HeadCraft: Modeling High-Detail Shape Variations for Animated 3DMMs

Artem Sevastopolsky[1] Philip-William Grassal[2] Simon Giebenhain[1]
ShahRukh Athar[3] Luisa Verdoliva[4,1] Matthias Nießner[1]

[1] Technical University of Munich (TUM), Germany [2] Copresence AG, Germany
[3] Stony Brook University, US [4] University of Naples Federico II, Italy

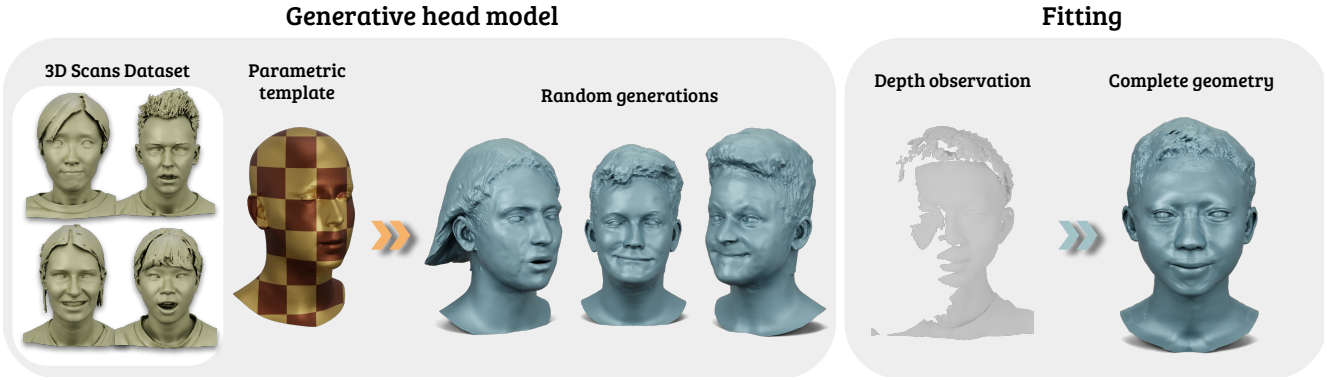

Figure 1. We present HeadCraft, a generative model for highly-detailed human heads, ready for animation. Our method is trained on 2D displacement maps collected by registering a parametric template head with free surface displacements to a large set of 3D head scans. The resulting model is highly versatile and its latent code can be fit to an arbitrary depth observation.

Abstract

Current advances in human head modeling allow to generate plausible-looking 3D head models via neural representations, such as NeRFs and SDFs. Nevertheless, constructing complete high-fidelity head models with explicitly controlled animation remains an issue. Furthermore, completing the head geometry based on a partial observation, e.g. coming from a depth sensor, while preserving a high level of detail is often problematic for the existing methods. We introduce a generative model for detailed 3D head meshes on top of an articulated 3DMM which allows explicit animation and high-detail preservation at the same time. Our method is trained in two stages. First, we register a parametric head model with vertex displacements to each mesh of the recently introduced NPHM dataset of accurate 3D head scans. The estimated displacements are baked into a hand-crafted UV layout. Second, we train a StyleGAN model in order to generalize over the UV maps of displacements, which we later refer to HeadCraft. The decomposition of the parametric model and high-quality vertex displacements allows us to animate the model and modify the regions semantically. We demonstrate the results of unconditional sampling, fitting to a scan and editing. The code and data are available at https://seva100.github.io/headcraft.

1. Introduction

The ability to create lifelike 3D head models is crucial for many applications, ranging from video game character design to virtual try-on experiences and medical simulations. 3D Morphable Models (3DMMs) [6] constitute an essential and robust tool for basic 3D head geometry estimation and tracking that enables its further reconstruction and animation. Furthermore, 3DMMs, along with their consistent UV parameterization of the surface, are commonly used in constructing virtual avatars as an approach for approximate surface representation or regularization [4, 22, 33, 60].

Constructing (neural) 3DMMs capable of representing the diverse distribution over human heads while disentangling identity and expressions and capturing a high degree of detail is challenging. At the same time, such neural representation would ideally be compatible with well-established methods for animation and tracking, which are usually mesh-based and are required to fulfill certain real-time constraints.

Despite the recent progress of implicit representations, such as neural radiance fields (NeRFs) [36] and signed distance functions (SDFs) [38], the most prominent 3DMMs [6, 32, 40] are based on a template mesh (i.e. feature explicit geometry), and principal component analysis (PCA), to represent identity and expression varia-

tions. Mesh-based 3DMMs can usually be robustly fitted to videos, easily animated and integrate well with established graphics pipelines. However, these approaches are fundamentally limited by both the mesh resolution and representational capacity of its underlying (multi-)linear statistical model. In our work, we improve on both of these aspects and leverage recent advances in generative image modeling by using StyleGAN2 [27] to predict highly detailed geometry in the UV space, which is independent of the mesh resolution.

A different line of work attempts to build 3DMMs based on neural SDFs instead of meshes (e.g. [20, 56]), thereby enabling reconstruction at arbitrary resolutions. However, the incorporation of such implicit representations is limited due to a lack of compatibility with standard graphics systems and animation tools. Furthermore, SDF-based approaches often require costly evaluations, e.g. via marching cubes [35], to extract an implicit surface, which can hinder the real-time application of these methods.

