Pix2NPHM: Learning to Regress NPHM Reconstructions From a Single Image

Simon Giebenhain[1] Tobias Kirschstein[1] Liam Schoneveld[2]
Davide Davoli[3*] Zhe Chen[2] Matthias Nießner[1]

[1]Technical University of Munich [2]Woven by Toyota [3]Toyota Motor Europe

Fig. 1: Pix2NPHM predicts NPHM [15] latent codes given a single image. We achieve high-fidelity 3D reconstructions of diverse head shapes and expressions, shown as mesh overlays, even under strong lighting conditions and occlusions. Website: https://simongiebenhain.github.io/Pix2NPHM/

Abstract. Neural Parametric Head Models (NPHMs) are a recent advancement over mesh-based 3d morphable models (3DMMs) to facilitate high-fidelity geometric detail. However, fitting NPHMs to visual inputs is notoriously challenging due to the expressive nature of their underlying latent space, which heavily limits NPHM's practical use. To this end, we propose Pix2NPHM, a vision transformer (ViT) network that directly regresses NPHM parameters, given a single image as input, finally bridging the gap from theoretical to practical use. Compared to existing 3DMM regressors, the neural parametric space allows our method to reconstruct more recognizable facial geometry and accurate facial expressions. For broad generalization, we exploit domain-specific ViTs as backbones, which are pretrained on geometric prediction tasks. We train Pix2NPHM on a mixture of 3D data, including a total of over 100K NPHM registrations, and large-scale 2D video datasets, for which normal estimates serve as pseudo ground truth. While Pix2NPHM runs at interactive frame rates, it is possible to improve geometric fidelity by a subsequent optimization against estimated surface normals and canonical point maps. As a result, we achieve unprecedented face reconstruction quality that can run at scale on in-the-wild data.

Keywords: 3D Face Reconstruction · 3DMM parameter estimation · Feed-Forward 3DMM regressor

*Providing contracted services for Toyota.

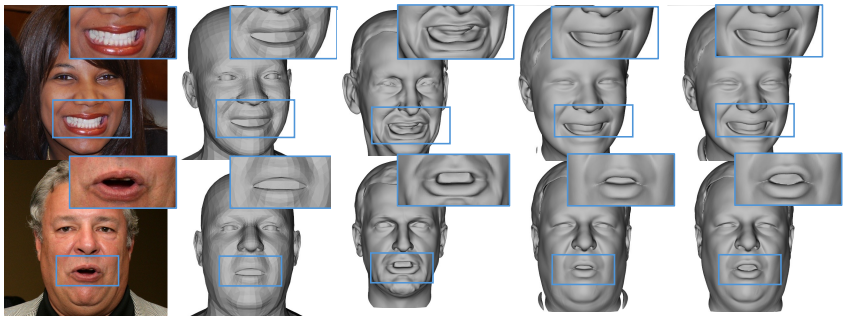


Fig. 2: Motivation: 3DMM regressors are limited by their underlying 3DMM. Higher fidelity can be obtained by replacing FLAME [32] with MonoNPHM [16], which is notoriously difficult to fit without a feed-forward prediction as initialization (see middle). Running inference-time optimization further increase fidelity (see right).

1 Introduction

Reconstructing faces in 3D, tracking facial movements, and ultimately extracting animation signals for virtual avatars are fundamental problems in many domains such as the computer games and movie industry, telecommunication, and AR/VR. Arguably the most relevant sub-task is 3D face reconstruction from a single image due to the vast availability of image collections as well as straightforward extensions to sequential tracking.

In order to solve the underconstrained reconstruction problem, 3d morphable models (3DMMs) [3] have evolved as industry and research standard due to their concise low-dimensional parametric representation, which lead to a plethora of algorithms build on top of 3DMMs. With the advancement of deep-learning methods, photometric tracking [56] approaches, have been augmented with additional priors, such as facial landmark detection, or direct 3DMM parameter regression from RGB signal [10,12,49,50,70,77]. Recently, additional priors, such as dense landmarks [41,52,65] and surface normals [2,18] have further improved reconstructions. Due to such methods that enable fitting in even the most challenging scenarios, 3DMMs have become an essential component of photo-realistic avatars [17, 43], generalized avatars [5, 25, 30, 31, 39], and even controllable generative diffusion models for faces [14, 29, 42, 53, 54, 66].

While 3DMMs have achieved great success in these domains, we argue that their concise parametric representation comes at the cost of geometric expressiveness, i.e. modern 3DMMs, such as FLAME [32], are unable to model high-fidelity geometric detail. Therefore, a more recent line of work has developed neural parametric head models (NPHMs) [15,16,68,71] for increased representational capacity, as shown in Fig. 2. This increased model capacity, however, makes image-based reconstruction challenging due its expressive parameter space. MonoNPHM [16] has attempted to reconstruct NPHM parameters from a single image. However, their purely photometric fitting approach remained slow and brittle

in real-world applications (see Fig. 2). To this end, we propose a robust and high-fidelity fitting frame-work, yielding a first-class tool for face reconstruction and tracking based on NPHM [15, 16].
Our approach addresses the two main challenges of neural parametric model fitting: underconstrained optimization and reconstruction speed. This is achieved by tailoring a transformer-based feed-forward predictor for NPHM parameters from a single image. As a highly data-driven approach, we shift the complexity from inference-time to training-time, ultimately enabling easy practical use. Thus we require large-scale high-quality training data. To this end, we curated a large collection of publicly available 3D face datasets and fitted MonoNPHM against it, resulting in a total of 102K registrations, which will soon be shared with the research community. Despite these efforts, we find that training on large-scale 2D video datasets using a self-supervised geometric loss based on estimated surface normals [18] further improves generalization. Furthermore, we observe strong generalization improvements by replacing a generic DINOv2 [36] backbone with a ViT [11] pre-trained on per-pixel geometric prediction tasks, such as surface normal or canonical point map regression. Our feed-forward estimator produces state-of-the-art (SotA) results, and makes Pix2NPHM a first-class choice for 3D face reconstruction, due to its fidelity, ease of use, and robustness to diverse input scenarios, as showcased in Fig. 1. Moreover, we show that our results can be refined by a few optimization steps at inference time against estimated surface normals to increase fidelity, as shown in Fig. 2. Together, these insights improve 3D reconstruction on the NeRSemble single-image face reconstruction (SVFR) [18] benchmark by 21%, measure as L1 chamfer distance, and neutral reconstruction improves by 6% compared to the best public method on the NoW [49] benchmark. To summarize, our main contributes are as follows:

- Pix2NPHM is the first feed-forward regressor for NPHM parameters, enabling, for the first time, robust and fast NPHM fittings from a single image.
- We curate a large mixture of high-quality 3D datasets, with over 100K NPHM registrations.
- We formulate a novel 2D self-supervised loss via estimated normals.
- We show that geometrically pre-trained, face-specific ViTs outperform DINOv2 as image encoding backbone.

