

LEONIDAS GUIBAS, Stanford University, USA

GORDON WETZSTEIN, Stanford University, USA

SAMEH KHAMIS, NVIDIA, USA
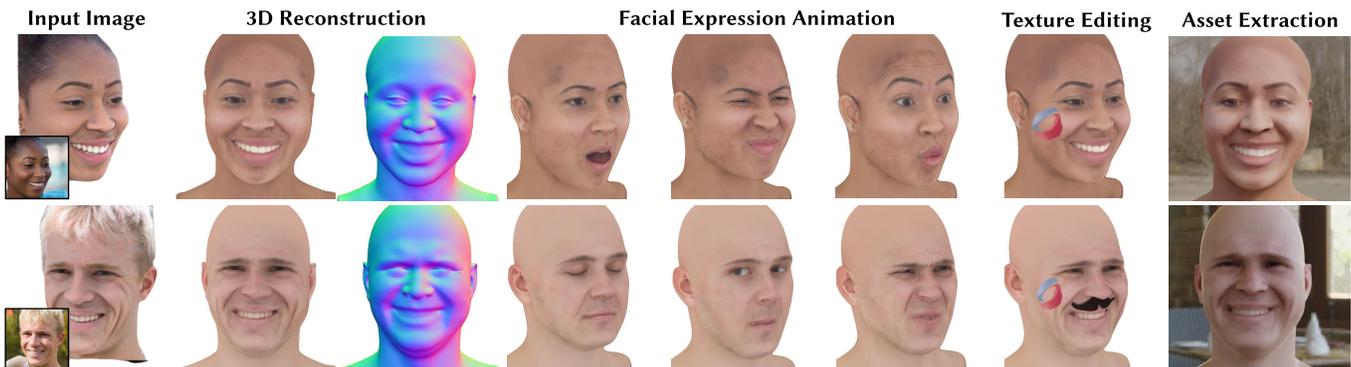
Fig. 1. Given a single input image, our method reconstructs a high-quality editable 3D digital avatar (columns 2 and 3) by combining implicit geometry representations with explicit texture maps. The proposed approach naturally supports novel view synthesis from large pose shifts, an expressive and non-linear facial animation space (columns 4 through 6), direct user access to texture map editing (column 7), and 3D asset extraction for further downstream applications such as relighting (column 8).

There is a growing demand for the accessible creation of high-quality 3D avatars that are animatable and customizable. Although 3D morphable models provide intuitive control for editing and animation, and robustness for single-view face reconstruction, they cannot easily capture geometric and appearance details. Methods based on neural implicit representations, such as signed distance functions (SDF) or neural radiance fields, approach photo-realism, but are difficult to animate and do not generalize well to unseen data. To tackle this problem, we propose a novel method for constructing implicit 3D morphable face models that are both generalizable and intuitive for editing. Trained from a collection of high-quality 3D scans, our face model is parameterized by geometry, expression, and texture latent codes with a learned SDF and explicit UV texture parameterization. Once trained, we can reconstruct an avatar from a single in-the-wild image by leveraging the learned prior to project the image into the latent space of our model. Our implicit morphable face models can be used to render an

avatar from novel views, animate facial expressions by modifying expression codes, and edit textures by directly painting on the learned UV-texture maps. We demonstrate quantitatively and qualitatively that our method improves upon photo-realism, geometry, and expression accuracy compared to state-of-the-art methods.

CCS Concepts: • **Computing methodologies → Modeling/Geometry**.

Additional Key Words and Phrases: Neural Avatars, Implicit Representations, Texture Maps, Animation, Inversion

*Work done during an internship at NVIDIA.

1 INTRODUCTION

Personalized avatar creation—the ability to map one's facial features to a 3D virtual replica that can be animated, customized, and rendered—is an emerging technology with great promise for cinema, the metaverse, and telepresence. Advances in this area may lead to digital twins with greater verisimilitude in detail and in animation that are more easily integrated into downstream applications and

pipelines. Single-shot personalized avatar creation enables reconstructing face avatars from individual RGB images with greater convenience and flexibility than methods that require more specialized capture setups or procedures.

Traditional approaches to animatable 3D avatar creation are often based on 3D Morphable Models (3DMM) [Blanz and Vetter 1999], which disentangle shape and appearance variation into a low-dimensional face representation. Building on these, more recent approaches often leverage either explicit (textured) template meshes [Daněček et al. 2022; Feng et al. 2021; Grassal et al. 2022; Khakhulin et al. 2022; Li et al. 2017; Tran and Liu 2019] or neural implicit representations [Mildenhall et al. 2021; Park et al. 2019; Sitzmann et al. 2019]. Template-based approaches enable easy asset extraction and intuitive editing, but are often unable to capture high-quality geometry and textures. Emerging implicit face models can achieve greater realism by modeling more complex geometric features such as hair [Cao et al. 2022b; Giebenhain et al. 2022; Zheng et al. 2022a]. However, implicit face representations often compromise on interpretability and are less intuitive to control; the entangled latent spaces learned by these highly parameterized models are difficult to edit.

Our approach aims to combine the interpretability and editability advantages of template-based 3DMMs with the quality and topological flexibility of implicit 3D representations. Crucially, we decouple appearance and geometry into two branches of our network architecture. By incorporating a UV parameterization network to learn continuous and consistent texture maps, we can export avatars as textured meshes to support downstream applications such as texture map editing and relighting in a traditional graphics pipeline (See Figure 1). On the other hand, by representing geometry with an implicit signed distance field (SDF), our facial shape is less limited by resolution and topology compared to mesh-based approaches.

We show that our proposed hybrid representation effectively captures the geometry, appearance, and expression space of faces. We demonstrate that single-shot in-the-wild portrait images can be effectively mapped to avatars based on our proposed representation, and that these avatars improve upon the previous state-of-the-art in photo-realism, geometry, and monocular expression transfer. Moreover, we demonstrate compelling capability for enabling direct texture editing and disentangled attribute editing such as facial geometry and appearance attributes.

In summary, contributions of our work include:

- We propose a hybrid morphable face model that combines the high-quality geometry and flexible topology of implicit representations with the editability of explicit UV parameterized texture maps.
- We present a single-shot inversion framework to map a single in-the-wild RGB image to our implicit 3D morphable model representation. The inverted avatar supports novel view rendering, non-linear facial reanimation, disentangled shape and appearance control, direct texture map editing, and textured mesh extraction for downstream applications.
- We demonstrate that our method achieves state-of-the-art reconstruction accuracy for photo-realistic rendering, geometry, and expression accuracy in the single-view reconstruction setting.

Table 1. Comparison to recent prior work. To the best of our knowledge, our method is the first implicit 3D face model to generalize across single-image inputs while supporting flexible topology and explicit texture map control.

	Generalizable	Single-Image	Implicit Representation	Explicit Texture Control
EMOCA [2022]	✓	✓	✗	✓
ROME [2022]	✓	✓	✗	✗
Neural Parametric Head Models [2022]	✗	✗	✓	✗
IM-Avatar [2022a]	✗	✗	✓	✗
Neural Head Avatars [2022]	✗	✗	✓	✓
Volumetric Avatars from a Phone Scan [2022b]	✓	✗	✓	✓
HeadNeRF [2022]	✓	✓	✓	✗
Ours	✓	✓	✓	✓