Inspired by the combination of these ideas, in this research, we introduce a generative model that allows for animation and tracking and preserves a high level of detail. At the heart of our approach lies the idea of combining an explicit parametric head model (FLAME [32]) with surface displacements complementing the low geometry detail of the head model. FLAME is an example of a 3D Morphable Model [6, 15] with a fixed set of vertices and fixed topology, constructed as a linear statistical model over the heads with point-to-point correspondence and further controlled by shape and expression latent codes. We register a highly subdivided FLAME mesh template with free vertex displacements to all 3D head scans in the NPHM [20] dataset to obtain the necessary training data. To facilitate as high level of detail in the displacements as possible, they are fitted in two steps. First, the optimization problem is solved for vector displacements with strong regularization that penalizes very hard for self-intersections of the deformed mesh regions. Afterwards, a separate optimization step refines the displacements only along the normals of the deformed vertices while keeping the regularization weight low. These displacements are baked into a predefined UV layout. Finally, we train a StyleGAN2 [27] model to generalize over this set of baked 2D displacement maps. This novel architecture allows us to operate at a resolution higher than the conventional FLAME template, enabling the generation of highly detailed and animatable 3D head models.

To validate the efficacy and practical utility of our approach, we evaluate it in several settings. The diversity and fidelity of the generated 3D head meshes are quantitatively and visually compared to other methods w.r.t. the real head scans from the FaceVerse dataset [51], both in UV space and rendered image space. We also explore the applicability of our approach in fitting the latent representation of the generative model to complete or incomplete point cloud data and demonstrate its animation and manipulation capabilities.

To summarize, our contributions are as follows:

- We introduce a two-stage registration procedure to craft high-detail displacement maps on top of 3DMMs from 3D scanning data. This enables the application of 2D generative models to tackle the 3D generative task.
- We propose a generative model operating in the displacement maps domain to enhance the low-frequency geometry of FLAME with details and extend its shape space to a range of variations.
- We demonstrate the versatility of our method through unconditional sampling, interpolation, semantic geometry transfer, and 3D completion based on partial depth observations.

2. Related Work

Many recent solutions to computer vision problems involving human bodies and heads are built on statistical body models, forming the foundation for building personalized avatars [1, 2, 19, 22, 60], motion tracking [17, 47], scan registration [20], controlling image synthesis [46], and more. Their line of research is divided into two major branches.

Mesh-based Models. Pioneering work in the field [6] proposed 3D morphable models (3DMMs) for the human faces' identity, expression, and appearance representation. Their model is built around a 3D template mesh and linear parametric blendshapes derived from PCAs over 3D scan data. With new datasets and registration procedures, their work has been extended from faces to heads [31, 32, 42], hands [45], full bodies [34], or combinations of these [39, 53]. The template mesh has a fixed topology which provides consistent UV unwrapping and enables fitting to know surface correspondences, e.g. semantic regions and landmarks. Yet, it limits the representative power w.r.t. the overall shape and level of detail beyond what the template provides. Downstream approaches compensate this by optimizing displacements [1, 2, 8, 22, 28, 31, 55] or additional implicit geometry [9, 18] on top of the mesh. Displacements are applied either per-vertex individually [1, 2, 22, 28] or as a displacement map over the whole surface using the consistent UV unwrapping of the template [17, 31, 55]. Some approaches infer the displacement maps from images or texture reprojections for specific individuals [17, 55]. By making use of high-quality scans, the authors of [31] demonstrate that GANs [21, 25, 27] for image generation can be leveraged to learn a generative model over displacement maps in the face region. Our method takes this idea further by learning a generative model for displacements over the whole head. It showcases that even the surface geometry of large and complicated hairstyles can be represented with high fidelity. The resulting outputs of our full-head

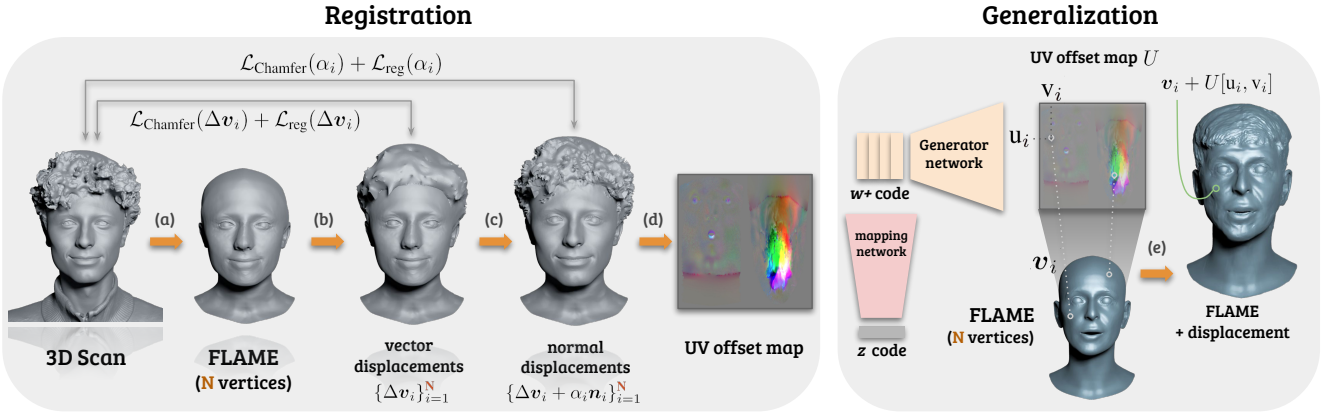

Figure 2. An overview of the method. In the registration stage, we **(a)** fit the FLAME template by the face landmarks to the scan from the NPHM dataset and highly subdivide it, **(b)** optimize for the vertex displacements in $\mathbb{R}^3$ to fit the rough geometry with strong regularizations, **(c)** optimize for the scalar refinements of the displacements along the normal directions, and **(d)** bake the displacements into a UV offset map. To generalize over the UV offset maps, we train a StyleGAN2 [27] model. After training, the offsets can be applied to an arbitrary FLAME template by subdividing it and **(e)** querying the generated UV offset map with the (u, v) locations of the FLAME vertices.

generative model provide quality approaching the scan level while exploiting the UV surface correspondences of the underlying 3DMM and enabling further animation through the rigging of the template.