2 Related Work

Optimization-Based Methods: 3D Morphable Models (3DMMs) [3, 24, 32] represent shape and appearance via linear PCA spaces. Fitting these models to images or videos is typically approached via optimization [3, 21, 43, 56, 77], usually by minimizing photometric error between renderings and the input. Because photometric losses are under-constrained and sensitive to illumination and noise, many works incorporate geometric priors such as sparse landmarks or semantic segmentation [19, 43, 56, 77] for stability. More recent methods leverage dense geometric cues to further improve robustness and accuracy, such as dense landmarks [65], UV flow [52], or a combination of UV maps and surface normals [18].

Regression-Based Methods: Regression-based approaches train deep networks to directly predict 3DMM parameters, avoiding optimization and enabling real-time inference. Fully supervised approaches [22, 45, 46, 57, 58, 77] rely on registered 3D scans or synthetic data, but are limited by data scarcity. Self-supervised methods learn from in-the-wild images using photometric constraints [10, 12, 13, 44, 49, 50, 55, 70]. EMOCA [10] extends this by incorporating emotion-aware supervision for more expressive results, and SPECTRE [13] integrated lipreading-based cues to better capture mouth movements and speech-related expressions. TokenFace [70] combines 3D and 2D supervision in a ViT-based framework. Recent work highlights the role of improved differentiable rendering: SMIRK [44] employs a neural renderer, while SHeaP [50] uses 2D Gaussian splats [23]. Compared to existing self-supervised FLAME regressors, we replace photometric cues with surface-normal supervision. While FLAME-based regressors have steadily improved throughout the last decade, our work builds training data, and 3D and 2D supervision signals completely from scratch. The learned regressor finally unlocks NPHM for in-the-wild and large scale use cases.

Neural 3DMMs: Classical 3DMMs [3, 24, 32] rely on fixed topology and linear PCA spaces, limiting detail and expressiveness. Neural 3DMMs address these issues by using implicit neural representations such as SDFs [15, 37, 40, 68, 71, 72] or occupancy grids [7], and sometimes maintain an explicit representation such as meshes [51, 62, 75] or 3D Gaussians [67, 76]. NPHM [15] models head shape via local SDF experts anchored at semantic keypoints, and uses a deformation field for expressions. MonoNPHM [16] proposed monocular NPHM fitting via volumetric SDF rendering and canonical appearance modeling. SIRA++ [6] has comparable capabilities, but lacks disentangled expression control. Other works explore photorealistic neural head models using NeRFs [34] and 3D Gaussian splats [27], such as [4, 27, 34, 60, 67, 73]. Compared to FLAME, NPHM's manifold has proven valuable for downstream tasks, including realistic avatars [17, 29] and high-fidelity audio-driven geometry [1]. Motivated by this, we develop a fast, robust, and accurate reconstruction approach that makes MonoNPHM practical for broader downstream use.

3 Pix2NPHM

Given a single input image I, it is our goal to estimate identity $\mathbf{z}_{\text{id}}$ and expression parameters $\mathbf{z}_{\mathbf{ex}}$ from which 3D geometry can be recovered using MonoNPHM [16] as decoder. After a brief background section on NPHMs (Sec. 3.1), we describe our face-specific pre-training strategy to obtain our image encoding backbone in Sec. 3.2. Next, in Sec. 3.3, we introduce Pix2NPHM, the first feed-forward NPHM parameter regressor. Fig. 3 provides an overview of our network architecture, and training strategy, which is further described in Sec. 3.4. Finally, Sec. 3.5 describes our inference-time optimization, which further improves reconstruction fidelity.

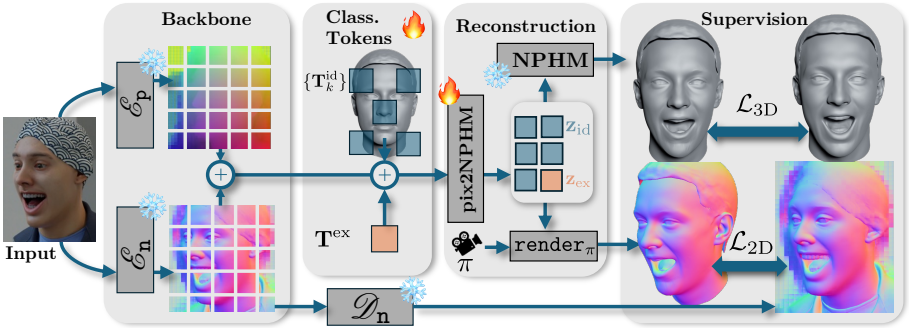


Fig. 3: Method Overview: We use pretrained ViTs, $\mathscr{E}_{\mathtt{n}}$ and $\mathscr{E}_{\mathtt{p}}$, as backbone, which encode the input into a token sequence. The resulting sequence is concatenated with learnable classifier token $\{\mathbf{T}^{\text{id}}_k\}$ and $\mathbf{T}^{\text{ex}}$, which are decoded into NPHM identity ($\mathbf{z}_{\text{id}}$) and expression ($\mathbf{z}_{\text{ex}}$) parameters using transformer network. We train using a 3D SDF loss, and a normal rendering loss against pseudo g.t. normals.

3.1 Background: Neural 3DMMs

In this paper we adopt the neural 3DMM formulation introduced by MonoNPHM [16]. It defines a neural field

$$\mathscr{F}_{\text{NPHM}} : (x, \mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}}) \mapsto \text{SDF}(x), \tag{1}$$

which predicts SDF-values for the 3D head geometry, given points $x \in \mathbb{R}^3$, and is conditioned on disentangled identity and expression latent codes. Internally, $\mathscr{F}_{\text{NPHM}}$ consists of a backward deformation field, conditioned on $\mathbf{z}_{\text{ex}}$ and $\mathbf{z}_{\text{id}}$ and a canonical SDF conditioned on $\mathbf{z}_{\text{id}}$. A mesh can be extracting using marching cubes [33], once latent codes have been obtained. The remainder of the paper will focus on the inference of ideal latent codes from a single image. Throughout all our experiments we use the public checkpoint from MonoNPHM [16].

3.2 Robust Encoder via Geometric Pre-Training

We tackle the inherent ambiguity of the single-image reconstruction task using a heavily data-driven approach. Our method starts with pre-training a ViT-based backbone on a per-pixel geometric reconstruction task. We follow the training strategy and encoder-decoder architecture

$$\mathscr{E}_{\mathtt{geo}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{L\times D} \tag{2}$$

$$\mathscr{D}_{\mathtt{geo}} : \mathbb{R}^{L\times D} \to \mathbb{R}^{H\times W\times 3} \tag{3}$$

from Pixel3DMM [18], where $\mathtt{geo} \in \{\mathtt{n}, \mathtt{p}\}$ represent estimators for surface normal and canonical point cloud. While normals have been proposed in Pixel3DMM [18], canonical point maps are a novel pixel-aligned objective. While regular point cloud prediction has to be handled using relative coordinate systems for arbitrary scenes, such as in DUSt3R [64] and VGGT [61], we exploit the possibility to define a *unique* coordinate system for 3D heads. Once pre-training is completed,

we discard $\mathscr{D}_{\texttt{geo}}$, and use the face-specific geometrically aware encoders $\mathscr{E}_{\texttt{geo}}$ as backbone for our regression network. Later we will show that this face-specific pre-training significantly outperforms DINOv2 image encodings.