2 RELATED WORK

2.1 Mesh-based 3D Morphable Models

The seminal work by Blanz and Vetter proposed a linear 3D Morphable Model (3DMM) [Blanz and Vetter 1999] that models facial shape and textures on a template mesh using linear subspaces computed by principal component analysis (PCA) from 200 facial scans. This low-dimensional facial shape and texture space makes 3DMMs suitable for robustly capturing facial animation as well as reconstructing 3D faces in monocular settings. To reconstruct shape, texture, and lighting from a photo, previous work employed continuous optimization using constraints such as facial landmarks and pixel colors [Cao et al. 2014, 2016; Garrido et al. 2013, 2016; Ichim et al. 2015; Li et al. 2017; Romdhani and Vetter 2005; Shi et al. 2014; Thies et al. 2016] and more recently deep learning-based inference [B R et al. 2021; Daněček et al. 2022; Deng et al. 2019b; Dib et al. 2021a,b; Dou et al. 2017; Feng et al. 2021; Genova et al. 2018; Luo et al. 2021; Tewari et al. 2019; Tewari et al. 2017; Tuan Tran et al. 2017; Wu et al. 2019]. While approaches relying on 3DMMs tend to be robust, they are ineffective for reconstructing high-fidelity geometry and texture details due to the linearity and low dimensionality of the model. Various other methods extended 3DMMs to capture non-linear shapes [Chandran et al. 2020; Li et al. 2020; Tewari et al. 2018; Tran et al. 2019; Tran and Liu 2018, 2019; Wang et al. 2022b], photo-realistic appearance using neural rendering or optimization [Gecer et al. 2019; Nagano et al. 2018; Saito et al. 2017; Thies et al. 2019], or reflectance and geometry details for relightable avatar generation [Chen et al. 2019; Huynh et al. 2018; Lattas et al. 2020; Yamaguchi et al. 2018]. Recent approaches predict geometry offsets over the template mesh to reconstruct non-facial regions such as hair [Grassal et al. 2022; Khakhulin et al. 2022]. We refer the reader to Egger et al. [2020] for an in-depth survey of 3DMM techniques and Tewari et al. [2022] for a report of recent advancements in neural rendering.

Since mesh-based 3DMMs represent geometry with a shared template mesh, their fixed topology limits the ability to scale the model to capture complex geometry such hair or fine-scale details. Additionally, their ability to synthesize photo-realistic facial textures may be limited by the resolution of the template mesh and discrete texture map. By parameterizing geometry with a signed distance function and color with a continuous texture map, our method is able to avoid such resolution issues and scale more efficiently with model capacity while retaining 3DMM-like intuitive parameters to individually control geometry and textures. Our consistent texture parameterization enables not only direct texture editing in UV space,

but also semantic correspondence between our face model and an input image via facial landmarks, which can be leveraged to improve single-shot reconstruction quality.

2.2 Implicit Representations for Modeling and Rendering

While single-shot 3D reconstruction methods have explored various explicit 3D representations such as voxels [Girdhar et al. 2016; Tulsiani et al. 2017; Wu et al. 2018; Yan et al. 2016; Yang et al. 2018; Zhu et al. 2017], point clouds [Fan et al. 2017], meshes [Xu et al. 2019], geometric primitives [Niu et al. 2018; Zou et al. 2017], and depth maps [Wu et al. 2020], implicit representations have recently been leveraged to achieve higher resolution reconstruction using occupancy or signed distance fields (SDFs) [Chen and Zhang 2019; Mescheder et al. 2019; Xu et al. 2019]. Implicit representations such as neural radiance fields (NeRFs) [Mildenhall et al. 2021] and signed distance fields (SDFs) [Park et al. 2019] have demonstrated high reconstruction quality for 3D shapes and volumetric scenes. PIFu [Saito et al. 2019] and follow-up works [Cao et al. 2022a; Saito et al. 2020] use implicit fields to model human bodies and clothing. AtlasNet [Groueix et al. 2018] demonstrated 3D shape generation by predicting a set of parametric surface elements given an input image or point cloud. NeuTex [Xiang et al. 2021] replaces the radiance prediction of NeRFs with a learned UV texture parameterization conditioned on lighting direction. Although our method also employs a UV cycle consistency loss, we 1) operate in a SDF setting and condition our parameterization on geometry and expression latent codes to generalize across samples rather than overfit to a single scene, 2) employ sparse facial landmark constraints to facilitate learning a semantically intuitive and consistent parameterization, and 3) explicitly leverage 2D to 3D facial landmark correspondences enabled by the learned consistent parameterization during single-image reconstruction. Implicit representations have also given rise to higher quality 3D generative models [Chan et al. 2022; Or-El et al. 2022; Xue et al. 2022], and follow-up work has studied inverting an image into the latent space of a pre-trained 3D GAN [Ko et al. 2023; Lin et al. 2022; Roich et al. 2022] for single-view 3D reconstruction. However, without careful optimization and additional priors [Xie et al. 2022; Yin et al. 2022], this 3D GAN inversion tends to be less robust due to unknown camera poses [Ko et al. 2023] and multi-view nature of NeRF training in the monocular setting. On the other hand, the compact face representation of our model provides robust initialization in the single-shot reconstruction setting.

2.3 Implicit Face Models

Compared to traditional mesh-based 3DMMs for face modeling, implicit representations naturally offer flexible topology and non-linear expression animation through latent code conditioning. While some approaches learn to reconstruct an implicit 3DMM from an input 3D face scan [Alldieck et al. 2021; Cao et al. 2022b; Giebenhain et al. 2022; Yenamandra et al. 2021; Zanfir et al. 2022; Zheng et al. 2022b], other works have explored modeling an implicit face model from RGB videos [Grassal et al. 2022; Ma et al. 2022; Zheng et al. 2022a,c]. However, the above approaches either do not support or cannot generalize to single-shot in-the-wild images. Multi-view methods have also been used to reconstruct implicit head models [Athar et al. 2021, 2022; Hong et al. 2022; Kellnhofer et al. 2021; Li et al. 2022;

Ramon et al. 2021; Wang et al. 2022a]. HeadNeRF [Hong et al. 2022] is the closest to our work and learns a parametric head model from multi-view images during training; at test-time, an input image can be inverted for 3D reconstruction. However, HeadNeRF performs volumetric rendering at a limited image resolution and relies on upsampling CNN modules, resulting in flickering artifacts from depth error during novel view synthesis. Furthermore, existing implicit morphable models do not support texture manipulation beyond interpolation; by contrast, our learned explicit texture paramterization enables intuitive and out-of-domain edits such as adding tattoos or mustaches (see Fig. 1).

3 METHOD

3.1 Implicit Morphable Face Parameterization

We disentangle each facial avatar into identity and expression, where identity is encoded by geometry and color latent codes while expression is captured by an expression latent code. To attain both high-quality geometry and interpretable texture, our model consists of an implicit geometry branch and a UV texture parameterization branch. The geometry branch contains a multilayer perceptron (MLP) that maps 3D points p to SDF values $SDF(p)$ during sphere tracing. The UV texture branch consists of a parameterization MLP that maps p to spherical coordinates $UV(p)$, a parameterization regularizer MLP that learns the inverse mapping from $UV(p)$ back to p, and a color network that predicts the output RGB at $UV(p)$. See Figure 2 for a diagram of our model pipeline. Please refer to the supplement for model architecture details.

We train our model on the Triplegangers [2022] 3D scan dataset for its volume and diversity of subjects and expressions. Although the RenderPeople [2022] dataset additionally models hair and clothing, it only contains 120 neutral expression subjects, making it less suitable for reconstructing an avatar from unconstrained in-the-wild photos. Our training samples consist of a 3D head mesh, UV diffuse texture map, and six diffusely lit frontal RGB images. The dataset contains 515 different subjects each with 20 expressions, for a total of 10,300 data samples. Our full model learns an AutoDecoder dictionary of 515 geometry codes, 515 color codes, and 10,300 expression codes, as subjects express the same sentiment differently. Different expressions for the same training subject share the same geometry and color codes, allowing the model to disentangle expression from the underlying geometry and texture. Please refer to the supplement for examples of our training data.

