

Pixel3DMM: Versatile Screen-Space Priors for Single-Image 3D Face Reconstruction

Simon Giebenhain¹ Tobias Kirschstein¹ Martin Rünz²
Lourdes Agapito³ Matthias Nießner¹

¹Technical University of Munich ²Synthesia ³University College London

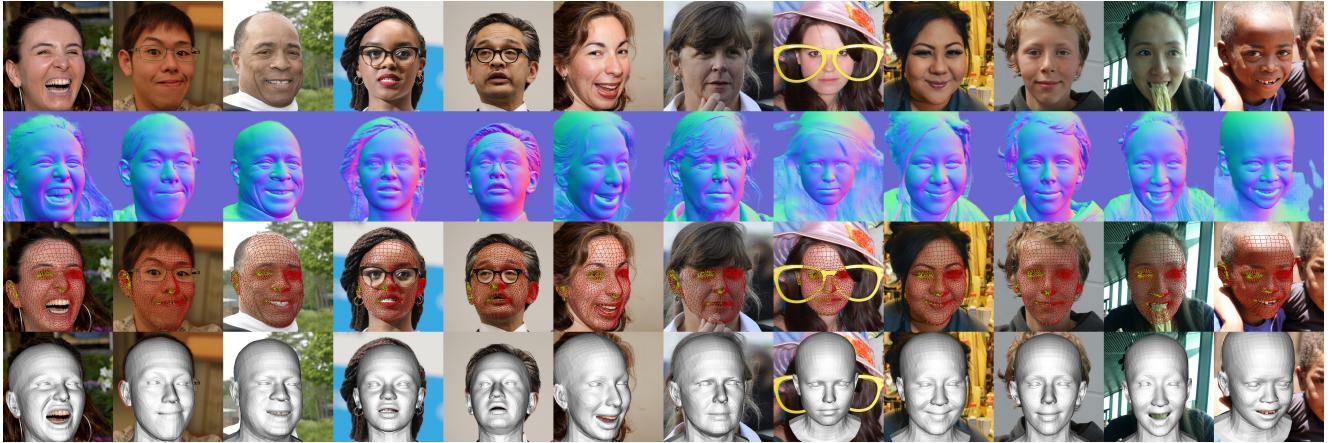


Figure 1. We present Pixel3DMM, a set of two ViTs [9], which are tailored to predict per-pixel surface normals and uv-coordinates. Here, we demonstrate the fidelity and robustness of our prior networks on examples from the FFHQ [20] dataset. From top to bottom we show input RGB, predicted surface normals, 2D vertices extracted from the uv-coordinate prediction, and our FLAME fitting results.

Abstract

We address the 3D reconstruction of human faces from a single RGB image. To this end, we propose Pixel3DMM, a set of highly-generalized vision transformers which predict per-pixel geometric cues in order to constrain the optimization of a 3D morphable face model (3DMM). We exploit the latent features of the DINO foundation model, and introduce a tailored surface normal and uv-coordinate prediction head. We train our model by registering three high-quality 3D face datasets against the FLAME mesh topology, which results in a total of over 1,000 identities and 976K images. For 3D face reconstruction, we propose a FLAME fitting optimization that solves for the 3DMM parameters from the uv-coordinate and normal estimates. To evaluate our method, we introduce a new benchmark for single-image face reconstruction, which features high diversity facial expressions, viewing angles, and ethnicities. Crucially, our benchmark is the first to evaluate both posed and neutral facial geometry. Ultimately, our method outperforms the most competitive baselines by over 15% in terms

of geometric accuracy for posed facial expressions.

1. Introduction

3D reconstruction of faces, tracking facial movements, and ultimately extracting expressions for animation tasks are fundamental problems in many domains such as computer games, movie production, telecommunication, and AR/VR applications. Recovering 3D head geometry from a single image is a particularly important task due to the vast amount of available image collections.

Unfortunately, reconstructing faces from a single input image is also inherently under-constrained. Not only depth ambiguity renders this task challenging, but also ambiguities between albedo and lighting/shadow effects. In addition, properly disentangling identity and expression information – which is critical for many downstream applications – makes the problem difficult. Finally, occlusions and unobserved facial regions further complicate the problem in real application scenarios, thus highlighting the need for strong data priors.