Implicit Models. The recent success of implicit SDFs [38] and neural radiance fields [36] in 3D modeling has also motivated applying them for statistical body models. Most implicit models learn shape and appearance in a canonical reference space [3, 20, 24, 43, 56] or as displacements on top of an existing model [57]. For better generalization and detail preservation, some approaches use a composition of implicit SDFs to model the canonical space [3, 20]. Articulation and animation are modeled either directly in canonical coordinates [24, 57], through implicit deformation fields [20, 56], explicit joints [3] or blendshape deformations borrowed from explicit 3DMMs [52, 59]. While the aforementioned methods rely on multi-view data and aligned 3D scans, a separate line of research demonstrates that statistical shape and appearance priors can also be learned from unstructured image collections [10, 11, 13, 37, 44].

Implicit approaches do not rely on topology and shape templates. This allows them to fit more detail and complex shapes such as hair [20, 56, 57] and even glasses [11, 37]. Yet, it prevents consistent surface correspondences between different samples, which need to be explicitly learned [3, 57]. As our approach uses a mesh-based template, it does not suffer from these limitations and has an explicit model for animation. Still, we are able to show that we can provide a comparable level of detail as implicit methods and also model hair surface geometry which has not been achieved with a generative, explicit shape model before.

3. HeadCraft

We cast the task of learning high-fidelity distribution over 3D heads as a 2D generative problem by leveraging the UV-space of an existing 3DMM [32]. This allows us to rely on the well-explored body of research on CNN-based 2D generative models [26, 27].

To this end, we register a dataset of high-end 3D head scans [20] in the FLAME head model [32] topology and bake the details into a displacement map in UV space, as explained further in Sec. 3.1. Subsequently, in Sec. 3.2, we train a 2D generative model to learn the distribution over the displacement maps, which can be converted into detailed 3D displacements compatible with the underlying FLAME model. An overview of HeadCraft is presented in Fig. 2.

3.1. Displacements registration procedure

The purpose of this step is to create a dataset of 2D displacement maps representing the geometric information beyond the representational capacity of classical 3DMMs. To this end, we compute displacements from a FLAME [32] mesh to the high quality 3D head scans from the NPHM dataset [20]. Let us consider a scanned mesh $\mathcal{P} = (V^{\text{gt}}, \mathcal{F}^{\text{gt}})$ with vertices $V^{\text{gt}} \in \mathbb{R}^{|V^{\text{gt}}| \times 3}$, and faces $\mathcal{F}^{\text{gt}} \in \{1, \dots, |V^{\text{gt}}|\}^{|\mathcal{F}^{\text{gt}}| \times 3}$.

In order to find the appropriate FLAME parameters for the scan, we follow the rigid alignment optimization procedure outlined in the NPHM work [20]. This procedure requires face landmarks to be known, which can be annotated manually or, as provided with the dataset in our case, calculated via 2D face landmark detectors on the projections of the colored scans and lifted to 3D. This way, we obtain a FLAME template, corresponding to the given scan, and subdivide it via Butterfly algorithm [14]. We will refer to the template after subdivision as to $F = (V, \mathcal{F}, \mathcal{U}_{\mathcal{F}})$, where