3.2 Training Losses

Our model is trained on geometry, color, and regularization losses:

$$\mathcal{L} = \mathcal{L}_{geom} + \mathcal{L}_{color} + \mathcal{L}_{reg} \tag{1}$$

Following Figure 2, let f be the SDF MLP, g the UV parameterization MLP, g^{-1} the inverse UV parameterization MLP, and X the set of randomly sampled surface points during training. The geometry loss consists of surface, Eikonal [Gropp et al. 2020], normal, and UV losses. The surface loss ℓ_{surf} optimizes the SDF zero level set, the Eikonal loss $\ell_{eikonal}$ regularizes the SDF gradients, and the normal loss ℓ_{normal} aligns the SDF gradients with the ground truth mesh normals $\hat{n}$. The UV loss ℓ_{uv} regularizes the learned mapping to
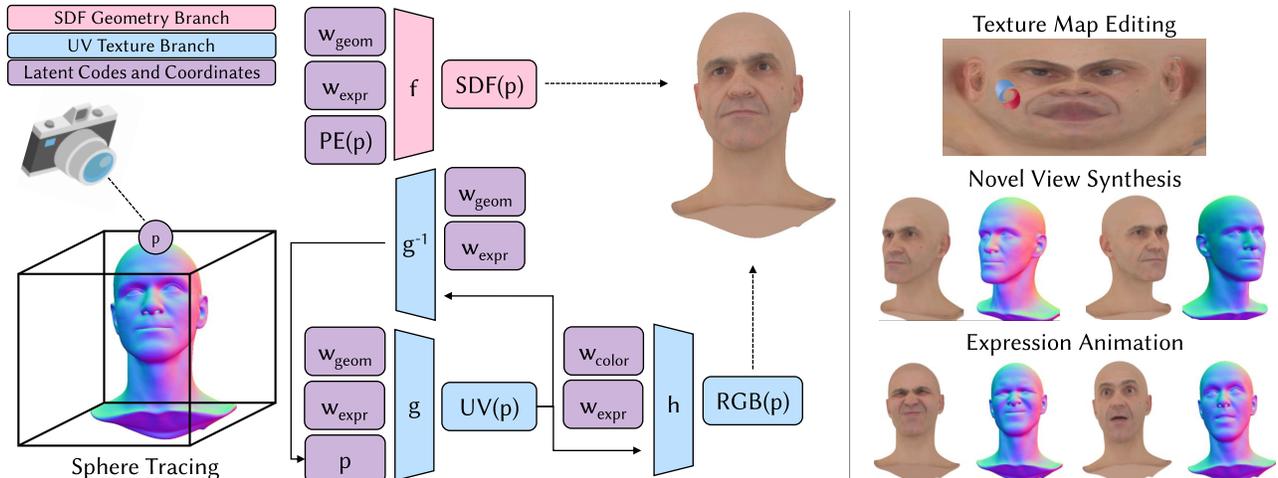
Fig. 2. Our Pipeline. Avatars are represented by geometry, expression, and color latent codes $\{w_{geom}, w_{expr}, w_{color}\}$ with each being 512 dimensional. At each 3D coordinate p during sphere tracing, the SDF network f and UV parameterization network g are conditioned on w_{geom}, w_{expr}, and positional encoding $PE(p)$ to predict the signed distance $SDF(p)$ and UV coordinates $UV(p)$, respectively. The inverse UV parameterization network g^{-1} regularizes the learned mapping to be a surface parameterization $g^{-1}(UV(p); w_{geom}, w_{expr}) = p$, while the color network h predicts the associated RGB texture $RGB(p) = h(UV(p); w_{color}, w_{expr})$. After training, the avatar can be rendered freely with direct control over its texture and facial expression, or extracted as a stand-alone textured mesh asset.

follow an invertible surface parameterization, which enables correspondences between texture and geometry used in our single-shot inversion pipeline, described in Section 3.5.

$$\ell_{surf} = \frac{1}{|X|} \sum_{x \in X} |f(x)| \tag{2}$$

$$\ell_{eikonal} = \mathbb{E}_x(\|\nabla_x f(x)\| - 1)^2 \tag{3}$$

$$\ell_{normal} = \frac{1}{|X|} \sum_{x \in X} \|\nabla_x f(x) - \hat{n}(x)\|^2 \tag{4}$$

$$\ell_{uv} = \frac{1}{|X|} \sum_{x \in X} \|x - g^{-1}(g(x))\|^2 \tag{5}$$

$$\mathcal{L}_{geom} = \ell_{surf} + \ell_{eikonal} + \ell_{normal} + \ell_{uv} \tag{6}$$

The color loss consists of a reconstruction loss ℓ_{tex} on the ground truth texture $\hat{T}$, as well as perceptual [Zhang et al. 2018] and reconstruction losses ℓ_{img} over the facial region I_{face} between the ground truth image $\hat{I}$ and rendered image I obtained via sphere tracing:

$$\ell_{tex} = \frac{1}{|X|} \sum_{x \in X} \|\hat{T}(x) - h(g(x))\|^2 \tag{7}$$

$$\ell_{img} = LPIPS(\hat{I}_{face}, I_{face}) + \|\hat{I}_{face} - I_{face}\|^2 \tag{8}$$

$$\mathcal{L}_{color} = \ell_{tex} + \ell_{img} \tag{9}$$

Finally, we enforce the compactness in the learned latent space by penalizing the magnitude of the geometry, color, and expression codes:

$$\mathcal{L}_{reg} = \|w_{geom}\|^2 + \|w_{color}\|^2 + \|w_{expr}\|^2 \tag{10}$$

3.3   Learning UV Parameterizations

To learn an interpretable texture space and coherent semantic correspondence across subjects, we add an auxiliary loss term to $\mathcal{L}_{reg}$
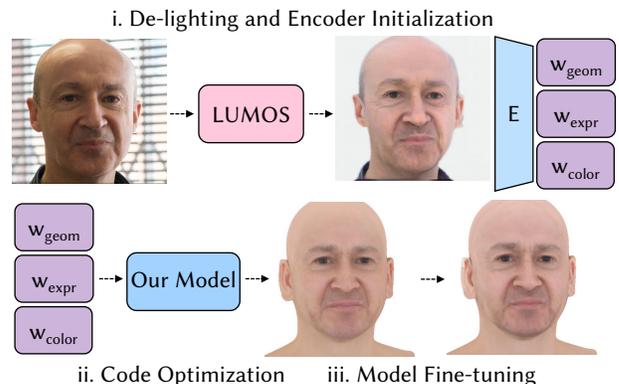

Fig. 3. Single-shot inversion pipeline. We de-light the input image and initialize the latent codes using a pre-trained encoder (top row). We then perform PTI [Roich et al. 2022] to get the final reconstruction (bottom row).

that enforces the parameterization to be consistent through a sparse set of facial landmark constraints:

$$\ell_{landmark} = \frac{1}{|L|} \sum_{x \in L} \|\hat{g}(x) - g(x)\|^2 + \|x - g^{-1}(g(x))\|^2 \tag{11}$$

The first term enforces the learned UV mapping to match the ground truth UV mapping $\hat{g}$ for the set of 3D facial landmark points L, and the second term enforces this mapping to be invertible. Fig. 8 demonstrates the consistency of our learned UV parameterization. Although mostly consistent, it is difficult to obtain perfect registrations around the inner mouth and eyes due to the billboard geometry and errors originating from the ground truth data.
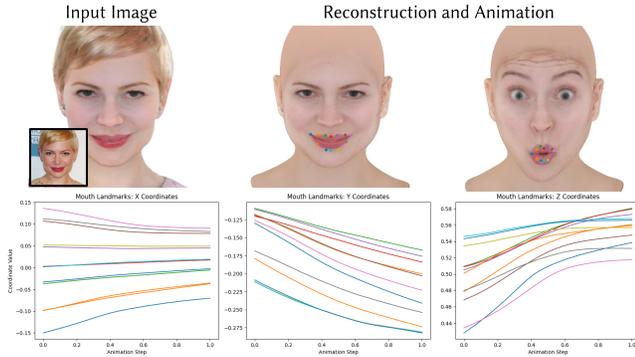
Fig. 4. Non-linear animation space. By linearly interpolating between source and target expression codes, our model exhibits non-linear deformation trajectories on the 3D mouth vertices visualized.

3.4 Animation

After training, an avatar can be animated by manipulating its expression latent code. For a source subject with expression code w_{expr}, target expression code w'_{expr}, and animation timesteps $t \in [0, 1]$, we define the expression animation trajectory by:

$$w_{expr}(t) = w_{expr} + t * (w'_{expr} - w_{expr}) \tag{12}$$

Unlike traditional linear 3DMM approaches, our expression space follows non-linear trajectories learned from high-quality 3D scans, as shown in Fig. 4.

3.5 Single-Shot Inversion

In order to reconstruct and animate unseen subjects, we project an input RGB image into the latent space of our pre-trained model and lightly fine-tune the model weights similar to Pivotal Tuning Inversion (PTI) [Roich et al. 2022]. To handle unseen lighting conditions, we de-light the input image using LUMOS [Yeh et al. 2022] and initialize the geometry, color, and expression codes through a separately trained encoder. We empirically find this encoder initialization to be important in obtaining robust results for in-the-wild input images (See Figure 9).