A typical approach to address single-image face reconstruction is to exploit 3D parametric head models

(3DMMs) [3, 26] which provide a comparatively low-dimensional parametric representation for the underlying 3D geometry. Optimizing within a 3DMM’s disentangled parameter space heavily constrains the search space with built-in assumptions about plausible facial structure and expressions, and allows to extract disentangled identity and expression information. Nonetheless, despite relying on 3DMMs, many ambiguities remain and their simplifying assumptions about our world often cannot explain the complexity of an observed RGB signal. This necessitates additional priors in order to obtain compelling fitting results such as sparse [37] and dense [5, 47] facial landmarks, or UV coordinate predictions [41].

In recent years, we have also seen significant progress in feed-forward 3DMM regressors [8, 10, 34, 38, 50, 53]. However, it is complicated to extend feed-forward regressors, *e.g.* to a multi-view or temporal domain, and, as we will show later, they fall behind optimization-based approaches on inputs with strong facial expressions. Overall, accurate 3D face reconstruction from single images remains a challenging and highly relevant problem.

Therefore, we propose Pixel3DMM, a novel optimization-based 3D face reconstruction approach. Our main idea is to exploit and further develop broadly generalized and powerful foundation models to predict pixel-aligned geometric cues that effectively constrain the 3D state of an observed face. Given a single image at test time, we propose normal and uv-coordinate predictions as optimization constraints from which we fit a 3D FLAME model. Instead of a simple rendering loss of uv-coordinates, we then transfer the information into a 2D vertex loss, which offers a wider basin of attraction during optimization. We argue that this strategy is superior to traditional photometric terms, or sparse landmarks, which often struggle with extreme view points and facial expressions. In order to train our approach, we unify three recent, high-fidelity 3D face datasets [13, 29, 52] by registering them against the FLAME [26] model. Our approach outperforms all available normal estimators for human faces in the NeRSemle [24] dataset.

In order to advance the evaluation of single-image 3D face reconstruction methods, we further propose a new benchmark based on the multi-view video dataset NeRSemle [24], which includes a wider variety of facial expressions than existing benchmarks [6, 11, 38, 52]. Our benchmark is the first to allow for the simultaneous evaluation of posed and neutral facial geometry. This enables a more direct comparison of methods, especially regarding fitting fidelity and ability to disentangle expression and identity information. Finally, we show that compared to our strongest baselines, our approach improves the L2-Chamfer reconstructions loss by over 15% for posed geometry, while slightly improving over neutral geometry predictions.

To summarize, our main contributions are as follows:

- A new formulation to exploit foundation model features for 3D-related, pixel-aligned predictions, facilitating state-of-the-art normal estimations for human faces.
- A novel 3D face reconstruction approach based on predicted uv-map correspondences and surface normals.
- A 3D face reconstruction benchmark and evaluation protocol from high-fidelity multi-view face captures.

We plan to make the model, code, and our new benchmark publicly available to promote progress in single image 3D face reconstruction and encourage quantitative benchmarking on challenging facial expressions.

2. Related Work

Single-Image 3DMM Fitting Tracking morphable models from single images is a well-studied problem in the context of 3D face reconstruction and tracking. Early works [2, 26, 31], introduced statistical shape and texture priors to estimate 3D face geometry from 2D images. Such methods rely on photometric fitting and subsequent approaches improve modeling capabilities using learned implicit representations [14, 27]. While some methods [15, 43] favor a high tracking frame rate for real-time applications, others favor reconstruction accuracy [53].

Facial Landmark Prediction Numerous reconstruction methods [5, 26] for faces rely on accurate landmark predictions, which are usually coupled with vertices of a template mesh. Pioneering work on detecting such landmarks already relies on statistical learning [7] and more recent models exploit large datasets [46, 48] and neural networks to improve the performance [1, 4]. MediaPipe [1], for instance, uses a convolutional network inspired by MobileNet [17].

Another line of work focuses on densely aligning template mesh and 2D predictions. To achieve this FlowFace [41] employs a vision-transformer backbone and iteratively refines the flow from UV to image space.