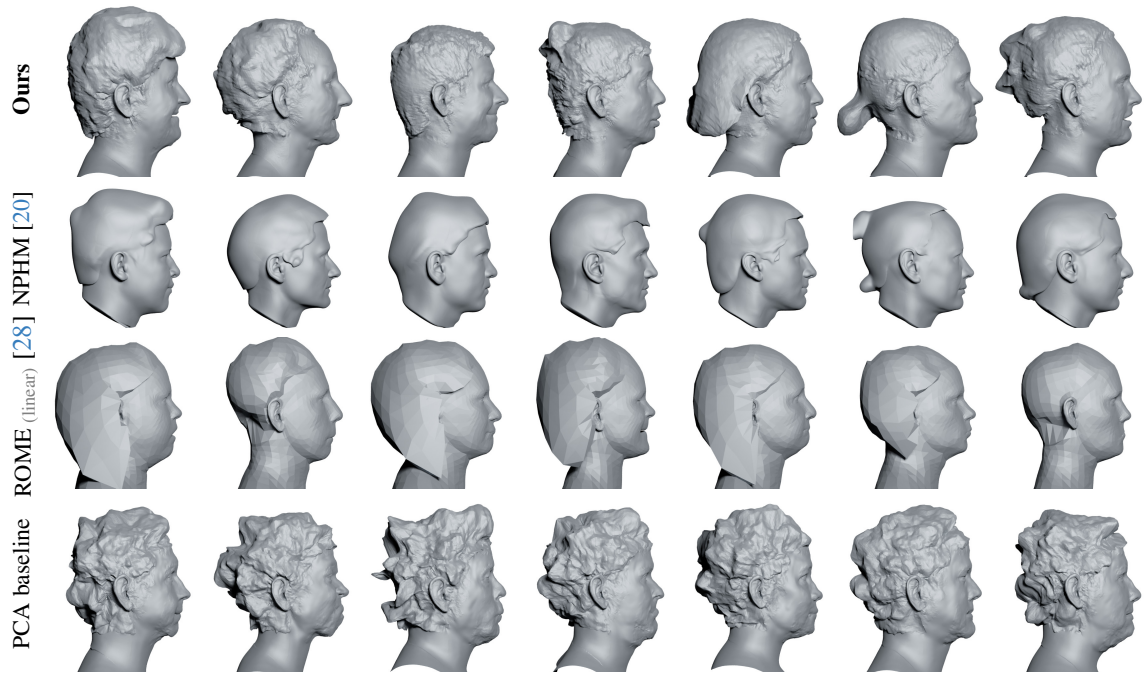

Figure 3. Visual comparison of fidelity and diversity of the meshes generated by various methods. For *Ours*, random FLAMEs are sampled from Gaussian distribution with statistics calculated over the NPHM dataset; same for the *PCA baseline* pre-fitted to our UV registrations. Meshes from *NPHM* are obtained by sampling the latent codes and running marching cubes over the generated SDF representations. We demonstrate higher variability of produced head geometry and better details than the other methods.

$V \in \mathrm{R}^{|V|\times 3}$ are the vertices coordinates, $\mathcal{F} \in \mathrm{R}^{|\mathcal{F}|\times 3}$ are the corresponding faces, and $\mathcal{U}_\mathcal{F} \in \mathrm{R}^{|\mathcal{F}|\times 3\times 2}$ are the texture coordinates of each vertex in a triangle. Note that using triangle coordinates instead of vertex coordinates is important due to the presence of a seam in the FLAME model, thus making UVs for the seam vertices ambiguous.

As FLAME basis does not represent hair or face details, we define these in a form of vertex displacements and learn them in two stages. During the first stage, we optimize the loss function $\mathcal{L}_{\text{Stage 1}}(D)$ for additive vector displacements $D_{\text{Stage 1}} \in \mathbb{R}^{|V|\times 3}$ of the vertices:

$$\begin{aligned}\mathcal{L}(D, V, \mathcal{F}, V^{\text{gt}}|\boldsymbol{\lambda}) = \lambda^{\text{Chamfer}}\,\mathcal{L}_{\text{Chamfer}}(V+D, V^{\text{gt}}) \\ + \lambda^{\text{edge}}\,\mathcal{L}_{\text{edge}}(V+D, \mathcal{F}) \\ + \lambda^{\text{lapl}}\,\mathcal{L}_{\text{lapl}}(V+D, \mathcal{F})\end{aligned} \tag{1}$$

$$\mathcal{L}_{\text{Stage 1}}(D) = \mathcal{L}(D, V, \mathcal{F}, V^{\text{gt}} \,|\, \boldsymbol{\lambda}_{\text{Stage 1}}) \tag{2}$$

Hyperparameters $\boldsymbol{\lambda}_{\text{Stage 1}} = (\lambda^{\text{Chamfer}}_{\text{Stage 1}}, \lambda^{\text{edge}}_{\text{Stage 1}}, \lambda^{\text{lapl}}_{\text{Stage 1}})$ define the Chamfer matching term weight, the weight of edge length regularization and standard Laplacian regularization. In this stage, the weight of regularizations is high in order to prevent self-intersections that can occur when regressing vector displacements. Also, we only optimize the vector displacements for the hair region.

In the second stage, we optimize the loss function $\mathcal{L}_{\text{Stage 2}}(\boldsymbol{\alpha})$ for displacements $D_{\text{Stage 2}} \in \mathbb{R}^{|V|\times 3}$ that are only allowed to move over the normals of the previously displaced vertices:

$$D_{\text{Stage 2}} = D_{\text{Stage 1}} + N \odot \boldsymbol{\alpha}, \tag{3}$$

where $N \in \mathbb{R}^{|V|\times 3}$ corresponds to the normals, calculated by numerical difference for vertices deformed after the Stage 1, and $\odot$ defines the element-wise product of rows of N and elements of α (each normal $\boldsymbol{n}_i$ is multiplied by the respective amplitude α_i). $\mathcal{L}_{\text{Stage 2}}(\boldsymbol{\alpha})$ is expressed through the same basic loss expression:

$$\mathcal{L}_{\text{Stage 2}}(\boldsymbol{\alpha}) = \mathcal{L}(D_{\text{Stage 1}} + N \odot \boldsymbol{\alpha}, V, \mathcal{F}, V^{\text{gt}} \,|\, \boldsymbol{\lambda}_{\text{Stage 2}}), \tag{4}$$

while hyperparameters $\boldsymbol{\lambda}_{\text{Stage 2}}$ are selected with relatively lower regularization weights. This allows for fitting high-frequency details while maintaining the same rough shape of the regressed shape. At this stage, we allow both hair and face regions to deform, while subtle parts such as ears and eyeballs are fixed from moving.

Finally, we bake the displacements $D_{\text{Stage 2}}$ into a UV map $U \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ by rendering it onto the UV space with known texture coordinates $\mathcal{U}_\mathcal{F}$ and triangles $\mathcal{F}$.

The registration procedure is repeated for the dataset consisting of multiple 3D scans, resulting in a set of UV displacement maps $(U_1, \ldots, U_S)$.