Image Encoder. We attain latent code initializations by training a DeepLabV3+ [Chen et al. 2018] encoder to reconstruct each training image $\hat{I}$ and its corresponding latent codes $\hat{W}$ already computed from the previous AutoDecoder training stage:

$$\mathcal{L}_{enc} = \|\hat{I} - I\|^2 + \|\hat{W} - W\|^2 \tag{13}$$

$$W = [w_{geom}; w_{color}; w_{expr}] \tag{14}$$

One major challenge when inverting in-the-wild images is handling unseen identities, accessories, hairstyles, and occlusion present in real-world images, as Triplegangers contain limited identities with no variations in hairstyles or background. Therefore, we augment the encoder's training dataset with synthetically augmented Triplegangers images from [Yeh et al. 2022], which improves the robustness of the initialization and final inversion reconstruction, shown in Fig. 9.

Optimization. After initializing the latent codes for an input image $\hat{I}$ using our encoder, we freeze the model weights and optimize

the latent codes while minimizing image, silhouette, multi-view consistency, facial landmark, and regularization losses:

$$\ell_{img} = LPIPS(\hat{I}_{face}, I_{face}) + \|\hat{I}_{face} - I_{face}\|^2 \tag{15}$$

$$\ell_{silhouette} = \sum_{x \in \hat{I}_{face} \wedge x \notin I_{face}} f(x) \tag{16}$$

$$\ell_{ID} = ArcFace(\hat{I}, I, I_{rand}) \tag{17}$$

$$\ell_{landmark} = \sum_{d \in D(\hat{I})} \|d - proj_{2D}(g^{-1}(\hat{d}))\|^2 \tag{18}$$

$$\ell_{reg} = \|w_{geom}\|^2 + \|w_{color}\|^2 + \|w_{expr}\|^2 \tag{19}$$

where the silhouette loss $\ell_{silhouette}$ iterates over points contained in the ground truth face region $\hat{I}_{face}$, but not in the predicted face region I_{face}, to bring the points closer to the SDF zero level set. ArcFace [Deng et al. 2019a] measures the face similarity between different views and I_{rand} is a predicted render from a randomly perturbed camera pose. D is an off-the-shelf facial landmark detector [King 2009] and $\hat{d}$ is the ground truth facial landmark UV mapping enforced in Eq. 11. Note that our consistent UV parameterization directly enables correspondences for the facial landmark alignment loss $\ell_{landmark}$; Fig. 10 demonstrates the benefits of incorporating this loss. The regularization loss ℓ_{reg} is important to ensure that the optimized codes stay near the manifold of the pre-trained latent space for expression animation. We obtain face masks using a pre-trained BiSeNet [Yu et al. 2018] and optimize for 800 steps.

Fine-tuning. To reconstruct finer details in the input image, we freeze the latent codes after optimization and fine-tune the model weights on the above losses. We omit the silhouette loss, as we find it tends to bloat the geometry when the model weights are unfrozen. Although fine-tuning the model improves reconstruction quality, it may also hinder its capability for animation or novel view synthesis. Therefore, we only perform model fine-tuning for 60 steps.

4 RESULTS

We present results of our proposed method with comparisons to EMOCA [Daněček et al. 2022], ROME [Khakhulin et al. 2022] and FaceVerse [Wang et al. 2022b], three recent mesh-based approaches for single-shot 3D avatar generation, and HeadNeRF [Hong et al. 2022], an implicit approach using neural radiance fields. Our method achieves higher fidelity texture and geometry reconstruction in the facial region compared to the baselines. Qualitatively and quantitatively, our method also demonstrates more faithful expression and pose transfer between in-the-wild source and target images. Finally, our learned texture map is intuitive to edit and propagates naturally during animation.

4.1 Implementation Details

Our model is trained in two stages. In the first stage, we withhold the ground truth multi-view images, as we find that supervising with both texture maps and multi-view images negatively impacts the model's ability to learn a consistent UV mapping. In the second stage, we freeze the UV networks $\{g, g^{-1}\}$ and supervise using the multi-view images to fine-tune the learned texture maps while rendering image reconstructions at 768×512 resolution. Camera poses are

| Input Image | De-lit Image | **Ours** | HeadNeRF | ROME | EMOCA | **Ours** | HeadNeRF | ROME | EMOCA | Target Expression and Pose |

Reconstruction

Expression and Pose Transfer
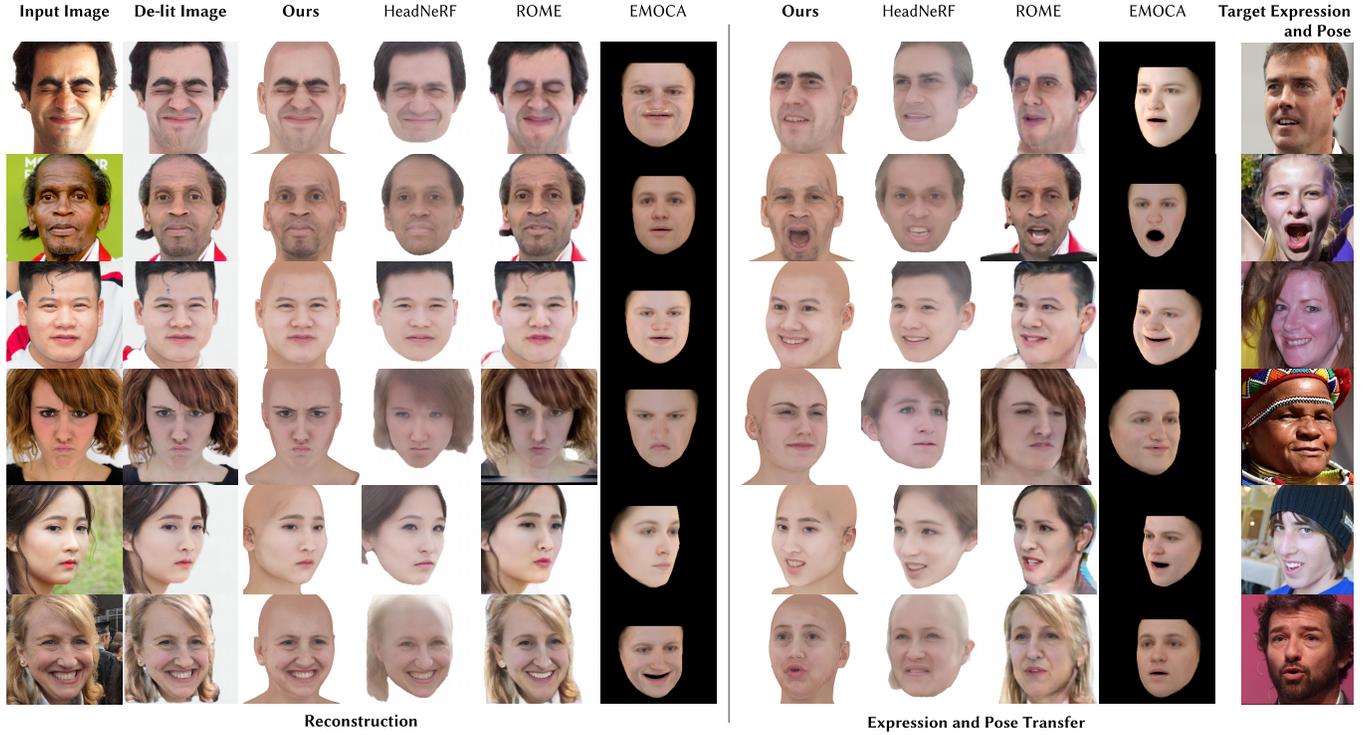
Fig. 5. Single-shot reconstruction on FFHQ with expression and pose transfer. On the left, we show the input FFHQ source image, de-lit input image using LUMOS [Yeh et al. 2022], and reconstruction results for each method. On the right, we show monocular performance capture and retargeting, where we reconstruct and transfer the expression and pose from a target image (right-most column) to the source image identity (left-most column).