3DMM Regression DECA [10] trains an encoder for 3DMM parameters that also outputs a displacement map for higher fidelity. An extension of this work is presented in EMOCA [8], which adds a head for facial expressions to the architecture and emphasises on the reconstruction of emotion-rich data. SPECTRE [12] too builds on top of DECA, but aims at temporal consistency and reconstructing lip motion truthfully. To improve the analysis-by-synthesis aspect of previous methods SMIRK [34] introduces a neural synthesis component, reducing the domain gap between real and rendered images. Since the aforementioned methods don’t assume 2D to 3D correspondences for training, it is easy to scale them to large datasets. As a downside, the lack of 3D information impedes accuracy and leaves depth ambiguity. In order to address this, MICA [53] supervises

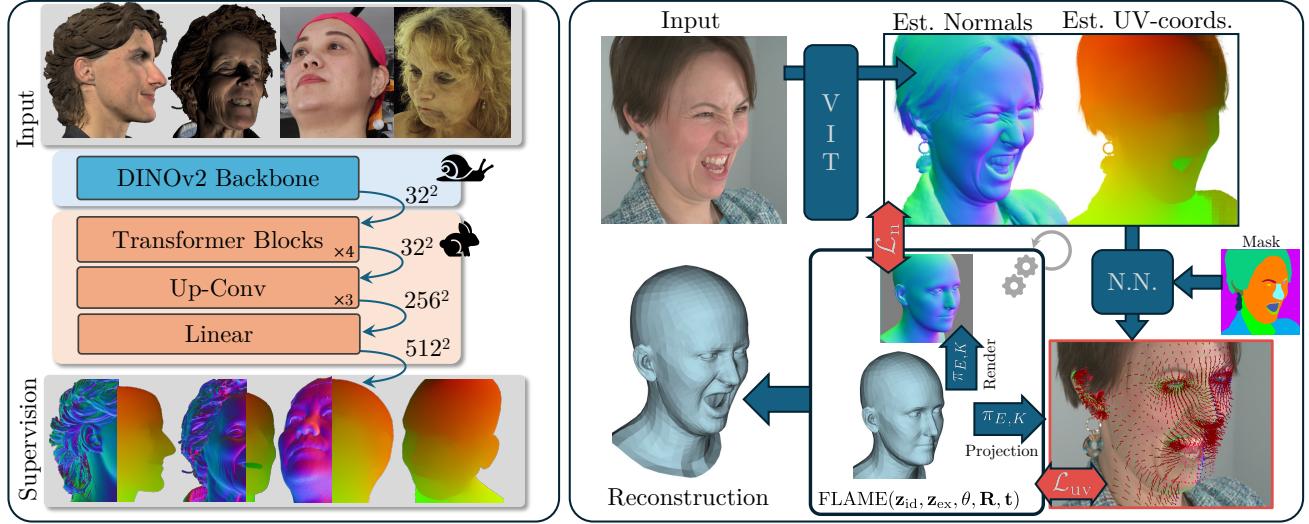


Figure 2. **Method Overview:** Pixel3DMM consists of (a) learning pixel-aligned geometric priors (left) and (b) test-time optimization against predicted uv-coordinates and normals (right). On the left we illustrate our network architecture and examples from the training set. On the right we illustrate to process finding per-vertex 2D locations using a nearest neighbor (N.N.) look up, and our loss terms.

directly on meshes and presents a 3D face dataset to do so. TokenFace [50] is a transformer-based hybrid method that can be trained on both 2D and paired 3D data.

Face Reconstruction Benchmarks Benchmarks with high-quality ground truth reconstructions are necessary to compare methods reliably. The Stirling [11] dataset contains 2000 images of 135 subjects. Unfortunately, ground truth reconstructions are only available for neutral poses in this dataset. Similarly, the NoW [38] benchmark provides reconstructions only in the neutral expression. It has 2054 images of 100 subjects and 3D models recorded with an active stereo 3dMD system. Both the FaceScape [52] and the REALY [6] dataset contain posed scans. While the former has 10 identities, the latter has 100 subjects. Neither of these two benchmarks measures disentanglement by additionally evaluating against neutral geometry.

3. Pixel3DMM

In this work we address the challenges of single-image face reconstruction by learning powerful priors of pixel-aligned geometric cues. In particular we train two vision transformer networks, which predict uv-coordinates and surface normals against which we fit FLAME [26] parameters at inference time. In Sec. 3.1 we describe our Pixel3DMM networks, our data acquisition, and how we train them for accurate surface normal and uv-coordinate prediction. Afterwards, in Sec. 3.2, we elaborate on our single-image fitting approach, which is purely based on our surface normal and uv-coordinate predictions.