3.2. Generative model

The described registration procedure allows us to relax the problem of 3D head geometry generation into a problem of generation of 2D UV displacement maps, which allows us to apply a 2D generative model. We have selected StyleGAN2 for that purpose due to its capability of generalizing over relatively small datasets of images [26, 41, 58] while maintaining close-to-SoTA image generation capabilities [27]. The model consists of a mapping network and a generator network, which we will refer to as $f(z)$ together, where $z \in \mathcal{Z} \subset \mathbb{R}^{\mathrm{D}}$ is a latent code sampled from a standard normal distribution during training (with truncation trick [7] at the inference time). The generator produces a UV displacement map $U = f(z)$, which we can apply to an arbitrary (anyhow densely subdivided) FLAME template $F = (V, \mathcal{F}, \mathcal{U}_{\mathcal{F}})$ by querying the map U with its texture coordinates $\mathcal{U}_{\mathcal{F}}$ to obtain the respective vertex displacements. The final mesh $M = (V + D(U), \mathcal{F}, \mathcal{U}_{\mathcal{F}})$ is different from the subdivided FLAME only in terms of the vertex locations. We later demonstrate visually that the generated displacements could be applied to an arbitrary template.

UV layout. Importantly, we modify the UV embedding of the FLAME template mesh into a 2D plane (see the Supplementary for the illustration). Doing so results in a more favorable layout for the 2D generative model with a limited receptive field. As we demonstrate further, the large distance in the UV space of the head's back left and right parts leads to inconsistent generations. The new layout allows us to stack the left- and right-hand sides into a single-channel displacement map with better seam alignment.

Post-processing. Since the UV map U is generated in the UV layout that contains a seam, we expect StyleGAN to resolve it in general, i.e. produce similar displacements in the face and scalp region near the same seam vertex. Still, there is no dedicated supervision during StyleGAN training that ensures that it always happens and that the border is preserved pixel-perfect. Because of that, we apply Laplacian smoothing [50] to the mesh M in the K-vertex vicinity of the seam. Additional smoothness of the face region is achieved by applying Laplacian smoothing to the facial skin, neck, scalp, eyeballs, and inner mouth region with different strength to the subdivided FLAME template in advance, before adding the displacements. Technical details are provided in the Supplementary Material.

4. Experiments

We evaluate HeadCraft's generative capabilities by examining its unconditional generation performance in Sec. 4.4 and ablate several important aspects in Sec. **??**. Additionally, in Sec. 4.5, we demonstrate the way HeadCraft can be beneficial in several important downstream applications, such as 3D head completion from a partial point cloud, its seamless integration with the FLAME expression space, and semantic geometry transfer. Before presenting these results, we describe implementation details, as well as our chosen metrics and baselines in Sec. 4.1, 4.2, and 4.3, respectively.

4.1. Implementation details

Datasets. Our method is trained on the 3D real scans of human heads from NPHM [20] dataset, namely, 6975 high-resolution scans of 327 diverse identities captured by two 3D scanners each. To quantitatively evaluate the generative performance of all compared methods, we use FaceVerse [51] dataset as a source of the ground truth scans not intersecting with our training data by identities.

Registration procedure. We use Adam optimizer with learning rate of $3 \cdot 10^{-2}$ for the first stage and $3 \cdot 10^{-4}$ for the second stage. The hyperparameters $\boldsymbol{\lambda}_{\text{Stage 1}} = (\lambda^{\text{Chamfer}}_{\text{Stage 1}}, \lambda^{\text{edge}}_{\text{Stage 1}}, \lambda^{\text{lapl}}_{\text{Stage 1}})$ equal to $(2 \cdot 10^3, 2 \cdot 10^5, 10^4)$. For the second stage, $\boldsymbol{\lambda}_{\text{Stage 2}} = (2 \cdot 10^4, 2 \cdot 10^4, 10^4)$. In the Chamfer loss, we additionally apply correspondences pruning by distance of 1.0, which defines that all the correspondences between source and target with the distance more than 1.0 in the NPHM coordinate system are automatically discarded. This has been introduced for more consistent gradual learning of displacements, such that at each optimization step, only the nearest points affect the deformation learning.

Generative modeling. As the 2D generative model, we choose StyleGAN2-ADA [26] with all augmentations turned off (since they wouldn't yield valid UV maps in our case) and 8 mapping network layers. To stablize the GAN training, we utilize a high gradient penalty of 4.0 for the discriminator. The learning rates are $2 \cdot 10^{-3}$ for the generator and $1 \cdot 10^{-3}$ for the discriminator. We train it for 95K steps with the batch size of 8 and 256×256 image resolution.

4.2. Metrics

To evaluate the quality of the generated samples in section 4.4, we rely both 2D metrics computed on renderings, as well as, on metrics directly compute in 3D. In the following we describe all employed metrics in details.

Firstly, to evaluate the visual plausibility of the generated

	FID ↓	KID ↓	MMD ↓	JSD ↓	COV ↑
Ours	**68.00**	**0.065**	**6.51**	21.33	**53.85%**
PCA	126.31	0.165	10.47	20.16	23.08%
NPHM	139.82	0.170	7.80	**19.06**	46.15%
ROME	169.65	0.204	10.02	23.19	32.69%
FLAME	198.85	0.262	12.95	23.89	5.77%

Table 1. The comparison of quality and diversity of random samples generated by each of the methods. FID and KID measure the similarity of the generated mesh renderings vs. the renderings of the ground truth meshes in FaceVerse dataset, while 3D metrics MMD, JSD, COV assess the similarity of generated and real point clouds distributions. MMD is multiplied by 10^3 and JSD by 10^2.

geometry, we render 2195 ground truth meshes from FaceVerse and the same number of meshes generated by each method with highly metallic material from eight distinct viewpoints, uniformly sampled along the circular trajectory in the horizontal plane. The FID [23] and KID [5] perceptual metrics are calculated for all generated and ground truth renderings from a given viewpoint and then macro-averaged over eight viewpoints.