Table 2. Quantitative results on single-shot in-the-wild reconstruction (left) and self-expression retargeting (right). **Left**: image, pose, and identity metrics are computed on 500 images sampled from FFHQ. Depth metrics are computed on the H3DS dataset. Image, identity, and depth metrics are computed only on the facial region. EMOCA is evaluated using its smaller face crop. **Right**: FACS coefficients and facial landmarks are computed after expression and pose transfer on 32 expression pairs sampled from the Triplegangers test split.

Reconstruction	LPIPS↓	DISTS↓	SSIM↑	Pose↓	ID↑	L1 Depth↓	RMSE Depth↓
EMOCA	0.1122	0.1268	0.9182	0.0681	0.0697	0.0300	0.0677
ROME	0.1054	0.1130	0.9317	0.0600	0.3866	0.0237	0.0513
HeadNeRF	0.1090	0.1199	0.9268	0.0606	0.2334	0.0379	0.0695
Ours (optimization-free)	0.1427	0.1465	0.9053	0.0549	0.1082	0.0357	0.0658
Ours (encoder-free)	0.0890	0.0921	0.9441	**0.0533**	0.4600	0.0241	0.0527
Ours	**0.0879**	**0.0905**	**0.9451**	0.0563	**0.4670**	**0.0228**	**0.0510**

Retargeting	FACS↓	Facial Landmarks↓
EMOCA	4.712	0.2088
ROME	3.204	0.1414
HeadNeRF	3.848	0.1641
Ours	**1.733**	**0.1165**
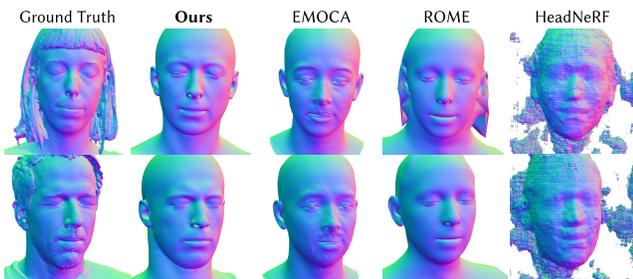
Fig. 6. Ground truth geometry comparison on the H3DS dataset in the single-view setting.

Table 3. Quantitative comparison with FaceVerse [Wang et al. 2022b] on 500 sampled FFHQ images for single-shot in-the-wild reconstruction.

Reconstruction	LPIPS↓	DISTS↓	SSIM↑
FaceVerse	0.1280	0.1119	0.9126
Ours	**0.0879**	**0.0905**	**0.9451**

provided with ground truth training data and we estimate camera poses for in-the-wild FFHQ images using Deep3DFaceRecon [Deng et al. 2019b]. We perform sphere tracing for 50 steps per ray and use a dimensionality of 512 for the geometry, color, and expression latent codes. We train our AutoDecoder for 1000 epochs (approx. one week) and our inversion encoder for 200 epochs (approx. one day)

across 8 NVIDIA A40 GPUs. We use a Triplegangers training/test split of 386/129 for the quantitative expression experiments. Sphere tracing takes 8.5 seconds and inversion takes 3 hours per image. See supplemental material for more details on training and model architectures.

4.2 Single-Shot 3D Face Reconstruction and Animation

Qualitative Results. We show qualitative comparisons for single-shot reconstruction followed by expression and pose transfer on FFHQ [Karras et al. 2019] images between the proposed method, EMOCA, ROME, and HeadNeRF in Fig. 5 and Fig. 13.

Overall, our method is more photo-realistic and achieves higher expression accuracy in facial reconstruction. EMOCA does not model the mouth interior and relies on a pre-trained FLAME [Li et al. 2017] albedo model for texture. Our model produces the most faithful expression transfer, demonstrating the diversity of its learned expression space and generalization capabilities of our method to in-the-wild data. HeadNeRF exhibits a large amount of identity shift during pose transfer, whereas our method remains view-consistent after large pose changes.

We also show a ground truth comparison of reconstructed geometry on the H3DS [Ramon et al. 2021] dataset between our method and the baselines in Fig. 6. HeadNeRF performs volumetric rendering at a low resolution and therefore produces noisy depth results. Our geometry captures higher fidelity facial geometry than ROME and captures the expression more faithfully (e.g., eye blink) compared to EMOCA.

Quantitative Results. We report quantitative reconstruction and self-reenactment expression transfer results in Table 2 and Table 3. The photometric (LPIPS [Zhang et al. 2018], DISTS [Ding et al. 2020], SSIM [Wang et al. 2004]), pose error, and MagFace [Meng et al. 2021] identity consistency (ID) metrics are calculated over a dataset of 500 images from FFHQ. We compute L1 and RMSE depth error over all subjects in the H3DS dataset. To evaluate self-reenactment expression error, we randomly sample 32 source–target expression pairs over a test split of the Triplegangers dataset and measure the L2 error for FACS [Ekman and Friesen 1978] coefficients and facial landmarks. For details related to how each metric is computed, please refer to the supplemental material.

On the FFHQ dataset, our proposed method achieves the best accuracy in terms of LPIPS, DISTS, SSIM, and ID score. The optimization-free ablation struggles to handle the considerably large domain shift between Triplegangers training data and FFHQ in-the-wild images. Our model also exhibits the lowest depth error on the H3DS dataset without relying on a 3D template mesh prior. Finally, our model has the lowest FACS and facial landmark errors, demonstrating the diversity of its learned expression space.

4.3 Ablations

In addition to the baselines mentioned, we compare our method to two ablations for single-shot reconstruction. The first ablation is an optimization-free inversion approach that only uses the learned encoder to directly map an input image to the geometry, color, and expression codes $\{w_{geom}, w_{color}, w_{expr}\}$. The second ablation is an encoder-free inversion approach that omits the encoder and
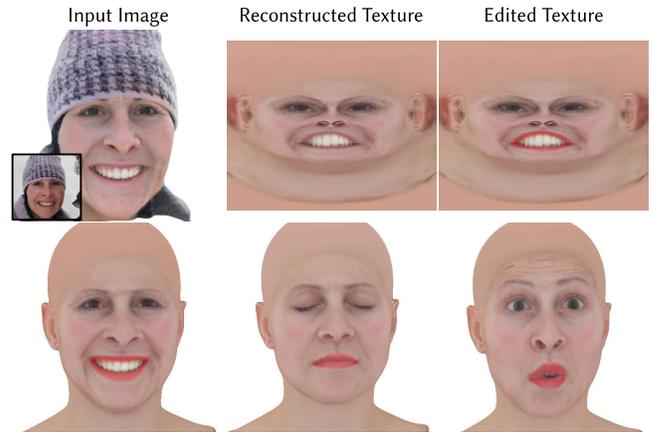

Fig. 7. Texture editing. Top row: input image, learned texture map, and user edited texture map. The learned texture map layout is intuitive and edits propagate naturally during facial animation as shown in the bottom row.

instead uses a mean initialization for $\{w_{geom}, w_{color}, w_{expr}\}$ over the learned AutoDecoder dictionary of latent codes.

Quantitative results for the ablations are reported in Table 2. The optimization-free approach produces significantly worse photometric and depth results, as there is a large domain gap between Triplegangers training data and in-the-wild images; this causes the encoder to produce a coarse reconstruction. The encoder-free approach performs better than the optimization-free approach but is still worse than our full method in image and geometry quality, demonstrating that the encoder initialization improves the optimization reconstruction. Both ablations and our full method perform similarly on pose accuracy.

Applications. As demonstrated in Fig. 5, our method directly supports monocular facial performance capture and expression retargeting. Our hybrid representation provides direct control over an intuitive texture map with a consistent layout. Fig. 7 demonstrates an example workflow: a user reconstructs an input image and modifies the learned texture map. The edits then continue to persist smoothly across different facial animations. Textured meshes can be extracted for further downstream applications such as re-lighting, as shown in the teaser. Fig. 11 and Fig. 12 further demonstrate our model's disentanglement between geometry, texture, and expression with its capability of shape and facial appearance transfer.

5 DISCUSSION

We have presented a new method for reconstructing 3D animatable and textured faces from a single RGB image. The proposed approach combines implicit representations with explicit texture maps to support explicit editing while achieving better photo-realistic rendering, geometry, and expression reconstruction than previous methods. We believe the proposed method makes important contributions towards accessible creation of high-fidelity avatars from in-the-wild images that are animatable, editable, and customizable for downstream applications.