3.3 Feed-Forward NPHM Parameter Prediction

Following the recent success of transformers [59], we employ learnable classifier tokens, similar to ViT [11] and TokenFace [70], which read out relevant identity and expression information via several transformer layers. We concatenate our encoded token sequence $[\mathscr{E}_{\texttt{n}}(I), \mathscr{E}_{\texttt{p}}(I)]$ with learnable classifier tokens

$$\mathbf{T}_{\text{CLS}} = \left[\mathbf{T}^{\text{ex}}, \{\mathbf{T}^{\text{id}}_k\}_{k=1}^{66}\right] \tag{4}$$

for expression and identity. Note that $\mathbf{T}^{\text{id}}$ is composed of 65 tokens, one for each local identity codes from MonoNPHM [16], and one for the global identity part. Finally, our regressor

$$\text{pix2NPHM} : ([\mathscr{E}_{\texttt{n}}(I), \mathscr{E}_{\texttt{p}}(I)], \mathbf{T}_{CLS}) \mapsto (\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}}) \tag{5}$$

consists of several transformer layers, and MLP prediction heads which map the resulting classification tokes tokens $\mathbf{T}'_{\text{CLS}}$ from the transformer to $\mathbf{z}_{\text{id}}$ and $\mathbf{z}_{\text{ex}}$, respectively. Finally, the predicted latent codes can be decoded into a 3D shape using the MonoNPHM model, which represents the 3D head geometry implicitly by conditioning an SDF, see Eq. (1)

3.4 Training Strategy

3D supervision in SDF Space On 3D datasets we aim to formulate the supervision signal as unambiguous as possible, and exploit the fact that ground truth NPHM codes $(\mathbf{z}^{\text{gt}}_{\text{id}}, \mathbf{z}^{\text{gt}}_{\text{ex}})$ can be obtained as a pre-processing step using our registration procedure, as described in Sec. 3.6. Given an image I, we predict NPHM latents $\text{pix2NPHM}(I, \mathbf{T}_{CLS}) = (\hat{\mathbf{z}}_{\text{id}}, \hat{\mathbf{z}}_{\text{ex}})$, and compute the loss

$$\mathcal{L}_{\text{3D}} = \sum_{x \in \mathcal{X}} \|\text{NPHM}(x; \hat{\mathbf{z}}_{\text{id}}, \hat{\mathbf{z}}_{\text{ex}}) - \text{NPHM}(x; \mathbf{z}^{\text{gt}}_{\text{id}}, \mathbf{z}^{\text{gt}}_{\text{ex}})\|_1, \tag{6}$$

between the induced SDFs, where $\mathcal{X}$ is a point cloud randomly sampled near the 3D surface. Directly supervising $\|\hat{\mathbf{z}} - \mathbf{z}\|$ did not lead to training convergence.

2D Self-Supervision: Adding Diversity Since 3D datasets barely cover all relevant modes of the valid input distribution, we add data diversity by training on large scale 2D video datasets using estimated surface normals $I^{\texttt{n}} = \mathscr{D}_{\texttt{n}}(\mathscr{E}_{\texttt{n}}(I))$ as pseudo ground truth. To this end, we modify the NeuS-based [63] rendering approach from MonoNPHM [16] to render surface normals using gradients of the SDF, and supervise via the cosine similarity

$$\mathcal{L}^{\texttt{n}}_{\text{2D}} = \sum_{\mathbf{p} \in \mathcal{P}} \langle \texttt{render}_{\pi}\left(\nabla \mathscr{F}_{\text{NPHM}}; \hat{\mathbf{z}}_{\text{id}}, \hat{\mathbf{z}}_{\text{ex}}\right)_{\mathbf{p}}, I^{\texttt{n}}_{\mathbf{p}} \rangle \tag{7}$$

between rendered and estimated normals, where camera parameters π are provided by our data registration. Note that due to the memory-intensive nature of MLP-based volumetric rendering (we find that 32 samples per ray are sufficient), we only supervise a random subset of 50 pixels $\mathbf{p} \in \mathcal{P}$, uniformly sampled in the facial area. Overall, we find this supervision to provide much more meaningful and stable gradients, compared to computing a photometric loss using spherical harmonics, which is the most established loss function in existing FLAME-based regressors [12, 50].

Full Training Objective Overall, we train our feed-forward predictor using

$$\mathcal{L}_{\text{total}} = \lambda_{\text{3D}}\mathcal{L}_{\text{3D}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{reg}}\mathcal{R} \tag{8}$$

as our complete training objective, which combines 3D and 2D losses with regularization terms, where λ_{3D} is set to 0 for 2D video datasets. Our regularization

$$\mathcal{R}(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}}) = \lambda^{\mathcal{R}}_{\text{id}}\|\mathbf{z}_{\text{id}}\|_2 + \lambda^{\mathcal{R}}_{\text{ex}}\|\mathbf{z}_{\text{ex}}\|_2 \tag{9}$$

simply punishes the norm of the predicted latents.

3.5 Test-Time Optimization: Increasing Fidelity

For even more accurate 3D reconstructions our feed-forward estimates can be naturally combined with inference-time optimization against per-pixel geometric predictions, similar to Pixel3DMM [18]. We optimize for

$$\underset{\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}}, \pi}{\arg\min}\, \lambda_{\text{n}}\mathcal{L}^{n}_{\text{2D}} + \lambda_{\text{p}}\mathcal{L}^{p}_{\text{2D}} + \lambda_{\text{reg}}\mathcal{R}\left(\mathbf{z}_{\text{id}} - \hat{\mathbf{z}}_{\text{ex}}, \mathbf{z}_{\text{ex}} - \hat{\mathbf{z}}_{\text{id}}\right), \tag{10}$$

where our feed-forward predictions serve as initialization and regularization target, and $\mathcal{L}^{\text{p}}_{\text{2D}}$ is defined similar to Eq. (7) but with an L_1-Loss. Note that we first need to estimate camera parameters π, using dense landmark predictions from Pixel3DMM, which are fine-tuned later alongside $\mathbf{z}_{\text{id}}$ and $\mathbf{z}_{\text{ex}}$.

3.6 High-Quality Training Dataset Curation

As a data-driven approach, a core requirement for success is sufficient high-quality training data. Therefore, we spend a considerable amount of effort on curating 3D and 2D datasets. We briefly describe our approach for different data types below, and provide an overview of datasets in table Tab. 1.

3D datasets: For 3D datasets, we estimate the canonical coordinate frame using FLAME registration similar to [15]. After transforming the ground truth 3D shapes into a unified coordinate system, we sample points $x \in \mathcal{X}$ on the ground truth surface and optimize for NPHM parameters

$$\underset{\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}}}{\arg\min} \sum_{x \in \mathcal{X}} \|\text{NPHM}(x; \mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}})\|_1 + \lambda_{\text{reg}}\mathcal{R}(\mathbf{z}_{\text{id}}, \mathbf{z}_{\text{ex}}) \tag{11}$$

Table 1: Training Data: We list number of identities, facial expressions, camera view-points, total images and total 3D shapes.