Secondly, we compare the distributions of point clouds sampled from the generated and ground truth meshes. To do that, we sample 10K points from each of the 2195 generated and the same number of ground truth meshes and calculate several 3D similarity metrics. Jensen-Shannon Divergence (JSD) is evaluated by comparing the distributions of generated and ground truth points, splat into a voxel grid (in our case, of 512^3 voxels). Minimum Matching Distance (MMD) is a measure of 3D object realism that, for each ground truth sample, involves evaluating the distance to the most similar sample in the generated set. Similarly, Coverage (COV) indicates the percentage of the ground truth samples, for which the nearest neighbor among all ground truth and generated samples falls into the generated set. More detailed description of the 3D metrics can be found in [16, 54].

4.3. Baselines

We compare against the recent state-of-the-art head model NPHM [20], which uses neural SDFs to represent the head geometry. We additionally compare against ROME [28], an alternative approach that models complete head geometry including hair using a FLAME template mesh. ROME is trained in an unsupervised fashion on the large-scale video dataset VoxCeleb2 [12].

Furthermore, we compare against a PCA-based baseline, whereas a linear PCA basis is fitted to our UV displacement maps, and provide the numbers for random FLAME samples without added displacements as a reference.

While NPHM, PCA baseline, and Ours have been fitted to exactly the same training dataset, for ROME, the checkpoint from the public repository has been used.

4.4. Results

Unconditional sampling. In Fig. 3, we compare the difference in details and diversity of the unconditional samples produced by our method to the ones produced by NPHM [20] and ROME [28] methods. For Ours, PCA baseline and ROME, a FLAME with random shape, expression and jaw parameters are sampled from normal distribution for every head mesh, in accordance with the statistics precalculated over the NeRSemble dataset [30]. For ROME, we sample the FLAME displacements from a linear model provided by the authors of ROME as the sampling strategy proposed by the ROME authors. Visually, we observe both higher diversity and better representation of details than for all baselines. The details of the facial region are generally the sharpest for ROME, PCA baseline, and Ours, due to the use of the FLAME template.

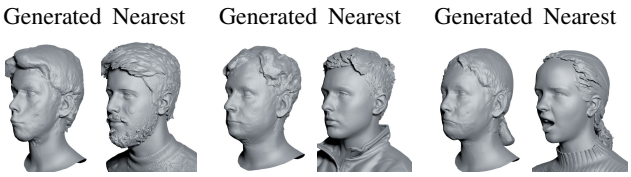

Figure 4. Randomly generated samples from HeadCraft and the corresponding nearest neighbors in the NPHM dataset among the scans used for training. L_2 distance over the scalp part of the displacement maps was used. Displacements were added to a random FLAME template for all samples.

In Table 1, we also quantify the level of detail and variety of the generated meshes w.r.t. the full head scans from the FaceVerse dataset [51] that has not been used for training. In addition, we demonstrate how much the generated samples deviate from the NPHM training set in Fig. 4. *Nearest* is found by comparing the generated displacement map to the maps of registered ground truth displacements for all training scans by L2 distance over the scalp.

The results in Table 1 indicate that the renderings from our method appear more realistic than of the other methods, with either PCA baseline or NPHM performing similar according to different subsets of the metrics. Close promixity to NPHM by MMD, COV, JSD could be explained by training on exactly the same dataset.

Ablating over the choice of the generative model architecture. We compare StyleGAN to other state-of-the-art generative model architectures, namely of VAE [29] and VQ-VAE [49] family, with ResNet-18 encoder and decoder. For VQ-VAE, the sampling from the latent space is implemented via training PixelCNN autoregressive model [48]. The results are presented in Table 2 and can be visually assessed in the Supplementary.

Behavior of the registration procedure. In Fig. 5, we demonstrate the advantage of the two-stage registration procedure, described in Subsec. 3.1, over omitting one of the stages. As can be seen, keeping only the vector displacement optimization results in too rough shape, and relaxing the regularization constraints yields significant artifacts

	FID ↓	KID ↓	MMD ↓	JSD ↓	COV ↑
Ours	**68.00**	**0.065**	**6.51**	21.33	**53.85%**
SG → VAE	112.03	0.130	6.78	21.67	47.12%
SG → VQ-VAE	124.17	0.151	7.15	21.93	43.27%
PCA	126.31	0.165	10.47	**20.16**	23.08%

Table 2. Ablation over the generative model design. VAE and VQ-VAE follow the ResNet-18 encoder and decoder architecture, while *Ours* is based on StyleGAN2. We also include PCA baseline scores here as a reference.

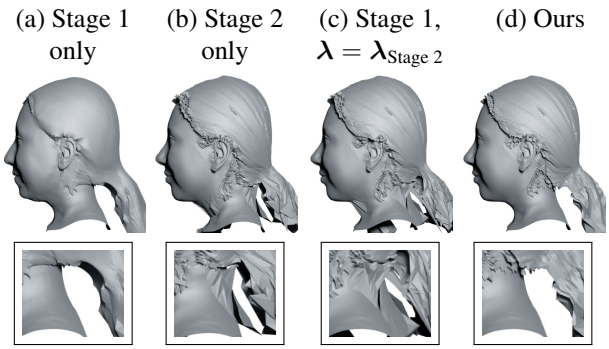

Figure 5. Ablation over the one-stage vs. two-stage registration. Regressing only vector displacements (a) yields too smooth geometry, and learning them only along the normals (b) introduces spikes – just like running the first stage with smaller $\boldsymbol{\lambda}$ (c).

such as self-intersections and spikes. Running the normal displacement stage without any preliminary vector displacement stage performs similarly to our two-stage procedure but produces artifacts for long hair that does not trivially project onto the surface. In turn, it can produce the mappings between template vertices and scan vertices, inconsistent across various samples for the long hair parts.