3.1. Learning Pixel-Aligned Geometric Cues

Despite recently released high-quality 3D face datasets [13, 24, 29, 52], such data is still relatively scarce, especially w.r.t. the number of different identities, ethnicities, age distribution and lighting variation. We therefore take inspiration from recent achievements on fine-tuning foundational and large generative models to become experts on a constrained domain, *e.g.* [18, 36].

In particular we train two expert networks

$$\mathcal{N} : \mathbb{R}^{512 \times 512 \times 3} \rightarrow [-1, 1]^{512 \times 512 \times 3} \quad (1)$$

$$\mathcal{U} : \mathbb{R}^{512 \times 512 \times 3} \rightarrow [0, 1]^{512 \times 512 \times 2} \quad (2)$$

which, given a single input image I , predict surface normals $\mathcal{N}(I)$ and uv-space coordinates $\mathcal{U}(I)$, respectively.

3.1.1. Network Architecture

We build Pixel3DMM on top of the foundational features from a pre-trained DINOv2 [30] backbone. As depicted in Fig. 2, we extend the ViT architecture using a simple prediction head. It consists of four additional transformer blocks, three up-convolutions which lift the feature map resolution from 32×32 to 256×256 . Finally, we use a single linear layer to increase the feature dimensionality and unpatchify the predictions to $512 \times 512 \times c$, where $c \in \{3, 2\}$ for normals and uv-coordinate prediction tasks, respectively.

3.1.2. Data Preparation

To train our networks, we opt for three recent, high-quality 3D face datasets: NPHM [13], FaceScape [52], and Ava256 [29]. We follow the non-rigid registration procedure from NPHM, to obtain the same kind of high-

quality registrations in FLAME topology for FaceScape and Ava256.

Fig. 2 shows pairs of input views with the associated supervision signal for surface normals and uv-coordinates. Since Ava256 does not provide high-fidelity geometry, we exclusively use it to supervise our UV-network \mathcal{U} . Since the NPHM dataset only consists of textured meshes, we render 40 random views randomly distributed on the frontal hemisphere using randomized intrinsics and camera distances. Additionally, we randomly sample lighting conditions (using point lights) and material parameters for each rendering.

Dataset Numbers In total, our dataset comprises 470 identities from NPHM in 23 expression and 40 renderings each (376K rgb, normal and uv images in total). For FaceScape we use 350 subjects, observed under 20 different expressions and 50 cameras each (350K rgb, normal and uv images in total). Since Ava256 is a video dataset, we leverage furthest point sampling to select the 50 most diverse expressions per person. For each person we choose a random subset of 20 cameras (250K rgb and uv images in total).

Diffusion-based Lighting Variations Since FaceScape and Ava256 are both studio datasets, which are captured at rather homogeneous lighting conditions, we leverage IC-Light [49], an image conditioned diffusion model [35], which alters the lighting condition based on a text prompt or background image.

3.1.3. Training

We train our models $\mathcal{M} \in \{\mathcal{N}, \mathcal{U}\}$ using a straight forward image translation formulation

$$\operatorname{argmin}_{\Psi_{\mathcal{M}}} \sum_{k \in \mathcal{D}} \sum_{p \in M^k} \|f(I^k)_p - Y_p^k\|_2, \quad (3)$$

where $\Psi_{\mathcal{M}}$ denotes the network's parameters, $k \in \mathcal{D}$ is a sample from our dataset, I^k and Y^k are input rgb and target images, respectively, and $p \in M^k$ are all pixels in the associated foreground mask.

Note, that instead of freezing the parameters of our DINOv2 backbone altogether, we set their learning rate ten times lower, in order to encourage prior preservation but enable stronger domain adoption.

Compared to Sapiens [22], a recent state-of-the-art foundation model for human bodies, training our models is cheap and can be realized using 2 GPUs and training for 3 days. Additionally, we highlight the fact that all data is publicly available. The relatively low computational burden and data accessibility, will hopefully inspire more research to follow in a similar direction. Finally, note that uv-coordinates are an abstract concept, without a strong correlation to rgb data, requiring a more global understanding of the input. Therefore, we demonstrate that the available data

is enough to achieve generalization on complicated, semi-global prediction tasks.