	#IDs	#Expr./#Views	#Images	#3D shapes
NPHM [15]	450	23 / 40	414K	10K
FaceScape [75]	300	20 / 50	300K	6K
DAViD [48]	65K	1 / 1	65K	65K
MimicMe [38]	2000	10 / 5	100K	20K
LYHM [9]	1200	1 / 2	2.4K	1.2K
Videos [8, 69, 74]	<50K	5 / 1	250K	0
Total 3D	69K	-/-	880K	102K
Total	119K	-/-	1.13M	102K

2D datasets: For large-scale, in-the-wild 2D datasets, obtaining reliable and 3D accurate NPHM fittings was a previously unsolved problem. Therefore, we simply estimate camera poses using the video tracker from Pixel3DMM [18], and purely rely on our rendering loss $\mathcal{L}_{\text{2D}}^{\text{n}}$.
Data Cleaning: To ensure high-quality training data, we devise a simple outlier filtering strategy, for both, 3D NPHM fittings and 2D FLAME video fittings. We define outlier thresholds, by analyzing the histograms of the norms of certain attributes, such as shape, expressions, neck and jaw parameters.

4 Experimental Results

To verify our proposed method experimentally, we compare against recent SotA methods on the NeRSemble single-view face reconstruction (SVFR) benchmark [18] (see Sec. 4.3), and on the established Now benchmark [49] (see Sec. 4.4). In Sec. 4.6 we conduct thorough ablations experiments to quantify the significance of our individual technical contributions. We highly encourage the reviewers to watch our supplementary video for more results. We will release our training data, model and full code base for research purposes.

4.1 Implementation Details

We implement our pipeline using PyTorch, and train our regression model using the Adam [28] optimizer, a batch size of 32 and learning rate of $1e^{-4}$ on a single A100-80GB GPU, which takes 4 days until convergence. We set λ_{3D}=10.0, λ_{2D}=1.0 and λ_{reg}=$1e^{-4}$. We use 8 transformer layers, each consisting of pre-norm LayerNorm, self-attention with 8 heads, a 2-layer MLP with GeLU activation, and a width of 1024.
We pre-train $\mathscr{E}_n$ and $\mathscr{E}_p$ separately for 3 days on 2 A6000 GPUs, and follow the network architecture and training strategy fom Pixel3DMM [18].
To register our training dataset, consisting of 102K 3D shapes, with the NPHM model, we invested roughly 2500 GPU-hours. We will release our training data to the community to drive future research. Note, that we explicitly exclude the NeRSemble-SVFR benchmark [18] identities from our NPHM training data, since both datasets share the same identities.

Table 2: Quantitaitive Results on the sinlg-image *posed* and *neutral* NeRSemble 3D face reconstruction benchmark [18] and NoW [49]. Methods above the line represent feed-forward networks, methods below require optimization. We also indicated public availability, ✓* indicates soon public release.

Method	NeRSemble Neutral			NeRSemble Posed			NoW Error			Avail.
	L1↓	L2↓	NC↑	L1↓	L2↓	NC↑	Median	Mean	Std	
MICA [77]	1.68	1.14	0.883	-	-	-	0.90	1.11	0.92	✓
TokenFace [70]	-	-	-	2.62	1.78	0.865	**0.76**	**0.95**	**0.82**	✗
DECA [12]	2.07	1.40	0.876	2.38	1.61	0.870	1.09	1.38	1.18	✓
EMOCAv2 [10]	2.21	1.49	0.873	2.63	1.78	0.860	-	-	-	✓
SHeaP [50]	1.86	1.26	0.882	2.08	1.41	0.876	0.95	1.18	0.99	✓
Ours (ffwd. only)	1.57	1.06	0.896	1.55	1.05	0.894	0.83	1.03	0.88	✓*
Metr. Tracker [77]	-	-	-	2.03	1.37	0.878	0.90	1.11	0.92	✓
FlowFace [52]	1.93	1.31	0.878	1.96	1.33	0.879	0.87	1.07	0.88	✗
Pixel3DMM [18]	1.66	1.12	0.883	1.66	1.11	0.884	0.87	1.07	0.89	✓
MonoNPHM [16]	2.32	1.56	0.878	2.50	1.68	0.870	-	-	-	✓
Ours	**1.54**	**1.04**	**0.897**	**1.37**	**0.92**	**0.897**	0.81	1.01	0.85	✓*

Runtime Analysis: Our feed-forward transformer runs at 8fps on an RTX3080 GPU, and the optional inference-time optimization, takes 85s. Related optimization-based methods MonoNPHM and Pixel3DMM take 150s and 30s, respectively. MICA, a CNN regressor, operates at 14fps. All numbers are measured on an RTX3080 GPU.For more details on the inference-time optimization procedure, we refer to our supplementary material.

4.2 Baselines

Feed-Forward FLAME Regressors: The first class of baselines are existing regressors for FLAME [32] parameters. We select DECA [12], MICA [77], EMOCAv2 [10], TokenFace [70] and SHeaP [50].

FLAME Optimization: The next class combines feed-forward predictions as a prior with inference-time optimization. As such, this category is closely related to our full approach, but uses a classical 3DMM representation. We select MetricalTracker [77], FlowFace [52] and Pixel3DMM [18].

NPHM Optimization: Finally, MonoNPHM [16] proposes a photometric fitting pipeline to obtain NPHM parameters. Since we also use the MonoNPHM model to decode latent parameters into 3D shapes, this comparison highlights the added robustness of our approach compared to photometric fitting.

Baseline Availability: Finally, we point out the poor public availability of recent SotA methods, *i.e.* TokenFace [70] and FlowFace [52], two of the top performing methods on NoW, remain unreleased, and haven't been replicated as of yet. Thus the numbers reported in this paper were obtain with the help of the respective paper's first authors.

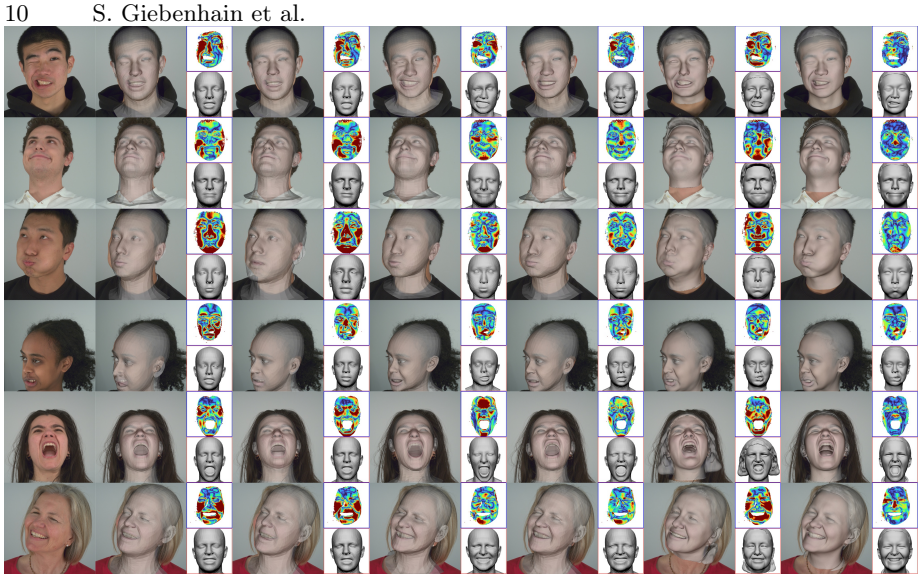


Fig. 4: Posed Reconstruction: We show overlays of the reconstructed meshes. Insets with a blue border depict L_2-Chamfer distance as an error map, rendered from a frontal camera. Red insets show the reconstructed mesh from the same camera. All our figures are best viewed digitally and zoomed-in.