However, there are still limitations to the method. Firstly, the current optimization process during inversion is significantly slower

than encoder-based methods. For real-time applications, more expressive representations such as neural feature fields can be explored to enable optimization-free inversion methods. Furthermore, the method relies on a de-lighting module from Lumos to process in-the-wild images to generate a diffusely lit input image, which may cause subjects to appear paler than expected. These limitations may be alleviated through lighting augmentations of the training dataset to reduce the domain gap and incorporating a lighting model such as spherical harmonics into the representation. Finally, the results shown in this paper do not capture hair or accessories due to limitations of the training dataset. While not perfect, we refer to the supplemental material for a preliminary demonstration of our representation's capacity to handle hair and clothing on the smaller RenderPeople dataset. As implicit representations such as neural radiance fields excel at capturing the geometry and texture of thin structures, it may be fruitful to combine our method with recent sparse view implicit hair models [Kuang et al. 2022; Wu et al. 2022].

REFERENCES

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5461–5470.

ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. 2021. Flame-in-nerf: Neural control of radiance fields for free view face animation. *arXiv preprint arXiv:2108.04913* (2021).

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. RigNeRF: Fully Controllable Neural 3D Portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20364–20373.

Mallikarjun B R, Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Complete 3D Morphable Face Models from Images and Videos. In *cvpr*.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced Dynamic Expression Regression for Real-Time Facial Tracking and Animation. (2014).

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. 2022b. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-Time Facial Animation with Image-Based Dynamic Avatars. (2016).

Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. 2022a. JIFF: Jointly-aligned Implicit Face Function for High Quality Single View Clothed Human Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2729–2739.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16123–16133.

Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. 2020. Semantic deep face models. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 345–354.

Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. 2019. Photo-Realistic Facial Details Synthesis from Single Image. In *iccv*.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5939–5948.

Radek Daněček, Michael J Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20311–20322.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019b. 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*. 0–0.

Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe Gosselin, Marco Romeo, and Louis Chevallier. 2021a. Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 153–164.

Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. 2021b. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12819–12829.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* (2020).

Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. 2017. End-To-End 3D Face Reconstruction With Deep Neural Networks. In *cvpr*.

Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)* 39, 5 (2020), 1–38.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

Haoqiang Fan, Hao Su, and Leonidas J Guibas. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 605–613.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

Pablo Garrido, Levi Valgaert, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. (2013).

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. (2016).

Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In *cvpr*.

Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. 2018. Unsupervised Training for 3D Morphable Model Regression. In *cvpr*.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2022. Learning Neural Parametric Head Models. *arXiv preprint arXiv:2212.02761* (2022).

Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. 2016. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*. Springer, 484–499.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099* (2020).

Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 216–224.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20374–20384.

Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. 2018. Mesoscopic Facial Geometry Inference Using Deep Neural Networks. In *cvpr*.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-Held Video Input. (2015).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. 2021. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4287–4297.

Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*. Springer, 345–362.

Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 2023. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2967–2976.

Zhiyi Kuang, Yiyang Chen, Hongbo Fu, Kun Zhou, and Youyi Zheng. 2022. Deep-MVSHair: Deep Hair Modeling from Sparse Views. In *SIGGRAPH Asia 2022 Conference Papers*. 1–8.

Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction "In-the-Wild". In *cvpr*.

Moran Li, Haibin Huang, Yi Zheng, Mengtian Li, Nong Sang, and Chongyang Ma. 2022. Implicit Neural Deformation for Sparse-View Face Reconstruction. (2022).

Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Kishore, Jun Xing, et al. 2020. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3410–3419.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

Connor Z Lin, David B Lindell, Eric R Chan, and Gordon Wetzstein. 2022. 3D GAN Inversion for Controllable Portrait Image Animation. *arXiv preprint arXiv:2203.13441* (2022).

Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. 2021. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11662–11672.

Li Ma, Xiaoyu Li, Jing Liao, Xuan Wang, Qi Zhang, Jue Wang, and Pedro V Sander. 2022. Neural parameterization for dynamic human head editing. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–15.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. 2021. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14225–14234.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, Hao Li, Richard Roberts, et al. 2018. paGAN: real-time avatars using dynamic textures. *ACM Trans. Graph.* 37, 6 (2018), 258–1.

Chengjie Niu, Jun Li, and Kai Xu. 2018. Im2struct: Recovering 3d shape structure from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4521–4529.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13503–13513.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5620–5629.

Renderpeople. 2022. *Renderpeople*. https://renderpeople.com

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)* 42, 1 (2022), 1–13.

S. Romdhani and T. Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *cvpr*.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2304–2314.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 84–93.

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2017. Photorealistic Facial Texture Inference Using Deep Neural Networks. In *cvpr*.

Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. 2014. Automatic Acquisition of High-Fidelity Facial Performances Using Monocular Videos. (2014).

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems* 32 (2019).

Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2019. FML: Face Model Learning from Videos. In *cvpr*.

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. 2022. Advances in neural rendering. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 703–735.

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2549–2559.

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *iccv*.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. (2019).

J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2016. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *cvpr*.

Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1126–1135.

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7346–7355.

Luan Tran and Xiaoming Liu. 2019. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 157–171.

Triplegangers. 2022. *triplegangers*. https://triplegangers.com

Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. 2017. Regressing Robust and Discriminative 3D Morphable Models With a Very Deep Neural Network. In

cvpr.

Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2626–2634.

Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022a. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022b. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20333–20342.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. 2019. MVF-Net: Multi-View 3D Face Morphable Model Regression. In *CVPR*.

Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. 2018. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 646–662.

Keyu Wu, Yifan Ye, Lingchen Yang, Hongbo Fu, Kun Zhou, and Youyi Zheng. 2022. NeuralHDHair: Automatic High-fidelity Hair Modeling from a Single Image Using Implicit Neural Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1526–1535.

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. 2020. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1–10.

Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. 2021. Neutex: Neural texture mapping for volumetric neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7119–7128.

Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. 2022. High-fidelity 3D GAN Inversion by Pseudo-multi-view Optimization. *arXiv preprint arXiv:2211.15662* (2022).

Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems* 32 (2019).

Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. 2022. GIRAFFE HD: A High-Resolution 3D-aware Generative Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18440–18449.

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-Fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image. (2018).

Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. 2016. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *Advances in neural information processing systems* 29 (2016).

Guandao Yang, Yin Cui, Serge Belongie, and Bharath Hariharan. 2018. Learning single-view 3d reconstruction with limited pose supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 86–101.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–21.

Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12803–12813.

Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Öztireli Cengiz, and Yujiu Yang. 2022. 3D GAN Inversion with Facial Symmetry Prior. *arxiv:2211.16927* (2022).

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 325–341.

Mihai Zanfir, Thiemo Alldieck, and Cristian Sminchisescu. 2022. PhoMoH: Implicit Photorealistic 3D Models of Human Heads. *arXiv preprint arXiv:2212.07275* (2022).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. 2022b. ImFace: A Nonlinear 3D Morphable Face Model with Implicit Neural Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20343–20352.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022a. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition. 13545–13555.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2022c. PointAvatar: Deformable Point-based Head Avatars from Videos. *arXiv preprint arXiv:2212.08377* (2022).

Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, and Simon Lucey. 2017. Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*. 57–65.

Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 2017. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 900–909.
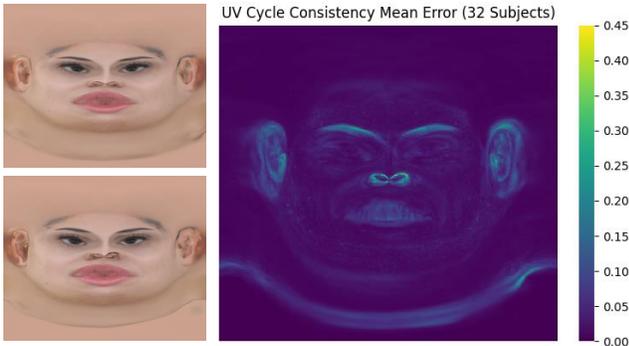
Fig. 8. UV parameterization consistency. We measure the mean L2 error over 32 FFHQ subjects between the learned texture map (top left) and the cycle texture map (bottom left) obtained by mapping from UV → 3D → UV.
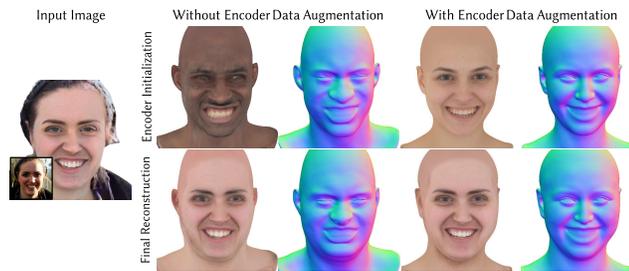
Fig. 9. Encoder training data augmentation ablation. Training the encoder with the synthetically augmented Triplegangers dataset [Yeh et al. 2022] significantly improves our initialization, which is important for converging to a high quality inversion result. Note the difference in the final reconstructed geometry.
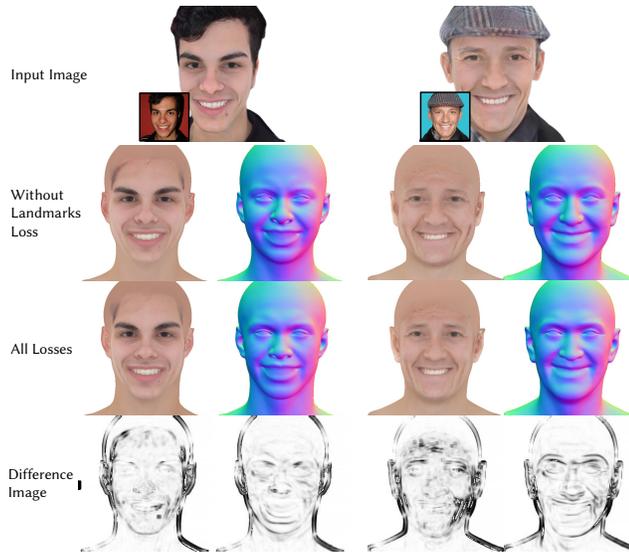
Fig. 10. Facial landmarks loss ablation. Removing the facial landmarks loss during inversion reduces reconstruction quality of the face contour (left and right jaws) and facial features such as the eyes (right).
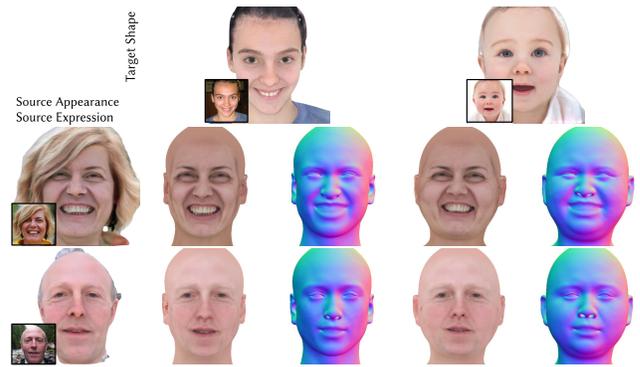
Fig. 11. Shape attribute transfer. We fix the color and expression codes for the source subject and directly replace the source geometry code with the target geometry code.

Fig. 12. Facial appearance attribute transfer. We fix the geometry and expression codes for the source subject and directly replace the source color code with the target color code.

Fig. 13. Zoomed in comparison with ROME [Khakhulin et al. 2022] from Fig. 5. Our model captures the target expression with higher fidelity and higher resolution textures (512×512) compared to ROME (256×256).

Input Image   De-lit Image   Reconstruction   Input Image   De-lit Image   Reconstruction

Fig. 14. Gallery of single-shot reconstruction results on FFHQ.

Single-Shot Implicit Morphable Faces with Consistent Texture Parameterization
Supplemental Material

CONNOR Z. LIN*, Stanford University, USA and NVIDIA, USA
KOKI NAGANO, NVIDIA, USA
JAN KAUTZ, NVIDIA, USA
ERIC R. CHAN*, Stanford University, USA and NVIDIA, USA
UMAR IQBAL, NVIDIA, USA
LEONIDAS GUIBAS, Stanford University, USA
GORDON WETZSTEIN, Stanford University, USA
SAMEH KHAMIS, NVIDIA, USA

AutoDecoder Architecture. The SDF network is a 8-layer MLP with an intermediate feature dimensionality of 512. Similar to NeRF [Mildenhall et al. 2021], we apply positional encoding to its 3D coordinate input using 10 frequency bands and an exponent base of 0.5. The SDF network weights are initialized using the geometric initialization from [Atzmon and Lipman 2020]. The UV parameterization and inverse mapping networks are 6-layer MLPs with an intermediate feature dimensionality of 512. Finally, the color network is an 8-layer MLP with an intermediate feature dimensionality of 512. We apply positional encoding to its spherical parameterization coordinate input using 10 frequency bands and an exponent base of 2.0. During training, we enforce landmark constraints on the first 60 facial landmarks obtained from [King 2009].

Inversion Encoder Architecture. The inversion encoder is a modified DeepLabV3+ [Chen et al. 2018] U-Net model. We remove the batch normalization layers and predict the geometry, color, and expression latent codes by passing its innermost features through 2 convolutional layers with stride 2, followed by a 3-headed linear layer decoder. During inversion, we uniformly perturb the camera pose by [0, 30] degrees in pitch and yaw for the ArcFace [Deng et al. 2019a] face similarity loss.

---

*Work done during an internship at NVIDIA.

De-lighting. To de-light input images for inversion, we provide LUMOS [Yeh et al. 2022] with a white diffuse dome environment to illuminate the input image.

Training Datasets. Fig. S5 shows examples of Triplegangers ground truth training data. Fig. S1 shows data augmentations of Triplegangers images for training our inversion encoder. Statistics about the demographics of the Triplegangers dataset can be found on the Triplegangers website: https://triplegangers.com.

Baseline Details. EMOCA adds an expression shape encoder to DECA [Feng et al. 2021] and encodes the input image into identity, expression, pose, and albedo parameters to reconstruct a coarse 3D mesh. The coarse mesh is then refined by a decoder conditioned on the regressed expression parameters. ROME encodes the input image into a neural texture and uses DECA to predict an initial mesh, which is further refined by an encoder conditioned on the neural texture. The final rendering is achieved by rendering the neural texture using a convolutional U-Net. FaceVerse [Wang et al. 2022] predicts an initial PCA base mesh model and refines its geometry and synthesized texture via optimization of detail and expression latent codes. HeadNeRF conditions a neural radiance field on identity, expression, albedo, and illumination parameters extracted from NL3DMM [Tran and Liu 2018] to reconstruct the input image. For all baselines, we used official pre-trained model and source code released by the authors.

Metric Details. For each FFHQ image, we perform a 512×512 resolution crop and estimate its pose using a pre-trained Deep3DFace [Deng et al. 2019b] model. We then obtain a face mask using a pre-trained BiSeNet [Yu et al. 2018]. The LPIPS, DISTS, and SSIM scores are computed between the facial regions of the 500 source and reconstructed images. For the pose error and identity consistency score, we randomly sample a camera pose following the FFHQ distribution for each of the 500 source images. The pose error is given by the L2 Euler angles error between the sampled novel view and the novel view reconstructed image, where pose is computed using Deep3DFace. Identity consistency is computed using dot product similarity of MagFace features extracted from the masked source image and masked novel view reconstructed image.

To compute the depth metrics, we sample a random frontal view from each subject in H3DS and render the ground truth depth map at 512×512 resolution. We then report the L1 and RMSE depth errors in the reconstructed facial region for each method.

For self-reenactment facial transfer, we evaluate on a separate version of our model trained on 386 out of the 515 total Triplegangers subjects, as FFHQ is not suitable for self-reenactment (no repeated subjects) and VoxCeleb [Nagrani et al. 2017] has lower image resolution. We randomly sample 32 source-target expression pairs over the test split and transfer the target pose and expression to the source image for all methods. We then extract FACS coefficients using an off-the-shelf facial expression tracker [NVIDIA 2022] and 68 facial landmarks using dlib [King 2009] for each ground truth target image and its corresponding source image after pose and expression transfer. The self-reenactment facial transfer error is given by the L2 distances for the FACS coefficients and facial landmarks.