3.2. Single-Image FLAME[26] Fitting

Given a single image I , we leverage our prior networks to obtain predicted surface normals $\mathcal{N}(I)$ and uv-coordinates $\mathcal{U}(I)$. Using these predictions we aim to recover 3DMM parameters. In particular, we optimize for FLAME [26] identity, expression, and jaw parameters, as well as, camera rotation, translation, focal length and principal point:

$$\Omega_{\text{FLAME}} = \{\mathbf{z}_{\text{id}} \in \mathbb{R}^{300}, \mathbf{z}_{\text{ex}} \in \mathbb{R}^{100}, \theta \in \mathcal{SO}(3)\} \quad (4)$$

$$\Omega_{\text{cam}} = \{\mathbf{R} \in \mathcal{SO}(3), \mathbf{t} \in \mathbb{R}^3, \mathbf{f} \in \mathbb{R}^+, \mathbf{pp} \in \mathbb{R}^2\}. \quad (5)$$

3.2.1. 2D Vertex Loss

Using the estimated uv-coordinates $\mathcal{U}(I)$, we aim to extract the 2d location p_v^* for each visible vertex $v \in V$ of the FLAME mesh. To this end we first run a facial segmentation network [51], in order to mask out the background, eyeballs and mouth interior. Then we find correspondences for each vertex $v \in V$ using a nearest neighbor lookup into $\mathcal{U}(I)$. To be more specific let $T_v^{\text{uv}} \in [0, 1]^2$ denote the uv-coordinate of v in the template mesh T . Then we find the pixel location

$$p_v^* = \operatorname{argmin}_{p \in P} \|T_v^{\text{uv}} - \mathcal{U}(I)_p\| \quad (6)$$

as the pixel with the closest uv prediction. Finally, we define

$$\mathcal{L}_{\text{uv}} = \sum_{v \in V} \mathbb{1}_{\|T_v^{\text{uv}} - \mathcal{U}(I)_p\| < \delta_{\text{uv}}} \cdot |p_v^* - \pi(v)| \quad (7)$$

to be our 2d vertex loss, where $\mathbb{1}$ denotes the indicator function which masks out vertices with a nearest neighbor distance larger than δ_{uv} . $V = \text{FLAME}(\Omega_{\text{FLAME}})$ is the current estimate of the FLAME parametric model, and π denotes the projection implied by the current estimate of the camera parameters Ω_{cam} .

3.2.2. Optimization

Next to the 2d vertex loss \mathcal{L}_{uv} , we include the normal loss $\mathcal{L}_n = |\mathcal{N}(I) - \text{render}_n(V)|$, where render_n denotes a rendering of surface normals of the FLAME mesh. The regularization term $\mathcal{R} = \lambda_{\text{id}} \|\mathbf{z}_{\text{id}} - \mathbf{z}_{\text{id}}^{\text{MICA}}\|_2^2 + \lambda_{\text{ex}} \|\mathbf{z}_{\text{ex}}\|_2^2$ completes our overall energy term

$$E = \lambda_{\text{uv}} \mathcal{L}_{\text{uv}} + \lambda_n \mathcal{L}_n + \mathcal{R}. \quad (8)$$

Here $\mathbf{z}_{\text{id}}^{\text{MICA}}$ denotes MICA's [53] identity prediction.

3.3. Monocular Video Tracking

Next to the single-image scenario, tracking faces in monocular videos is a fundamental task in computer vision. To address this problem, we simply extend our optimization strategy from Sec. 3.2.2 globally over all images in a video



Figure 3. 3D Face Reconstruction Benchmark Analysis. We show the 5 most diverse images from each benchmark dataset, as measured by the expression codes of EMOCA [8]. Our benchmark covers a richer diversity of facial expressions.