4.3 NeRSemble SVFR Benchmark

The recent NeRSemble SVFR benchmark [18] features diverse and challenging facial expressions, and is the first to simultaneously allow the evaluation of *posed* and *neutral* face reconstruction. Given a single image as input, the *posed* task measures the 3D reconstruction accuracy, while the *neutral* task measures how well the methods can disentangle expressions, by predicting a neutral face, from an image with an arbitrary facial expression. We provide quantitative and qualitative results in Tab. 2 and Figs. 4 and 6, respectively. We significantly outperform all baselines across all metrics, even without test-time optimization.

4.4 NoW Benchmark

The NoW benchmark [49] can only measure the *neutral* reconstruction task, but, compared to the NeRSemble SVFR benchmark, it covers a wider variety of identities, lighting conditions, hair styles, head accessories and other types of occlusion. Our method significantly outperforms all baselines, except for TokenFace [70], which is, however, not publicly available and performs poorly on the NeRSemble SVFR benchmark, as presented in Tab. 2. Qualitative results from the validation set are presented in Fig. 5.

4.5 AffectNet Benchmark

Although emotion recognition is not directly measuring 3D understanding, we further assess the degree to which the semantics of facial expressions are accurately captured by our feedforward network on AffectNet [35]. We extract shape

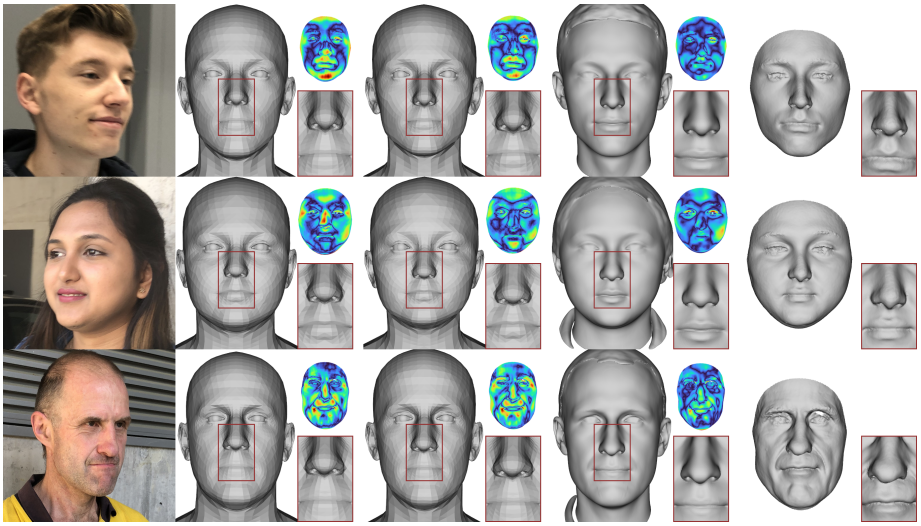


Fig. 5: Neutral Reconstruction, NoW: We show frontal mesh renderings in comparison to the ground truth mesh, as well as, error maps and zoom ins.

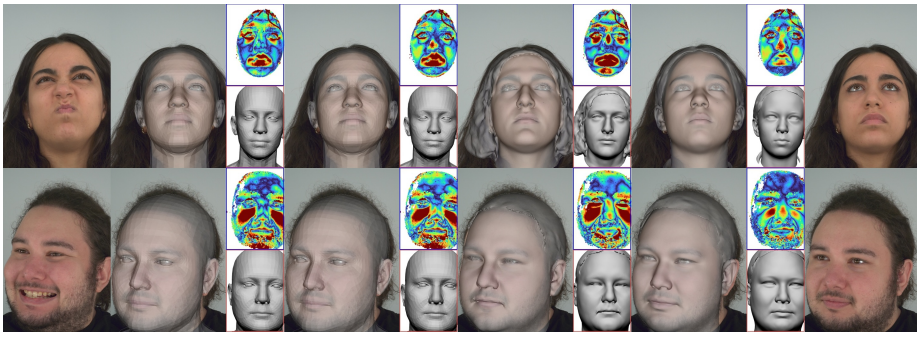


Fig. 6: Neutral Reconstruction, NeRSemble: Comparison against available SotA methods on top of neutral reference image.

and expression parameters using Pix2NPHM for all train and test images in the AffectNet dataset (note that this is done in a feedforward-only manner, without per-image optimization). Following EMOCA [10], we then fit a 4-layer MLP to the the parameters extracted from the train set, and evaluate on the test set. Pix2NPHM outperforms all competing methods on this benchmark (Table 3), showcasing our trained feedforward network's ability to extract detailed expression information.

Table 3: Emotion recognition: We report the concordance correlation coefficient (CCC) and root mean squared error (RMSE) on predicting Valence and Arousal, and, most importantly, we report accuracy in 8-way emotion classification on AffectNet [35].

Model	Arousal		Valence		Emo
	CCC↑	RMSE↓	CCC↑	RMSE↓	Acc.
SMIRK [44]	0.560	0.288	0.681	0.313	0.653
EMOCA [10]	0.577	0.282	0.70	0.307	0.676
SHeaP [50]	0.615	0.274	0.735	0.301	0.695
Ours-NPHM	**0.621**	**0.273**	**0.739**	**0.291**	**0.711**

Table 4: Ablations on NeRSemble-SVFR Benchmark [18] and NoW [49].

Method	Description					NeRSemble Neutral			NeRSemble Posed			Now
	2D	3D Input		Opt.	3DMM	L1↓	L2↓	NC↑	L1↓	L2↓	NC↑	
3D only		✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$		NPHM	1.64	1.11	0.894	1.65	1.11	0.890	1.074
3D only, no regi.		✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$		NPHM	1.93	1.30	0.892	2.05	1.38	0.887	-
2D only	✓		$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$		NPHM	1.79	1.21	0.893	1.67	1.13	0.893	1.238
RGB input		✓	DINO(I)		NPHM	1.89	1.28	0.891	2.10	1.42	0.883	1.165
Normals inp.	✓	✓	DINO(I^n)		NPHM	1.76	1.19	0.893	1.87	1.26	0.886	1.168
no $\mathscr{E}_\mathrm{p}$	✓	✓	$\mathscr{E}_n$		NPHM	1.67	1.13	0.895	1.65	1.11	0.891	1.053
ffwd. NPHM	✓	✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$		NPHM	1.57	1.06	0.896	1.55	1.05	0.894	1.016
ffwd. FLAME	✓	✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$		FLAME	1.71	1.15	0.883	1.81	1.22	0.881	1.073
opt. only w/o MICA	✓	✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$	✓	NPHM	1.76	1.18	0.894	1.61	1.09	0.893	1.137
opt. only w/ MICA	✓	✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$	✓	NPHM	1.60	1.08	0.890	1.50	1.01	0.895	1.029
Ours	✓	✓	$\mathscr{E}_n + \mathscr{E}_\mathrm{p}$	✓	NPHM	**1.54**	**1.04**	**0.897**	**1.37**	**0.92**	**0.897**	**0.988**