4.5. Applications

Fitting the latent code to a depth map. Our model can act as a prior for completing the partial observations, e.g. when they come from a depth sensor. To evaluate the performance of the model in that scenario, we demonstrate the completion capabilities of the model over a number of scans from NPHM corresponding to the subjects unseen during training. For each of these scans, we project their depth onto random viewpoints in the frontal hemisphere and project it back to 3D to construct partial point clouds. To obtain a partial UV map to be completed, we run our registration procedure with a few modifications to fit a part of the scan. Namely, we only fit the points within the convex hull of the partial point cloud, apply stronger edge length regularization weight, and constrain the points at the border of the allowed region from moving. The final mask of observed UV texels is refined by only selecting those points that turn out to be close to the partial point cloud. Finally, a latent code of HeadCraft explaining the partial UV map is found via StyleGAN inversion techniques. More technical details of the partial registration and inversion are provided in the Supplementary. The fitting quality can be evaluated by the visual comparison in Fig. 7.

Figure 6. Demonstration of the animation capabilities of the model. Each of the sequences is created by adding the randomly generated displacements from HeadCraft to the FLAME template with varying extreme expression parameters.

The capabilities of fitting the model to the full scan, e.g. created from Structure-from-Motion (SfM), are demonstrated as a part of the semantic editing experiments in Fig. 8 (top rows, the result of the latent fitting, $\lambda = \{0, 1\}$).

Animation. The decomposition of the parametric model and the displacements allows us to animate the complete head model. In our experiments, we take real multi-view video sequences with talking people from the NeRSemble dataset [30] and obtain shape, expression, jaw, and head pose parameters for each time frame of the speech by running a FLAME tracker for each sequence. For each of the sample shapes, estimated from the sequences, we reenact the corresponding FLAME with estimated expression parameters, subdivide the template and query a randomly presampled UV displacement map from HeadCraft. Since the template is also deforming over time, we rotate the displacements according to the changing surface normals of the template. The reenactment results on NeRSemble are available in the Supplementary Video. In Fig. 6, for higher clarity, we demonstrate rigging with randomly sampled FLAME shapes and a small number of extreme expressions generated artificially by randomly setting a subset of the first ten expression components of FLAME to ± 2.

Interpolation between the displacements. In Fig. 8, we show how interpolating the latent code of our generative model influences the change of the geometry. Further interpolations are presented in the Supplementary Video.

Semantic transfer from one scan to another. Access to the shared UV space allows us to modify the geometry semantically. In Fig. 8, the transfer of the scalp region from one ground truth NPHM scan, unseen during training, to the other is shown. The transfer is performed via fitting the latent representation of HeadCraft to the driver scan (the source of displacements) and feeding it to the model. The extracted displacements are later applied to the source scan.

5. Discussion

In this work, a generative model for 3D human heads is presented. We demonstrate the efficacy of the hybrid approach

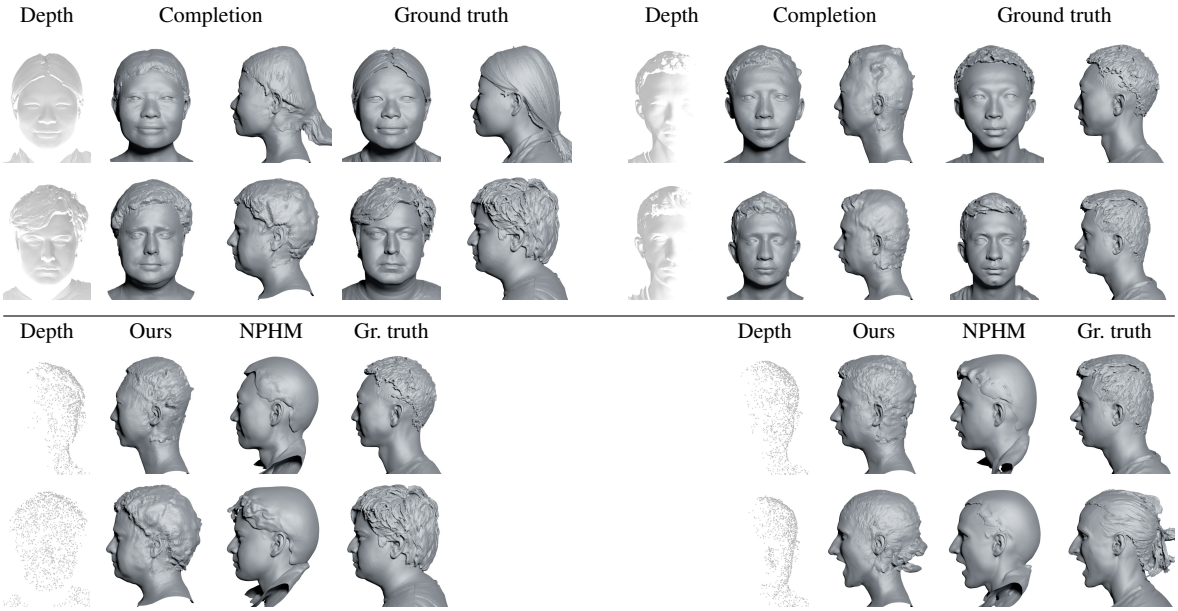

Figure 7. Demonstration of geometry completion aided by the HeadCraft model. Here, we extract depth maps from scans from the NPHM dataset, unseen during training, and try to complete them by finding the appropriate latent representation of StyleGAN. As a necessary intermediate step, we first apply our registration procedure to the partial point cloud to locate the points in the UV space of the template. The optimal latent is found by minimizing the discrepancy of the complete UV map and registered partial UV map in the observed regions. HeadCraft is also capable of estimating plausible details for a very sparse point cloud (1% of # points) – see the last row.