Asset Extraction. To extract 3D textured meshes, we run Marching Cubes [Lorensen and Cline 1987] at a 512^3 grid resolution to extract the zero level set of the learned SDF. We query the parameterization network at each mesh vertex to obtain its corresponding UV coordinates, and extract a texture map by sampling the parameterization and color networks on a 1024×2048 grid of UV coordinates. Fig. S2 shows how DCC tools like Blender can be used in tandem with our method to re-light and customize avatars.

Normalized Avatar Generalization. Our method can also be used for face normalization to recover a neutral expression avatar from a single-shot in-the-wild image, as shown in Fig. S3.

Stylized Portrait Images. Fig. S4 illustrates our model's ability to create avatars from out-of-domain stylized portrait images.

RenderPeople Results. Flexible toplogy is one of the key advantages of implicit representations over mesh-based approaches. In Fig. S6, we demonstrate our representation's ability to model hair and clothing on the RenderPeople [2022] dataset.

Societal Impact. This work enables the synthesis, animation, and editing of photo-realistic 3D face avatars from single-shot images. However, this research also has potential for misuse and we condemn its use for any nefarious purposes; we refer the reader to Tewari et al. [2020] for an in-depth discussion of forgery detection.
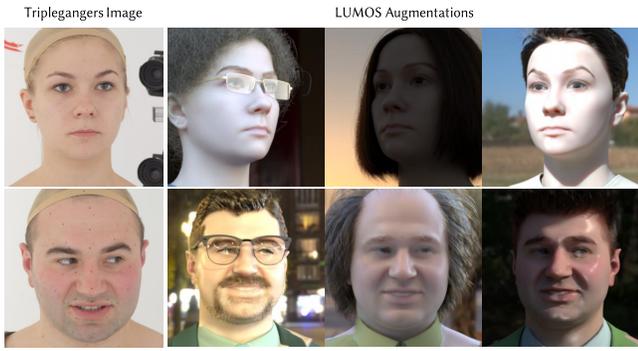
Fig. S1. Our inversion encoder is trained by the synthetic dataset from LUMOS [Yeh et al. 2022] that augments the Triplegangers dataset. The synthetic variations include different lightings, hairstyles, accessories, and camera poses and help make the encoder initialization robust to in-the-wild images. Please refer to LUMOS [Yeh et al. 2022] for the details of the synthetic dataset.

Fig. S2. The extracted assets (head mesh and texture) can be rendered in DCC tools like Blender for re-lighting and avatar customization with hair card models. Please see teaser for the extracted head meshes without hair.
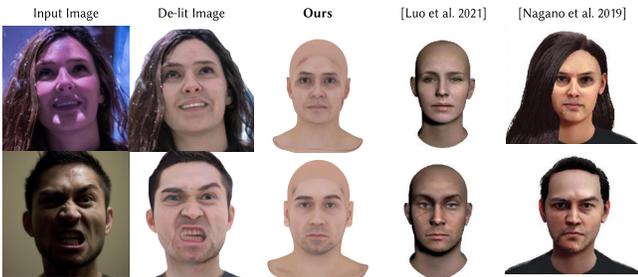

Fig. S3. Normalized avatar generation. From left to right, we show the input image, de-lit image, our avatar with expressions neutralized, state-of-the-art normalized avatar generation methods by [Luo et al. 2021], and by [Nagano et al. 2019]. To perform expression neutralization, we replace the recovered input expression code with a neutral expression code from a Triplegangers training subject.
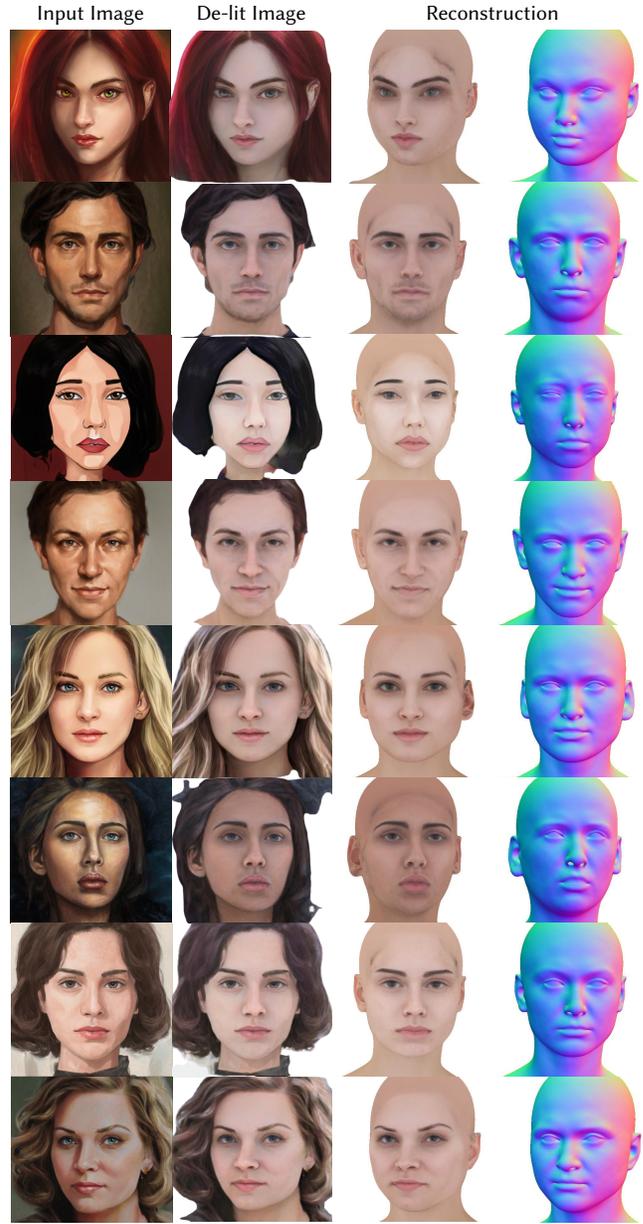

Fig. S4. Single-shot 3D reconstruction results on stylized portrait images. While our model is never trained with stylized portrait images, it can handle them reasonably well.
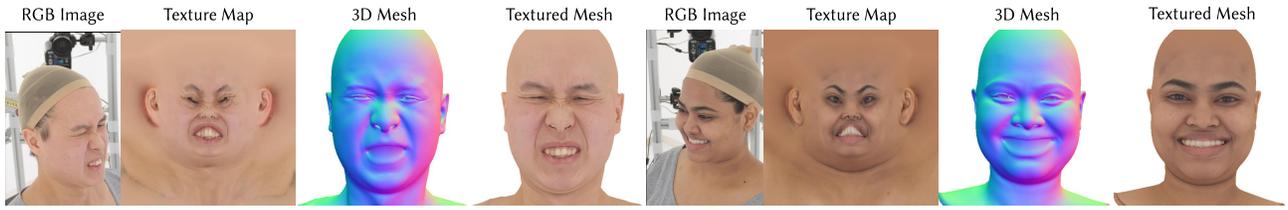
Fig. S5. Examples of training data from the Triplegangers dataset.

Fig. S6. Reconstruction results for training samples on the RenderPeople dataset.

REFERENCES

Matan Atzmon and Yaron Lipman. 2020. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2565–2574.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019a. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019b. 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*. 0–0.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

William E Lorensen and Harvey E Cline. 1987. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics* 21, 4 (1987), 163–169.

Huiwen Luo, Koki Nagano, Han-Wei Kung, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. 2021. Normalized avatar synthesis using stylegan and perceptual refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11662–11672.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. 2019. Deep face normalization. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

NVIDIA. 2022. *NVIDIA MAXINE AR SDK*. https://github.com/NVIDIA/MAXINE-AR-SDK

Renderpeople. 2022. *Renderpeople*. https://renderpeople.com

Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. 2020. State of the Art on Neural Rendering. *Eurographics Association* (2020).

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7346–7355.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022. FaceVerse: a Fine-grained and Detail-controllable 3D Face Morphable Model from a Hybrid Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20333–20342.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–21.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 325–341.