4.6 Ablation Studies

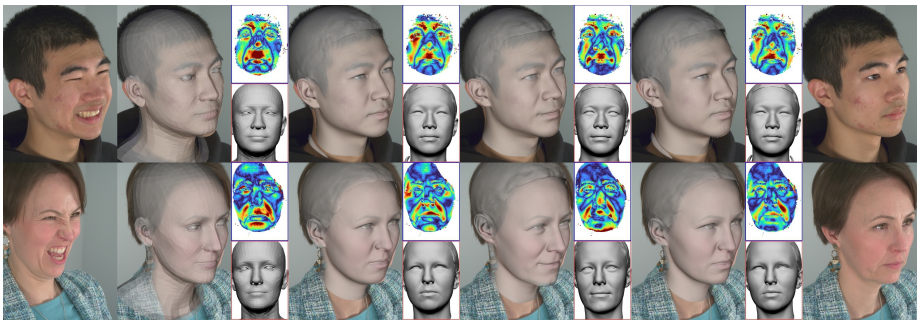
We demonstrate the effectiveness of our technical contributions by removing them one-by-one from our complete method. Quantitative and qualitative results are presented in Figs. 7 and 8, and Tab. 4, respectively. Additionally, Tab. 4 has a verbose description of the ablated components.

What impact does the data preprocessing have? The registration of the five different 3D datasets into uniform NPHM format results in significant performance improvements, as shown in Tab. 4. Without such registration, training becomes more complex; e.g., DAViD only provides single-view depth maps, FaceScape only non-watertight meshes with noisy back of the head and MimicMe provides incomplete meshes (only the frontal face part). To showcase the importance we implemented point-cloud based supervision using the IGR [20] loss. Tab. 4 shows significantly worse quantitative results. We parallellized the registration effort on many low-grade GPUs (GTX1080/RTX2080) and will release the registration to the public.

What effect does the underlying 3DMM have? The main motivation for Pix2NPHM is the increased representational capacity of neural 3DMMs, *i.e.* MonoNPHM in our case. To ablate the effect of the underlying 3DMM, we train a version of our approach with FLAME instead of MonoNPHM. Quantitative and qualitative results on the SVFR-Benchmark and NoW support our claim,

Input ffwd. FLAME ffwd. NPHM Opt. only w/ MICA Ours

Fig. 7: Ablations, Posed: NPHM feed-forward predictions exhibt more details compared to FLAME. Wihtout the feed-forward initialization our optimization sometimes fails to reconstruct extreme expressions (*e.g.* see rows 1 and 3).



Input ffwd. FLAME ffwd. NPHM Opt. only w/ MICA Ours Reference

Fig. 8: Ablations, Neutral: Without the feed-forward prediction as initialization, our optimization cannot properly disentangle identity and expression.

that leveraging a neural 3DMM increases reconstruction fidelity. Note that our FLAME-based feed-forward predictor outperforms the best available FLAME-based predictor MICA on the NoW validation set (1.073 vs. 1.109 mean error), and outperforms SHeaP, the best FLAME-based feed-fordward method on the posed NeRSemble SVFR task.

Inference-Time Optimization: While our feed-forward prediction already reaches SotA performance, reconstruction fidelity can be improved further, *e.g.* see Fig. 2. Especially posed reconstructions benefit from optimization, while

neutral reconstructions improve slightly (see Tab. 2). Importantly, we note that optimization without using our feed-forward as initialization sometimes fails to reconstruct complicated expressions, *e.g.* see rows one and three of Fig. 7. For this we distinguished two separate scenarios: "w/o MICA" uses no prior, while "w/ MICA" uses the regressed FLAME mesh as additional optimization constraint. See our supplementary for more details.

Image Encoding: Furthermore, we ablate several options on how the input image is encoded. Using DINOV2 [36] encodings of the RGB input I, similar to Pixel3DMM [18], performs the worst, and did not properly converged on 2D training data. Using DINOv2 encodings of estimated normals $I^{\mathtt{n}}{=}\mathscr{D}_{\mathtt{n}}\left(\mathscr{E}_{\mathtt{n}}(I)\right)$ resulted in similar performance on NoW, but increased performance on NeRSemble. Directly using the tokens from $\mathscr{E}_{\mathtt{n}}$ gives significantly better performance and adding tokens from $\mathscr{E}_{\mathtt{p}}$ further boosts results.

Training Data: Finally, we show the importance of training on a mixture of 2D and 3D data. We hypothesis that 2D data is essential to handle the input appearance diversity, while 3D data provides an unambiguous training signal.

Inference Time Optimization Analysis: Fig. 9 presents the trade-off between the number of inference-time optimization steps and the reconstruction fidelity. For this analysis, we report the metrics of the posed NeRSemble SVFR benchmark, indicating that 20 and 60 optimization steps provide good latency-quality trade-offs. Nevertheless, optimization fundamentally remains slow. For real-time inference, we believe that distilling Pix2NPHM into a ffwd-only model with slimmer architecture (w/o two ViT Backbones) is the most promising solution; one could use Pix2NPHM (w/ optimization) as a data generator.

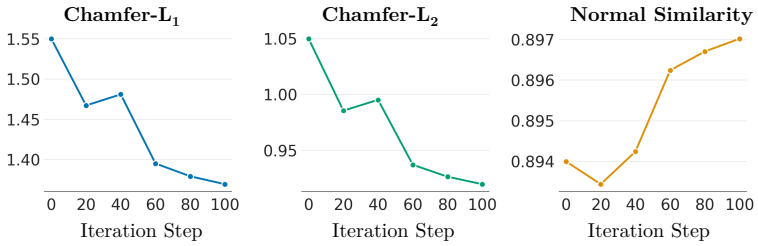


Fig. 9: Analysis of number of test-time optimization steps.

4.7 In-the-Wild Results

For results on in-the-wild images we refer to Fig. 1 and our supplementary material. Additionally, Fig. 2 compares our FLAME-based and NPHM-based regressors, and our full approach.

4.8 Downstream Application: Avatar Cross-Reenactment

As an example of a practical applications, we demonstrate that our feed-forward predictions of $\mathbf{z}_{\text{ex}}$ can be used to cross-reenact NPGA [17] avatars, as depicted

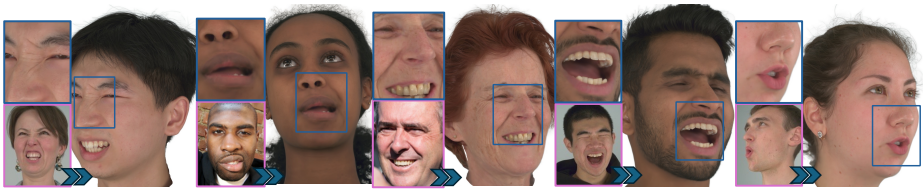


Fig. 10: Cross-Reenatment: Driving NPGA [17] avatars using our feed-forward predictions on the driving image (pink boundary). Zoom-ins are shown in blue boundaries.

in Fig. 10. We use driving images from the NeRSemble benchmark [18], as well as, in-the-wild images from FFHQ [26]. Note, that there is no identity leakage, and that the avatars maintain their own authentic expression characteristics.