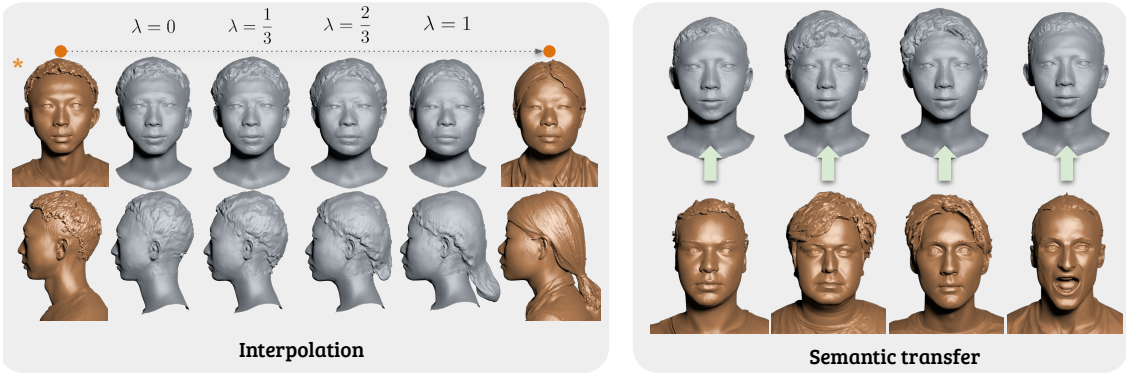

Figure 8. Semantic editing. Interpolating the latent representations of HeadCraft *(left)* allows us to smoothly change the person's appearance from one to another. To do that, we fit the latent codes for two real scans in the NPHM dataset (brown, top rows), unseen during training, and blend them together with a λ weight. Likewise, we can transfer surface variations from one person to another *(right)*. The source person is marked with $\star$ on the left and the driving person is in the bottom row on the right.

involving an underlying animatable parametric model and a neural vertex displacement modeller. Most importantly, our method allows to model high-quality shape variations while maintaining the realistic animation capability, and the inversion framework allows us to find a suitable latent representation to either represent a full head scan or a part of it that could come from e.g. the depth sensor. A direction of the possible future work could be focused on incorporating an appearance model for color and material-based relighting and a physical model of hair movement, based on, for instance, hair strands, to support more realistic animation. The code and the dataset of displacement registrations will be released to the public.

Acknowledgments. We gratefully acknowledge the support of this research by a TUM-IAS Hans Fischer Senior Fellowship, the ERC Starting Grant Scan2CAD (804724) and the Horizon Europe vera.ai project (101070093). We also thank Yawar Siddiqui, Alexey Artemov, Justus Thies for helpful advice, Tobias Kirschstein for his assistance with the NeRSemble, Taras Khakhulin for his help with the ROME baseline, Angela Dai for the video voiceover.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 2

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

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

[4] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems*, 35:19900–19916, 2022. 1

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[6] 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*, page 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 1, 2

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 5

[8] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10754–10764, 2021. 2

[9] 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. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 2

[10] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 3

[11] 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. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[12] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 6

[13] Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[14] Nira Dyn, David Levine, and John A Gregory. A butterfly subdivision scheme for surface interpolation with tension control. *ACM transactions on Graphics (TOG)*, 9(2):160–169, 1990. 3

[15] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 2

[16] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. HyperDiffusion: Generating implicit neural fields with weight-space diffusion. *arXiv preprint arXiv:2303.17015*, 2023. 6

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 2

[18] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 2

[19] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 2

[20] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023. 2, 3, 4, 5, 6

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[22] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 1, 2

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[24] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 3

[25] 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*, pages 4401–4410, 2019. 2

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adver-

sarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 3, 5

[27] 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*, pages 8110–8119, 2020. 2, 3, 5

[28] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 2, 4, 6

[29] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 6

[30] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads. *arXiv preprint arXiv:2305.03027*, 2023. 6, 7

[31] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3410–3419, 2020. 2

[32] 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):194–1, 2017. 1, 2, 3

[33] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), 2021. 1

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[35] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 2

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *CoRR*, abs/2003.08934, 2020. 1, 3

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

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1, 3

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[40] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. Genova, Italy, 2009. IEEE. 1

[41] Justin Pinkney. GitHub: Awesome Pretrained Stylegan. A collection of pre-trained StyleGAN models trained on different datasets at different resolution. https://github.com/justinpinkney/awesome-pretrained-stylegan, 2024. 5

[42] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4142–4160, 2020. 2

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

[44] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2022. 3

[45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2

[46] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2

[47] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2

[48] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 6

[49] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 6

[50] Jörg Vollmer, Robert Mencl, and Heinrich Mueller. Improved laplacian smoothing of noisy surface meshes. In *Computer graphics forum*, pages 131–138. Wiley Online Library, 1999. 5

[51] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3D face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vi-*

sion and pattern recognition, pages 20333–20342, 2022. 2, 5, 6

[52] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022. 3

[53] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[54] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3D point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019. 6

[55] 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/cvf conference on computer vision and pattern recognition*, pages 601–610, 2020. 2

[56] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3DMM: Deep implicit 3D morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021. 2, 3

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

[58] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in neural information processing systems*, 33:7559–7570, 2020. 5

[59] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. I M Avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 3

[60] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 1, 2