5 Limitations and Future Work

We believe that our scalable approach to 3D face reconstructions shows great potential, however, performance is currently limited by several aspects of MonoNPHM [16]. For example, its latent space prevents reliable registration of 3D hair styles, which leads to suboptimal geometry, even in the facial region. Moreover, the MLP-based volumetric rendering and marching cubes extraction introduce substantial computational overhead. Future work could develop an improved NPHM variant, *e.g.* leveraging 2DGS [23] for faster rendering and mesh extraction, and training on more diverse 3D hairstyles such as Difflocks [47]. Jointly fine-tuning NPHM with the feed-forward regressor is another promising direction. Finally, as single-image reconstruction remains inherently ambiguous, incorporating uncertainty estimation, probabilistic reconstruction, *e.g.* via conditional generation, or multi-image extensions, *e.g.* by leveraging sequential information as in VGGT [61], could resolve any remaining ambiguities.

6 Conclusion

We introduced Pix2NPHM, the first feed-forward framework for predicting NPHM parameters, enabling fast and high-fidelity 3D face reconstruction from a single input image. Despite considerable research effort on FLAME regression, our initial attempt on NPHM reconstruction demonstrates significant benefits. Key to this success are our large-scale 3D data registration and the self-supervised training strategy on 2D data using surface normal estimators. Moreover, we show that ViTs pretrained on face-specific geometric tasks capture facial structure far more effectively than DINOv2. Together, these insights establish a new path toward scalable, high-quality monocular 3D face reconstruction.

Acknowledgements

This work was supported by Toyota Motor Europe and Woven by Toyota. This work was also supported by the ERC Consolidator Grant Gen3D (101171131). We would also like to thank Angela Dai for the video voice-over.

References

1. Aneja, S., Thies, J., Dai, A., Nießner, M.: Facetalk: Audio-driven motion diffusion for neural parametric head models. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024) 4
2. Artru, N., Hussain, R., Got, E., Messier, A., Lindell, D.B., Dib, A.: Skullptor: High fidelity 3d head reconstruction in seconds with multi-view normal prediction (2026), https://api.semanticscholar.org/CorpusID:286000870 2
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999) 2, 3, 4
4. Buehler, M.C., Li, G., Wood, E., Helminger, L., Chen, X., Shah, T., Wang, D., Garbin, S., Orts-Escolano, S., Hilliges, O., Lagun, D., Riviere, J., Gotardo, P., Beeler, T., Meka, A., Sarkar, K.: Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In: ACM SIGGRAPH Asia 2024 Conference Paper (2024). https://doi.org/10.1145/3680528.3687580, https://doi.org/10.1145/3680528 4
5. Cao, C., Simon, T., Kim, J.K., Schwartz, G., Zollhoefer, M., Saito, S., Lombardi, S., Wei, S.E., Belko, D., Yu, S.I., Sheikh, Y., Saragih, J.: Authentic volumetric avatars from a phone scan. ACM Trans. Graph. **41**(4) (Jul 2022). https://doi.org/10.1145/3528223.3530143, https://doi.org/10.1145/3528223.3530143 2
6. Caselles, P., Ramon, E., García, J., Triginer, G., Moreno-Noguer, F.: Implicit shape and appearance priors for few-shot full head reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. **47**(5), 3691–3705 (May 2025). https://doi.org/10.1109/TPAMI.2025.3540542, https://doi.org/10.1109/TPAMI.2025.3540542 4
7. Chen, X., Jiang, T., Song, J., Yang, J., Black, M.J., Geiger, A., Hilliges, O.: gdna: Towards generative detailed neural avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20427–20437 (2022) 4
8. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer (2024) 8
9. Dai, H., Pears, N., Smith, W.A.P., Duncan, C.: Statistical modeling of craniofacial shape and texture. International Journal of Computer Vision **128**(2), 547–571 (2020) 8
10. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022) 2, 4, 9, 11, 12
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3, 6
12. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021) 2, 4, 7, 9, 10
13. Filntisis, P.P., Retsinas, G., Paraperas-Papantoniou, F., Katsamanis, A., Roussos, A., Maragos, P.: Visual speech-aware perceptual 3d facial expression reconstruction from videos (2022) 4
14. Gerogiannis, D., Papantoniou, F.P., Potamias, R.A., Lattas, A., Zafeiriou, S.: Arc2avatar: Generating expressive 3d avatars from a single image via id guidance. arXiv preprint arXiv:2501.05379 (2025) 2

15. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Learning neural parametric head models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21003–21012 (2023) 1, 2, 3, 4, 7, 8
16. Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., Nießner, M.: Mononphm: Dynamic head reconstruction from monocular videos. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 3, 4, 5, 6, 9, 10, 11, 15
17. Giebenhain, S., Kirschstein, T., Rünz, M., Agapito, L., Nießner, M.: Npga: Neural parametric gaussian avatars. In: SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3-6, Tokyo, Japan (2024). https://doi.org/10.1145/3680528.3687689 2, 4, 14, 15
18. Giebenhain, S., Kirschstein, T., Rünz, M., Agapito, L., Nießner, M.: Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction (2025), https://arxiv.org/abs/2505.00615 2, 3, 5, 7, 8, 9, 10, 11, 12, 14, 15
19. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18653–18664 (2022) 3
20. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3789–3799. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/gropp20a.html 12
21. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 3
22. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: European Conference on Computer Vision. pp. 152–168. Springer (2020) 4
23. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH 2024 Conference Papers. Association for Computing Machinery (2024). https://doi.org/10.1145/3641519.3657428 4, 15
24. IEEE: A 3D Face Model for Pose and Illumination Invariant Face Recognition (2009) 3, 4
25. Ji, X., Weiss, S., Kansy, M., Naruniec, J., Cao, X., Solenthaler, B., Bradley, D.: Fastgha: Generalized few-shot 3d gaussian head avatars with real-time animation. ArXiv **abs/2601.13837** (2026), https://api.semanticscholar.org/CorpusID:284909908 2
26. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 15
27. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/ 4
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014), http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015 8

29. Kirschstein, T., Giebenhain, S., Nießner, M.: Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5481–5492 (2024) 2, 4
30. Kirschstein, T., Giebenhain, S., Nießner, M.: Flexavatar: Learning complete 3d head avatars with partial supervision. ArXiv **abs/2512.15599** (2025), https://api.semanticscholar.org/CorpusID:283920720 2
31. Kirschstein, T., Romero, J., Sevastopolsky, A., Nießner, M., Saito, S.: Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12089–12100 (October 2025) 2
32. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813 2, 3, 4, 9
33. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. p. 163–169. SIGGRAPH '87, Association for Computing Machinery, New York, NY, USA (1987). https://doi.org/10.1145/37401.37422, https://doi.org/10.1145/37401.37422 5
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 4
35. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017) 10, 12
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 3, 14
37. Palafox, P., Božič, A., Thies, J., Nießner, M., Dai, A.: Npms: Neural parametric models for 3d deformable shapes. arXiv preprint arXiv:2104.00702 (2021) 4
38. Papaioannou, Athanasios an Baris, G., Cheng, S., Chrysos, G.G., Deng, J., Fotiadou, E., Kampouris, C., Kollias, D., Moschoglou, S., Songsri-In, K., Ploumpis, S., Trigeorgis, G., Tzirakis, P., Ververas, E., Zhou, Y., Ponniah, A., Roussos, A., Zafeiriou, S.: Mimicme: A large scale diverse 4d database for facial expression analysis. In: Proceedings of the European Conference on Computer Vision (2022) 8
39. Peng, C., Su, Z., Wang, L., Guo, C., Li, Z., Long, C., Lv, Z., Sun, J., Zhang, C., Liu, Y.: Flexavatar: Flexible large reconstruction model for animatable gaussian head avatars with detailed deformation. ArXiv **abs/2512.17717** (2025), https://api.semanticscholar.org/CorpusID:284057797 2
40. Potamias, R.A., Galanakis, S., Deng, J., Papaioannou, A., Zafeiriou, S.: Imhead: A large-scale implicit morphable model for localized head modeling (2025) 4
41. Pozdeev, D., Artemov, A., Bhattarai, A.R., Sevastopolsky, A.: Densemarks: Learning canonical embeddings for human heads images via point tracks. ArXiv **abs/2511.02830** (2025), https://api.semanticscholar.org/CorpusID:282749210 2
42. Prinzler, M., Zakharov, E., Sklyarova, V., Kabadayi, B., Thies, J.: Joker: Conditional 3d head synthesis with extreme facial expressions. arXiv preprint arXiv:2410.16395 (2024) 2

43. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024) 2, 3
44. Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: 3d facial expressions through analysis-by-neural-synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 4, 12
45. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 2016 fourth international conference on 3D vision (3DV). pp. 460–469. IEEE (2016) 4
46. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1259–1268 (2017) 4
47. Rosu, R.A., Wu, K., Feng, Y., Zheng, Y., Black, M.J.: DiffLocks: Generating 3d hair from a single image using diffusion models. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition(CVPR) (2025) 15
48. Saleh, F., Aliakbarian, S., Hewitt, C., Petikam, L., Xiao-Xian, Criminisi, A., Cashman, T.J., Baltrušaitis, T.: DAViD: Data-efficient and accurate vision models from synthetic data (2025), https://arxiv.org/abs/2507.15365 8
49. Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to regress 3D face shape and expression from an image without 3D supervision. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 7763–7772 (Jun 2019) 2, 3, 4, 8, 9, 10, 12
50. Schoneveld, L., Chen, Z., Davoli, D., Tang, J., Terazawa, S., Nishino, K., Nießner, M.: Sheap: Self-supervised head geometry predictor learned via 2d gaussians. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14162–14172 (October 2025) 2, 4, 7, 9, 12
51. Tarasiou, M., Potamias, R.A., O'Sullivan, E., Ploumpis, S., Zafeiriou, S.: Locally adaptive neural 3d morphable models (2024) 4
52. Taubner, F., Raina, P., Tuli, M., Teh, E.W., Lee, C., Huang, J.: 3D face tracking from 2D video through iterative dense UV to image flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1227–1237 (June 2024) 2, 3, 9, 10
53. Taubner, F., Zhang, R., Tuli, M., Bahmani, S., Lindell, D.B.: MVP4D: Multi-view portrait video diffusion for animatable 4D avatars (2025), https://arxiv.org/abs/2510.12785 2
54. Taubner, F., Zhang, R., Tuli, M., Lindell, D.B.: Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. arXiv preprint arXiv:2412.12093 (2024) 2
55. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 1274–1283 (2017) 4
56. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016) 2, 3
57. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.: Extreme 3d face reconstruction: Seeing through occlusions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3935–3944 (2018) 4

58. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5163–5172 (2017) 4
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf 6
60. Wang, D., Chandran, P., Zoss, G., Bradley, D., Gotardo, P.: Morf: Morphable radiance fields for multiview neural head modeling. In: ACM SIGGRAPH 2022 Conference Proceedings. SIGGRAPH '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3528233.3530753, https://doi.org/10.1145/3528233.3530753 4
61. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 5, 15
62. Wang, L., Chen, Z., Yu, T., Ma, C., Li, L., Liu, Y.: Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2022) (June 2022) 4
63. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. NeurIPS (2021) 6
64. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024) 5
65. Wood, E., Baltrušaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljević, N., Wilde, D., Garbin, S., Sharp, T., Stojiljković, I., et al.: 3d face reconstruction with dense landmarks. In: European Conference on Computer Vision. pp. 160–177. Springer (2022) 2, 3
66. Xu, C., Zhao, X., Deng, X., Sun, J., Su, Z., Di, D., Liu, Y.: Geodiff4d: Geometry-aware diffusion for 4d head avatar reconstruction (2026), https://api.semanticscholar.org/CorpusID:286171509 2
67. Xu, Y., Wang, L., Zheng, Z., Su, Z., Liu, Y.: 3d gaussian parametric head model. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024) 4
68. Yenamandra, T., Tewari, A., Bernard, F., Seidel, H.P., Elgharib, M., Cremers, D., Theobalt, C.: i3dmm: Deep implicit 3d morphable model of human heads. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12803–12813 (2021) 2, 4
69. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: CelebV-Text: A large-scale facial text-video dataset. In: CVPR (2023) 8
70. Zhang, T., Chu, X., Liu, Y., Lin, L., Yang, Z., Xu, Z., Cao, C., Yu, F., Zhou, C., Yuan, C., et al.: Accurate 3d face reconstruction with facial component tokens. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9033–9042 (2023) 2, 4, 6, 9, 10
71. Zheng, M., Yang, H., Huang, D., Chen, L.: Imface: A nonlinear 3d morphable face model with implicit neural representations. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 20343–20352 (2022) 2, 4
72. Zheng, M., Zhang, H., Yang, H., Chen, L., Huang, D.: Imface++: A sophisticated nonlinear 3d morphable face model with implicit neural representations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 4
73. Zheng, X., Wen, C., Li, Z., Zhang, W., Su, Z., Chang, X., Zhao, Y., Lv, Z., Zhang, X., Zhang, Y., Wang, G., Lan, X.: Headgap: Few-shot 3d head avatar via generalizable gaussian priors. arXiv preprint arXiv:2408.06019 (2024) 4
74. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: CelebV-HQ: A large-scale video facial attributes dataset. In: ECCV (2022) 8
75. Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Wu, M., Shen, Q., Yang, R., Cao, X.: Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2023) 4, 8
76. Zielonka, W., Bolkart, T., Beeler, T., Thies, J.: Gaussian eigen models for human heads (2025), https://arxiv.org/abs/2407.04545 4
77. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: European conference on computer vision. pp. 250–269. Springer (2022) 2, 3, 4, 9, 11