sequence $\{I_t\}_{t=1}^T$. Using our prior networks, we first obtain normal predictions $\{\mathcal{N}(I_t)\}$ and uv-predictions $\{\mathcal{U}(I_t)\}$. After obtaining an initial estimate for $\Omega_{\text{FLAME}}^{(0)}$ and $\Omega_{\text{cam}}^{(0)}$ on the first frame by optimizing for Eq. (8), we freeze \mathbf{z}_{id} , \mathbf{f}_l and \mathbf{pp} . We then sequentially optimize for all remaining attributes in $\Omega_{\text{FLAME}}^{(t)}$ and $\Omega_{\text{cam}}^{(t)}$. Using the results from the sequential optimization pass as initialization, we extend Eq. (8) to a batched version including a random sample of $B = \min(T, 16)$ frames. Note, that the parameters \mathbf{z}_{id} , \mathbf{f}_l and \mathbf{pp} are shared for all frames. In order to enforce smoothness across all per-frame optimization targets we add a smoothness term

$$\mathcal{L}_{\text{smooth}}^{\Phi} = \frac{\lambda_{\text{smooth}}^{\Phi}}{2 * B} \sum_{t \in B} \|\Phi^{(t-1)} - \Phi^{(t)}\|_2^2 + \|\Phi^{(t)} - \Phi^{(t+1)}\|_2^2 \quad (9)$$

to our optimization energy E , where we let $\Phi^{(t)} \in \{\mathbf{z}_{\text{ex}}^{(t)}, \theta^{(t)}, \mathbf{R}^{(t)}, \mathbf{t}^{(t)}\}$ denote any of the per-frame variables.

	Year	neutr.	expr. div.	view div.	#pers.	#Scans
Stirling [11]	2013	✓		✓	133	133
REALY [6]	2015				100	100
NoW [38]	2019	✓		✓	80	80
FaceScape [52]	2020			✓	✓	20
Ours	2023	✓	✓	✓	21	441

Table 1. Comparison of 3D Face Reconstruction Benchmarks. We compare data capture year, whether the benchmark evaluates disentanglement by predicting a neutral mesh from a posed image (neutr.), expression diversity (div. expr.), viewpoint diversity (div. views), number of persons (#pers.) and number of GT scans.

4. 3D Face Reconstruction Benchmark

Human face geometry is complex due to the presence of thin structures, different textures and diverse shapes. Further-

more, humans can deform their facial geometry in a remarkable way, performing a wide range of expressions and emotions. Consequently, building a robust 3D face reconstruction pipeline that covers all potential states of a human face is a challenging endeavor. Several 3D face reconstruction benchmarks have been previously proposed to rank reconstruction methods in terms of quality and robustness. Tab. 1 shows a comparison of popular benchmarks. However, we find that most existing benchmarks rarely evaluate extreme facial expressions, an important aspect of human face geometry. This can be seen in Fig. 3 where we retrieve the 5 most expressive images from the recent FaceScape benchmark [52] and the established NoW benchmark [38]. We do this by running EMOCA [8] on each image of the dataset, collecting the expression codes, and then performing furthest point sampling in EMOCA’s expression space, starting from the expression with highest norm. We find that FaceScape only contains 20 different but relatively articulated expressions while the NoW benchmark is dominated by mostly neutral and smiling expressions. We therefore propose a new benchmark for 3D face reconstruction that is sourced from images of the recently published multi-view video dataset NeRSemble [24]. For 21 diverse identities, we select 20 distinct expressions via furthest point sampling of 3D landmarks for a total of 420 images. The corresponding ground truth 3D geometries are obtained by running COLMAP [39] on the full resolution 3208x2200 images. Additionally, we compute one pointcloud for a neutral frame of each person, yielding 441 ground truth 3D geometries in total.

4.1. Task Description

Our benchmark consists of two 3D face reconstruction tasks: *posed* and *neutral* 3D face reconstruction. The posed reconstruction task aims to measure the fidelity of a 3D reconstruction. Given any expressive face image, the underlying geometry shall be recovered. This requires images with paired ground truth geometries which are available in NeRSemble through COLMAP. The neutral reconstruction task on the other hand is specific to the face domain and measures how well a reconstruction method can disentangle the effects of shape and expression on a human 3D face. Specifically, the task is to reconstruct the geometry of a person’s face under neutral expression given an image of the person under any arbitrary expression.

4.2. Evaluation Protocol

To measure the performance of a reconstructed posed or neutral 3D face, we follow established practice and first rigidly align the prediction to the ground truth pointcloud via landmark correspondences and ICP. Furthermore, we use segmentation masks [51] to remove non-facial areas (hair, neck, ears, and mouth interior) from the ground truth.

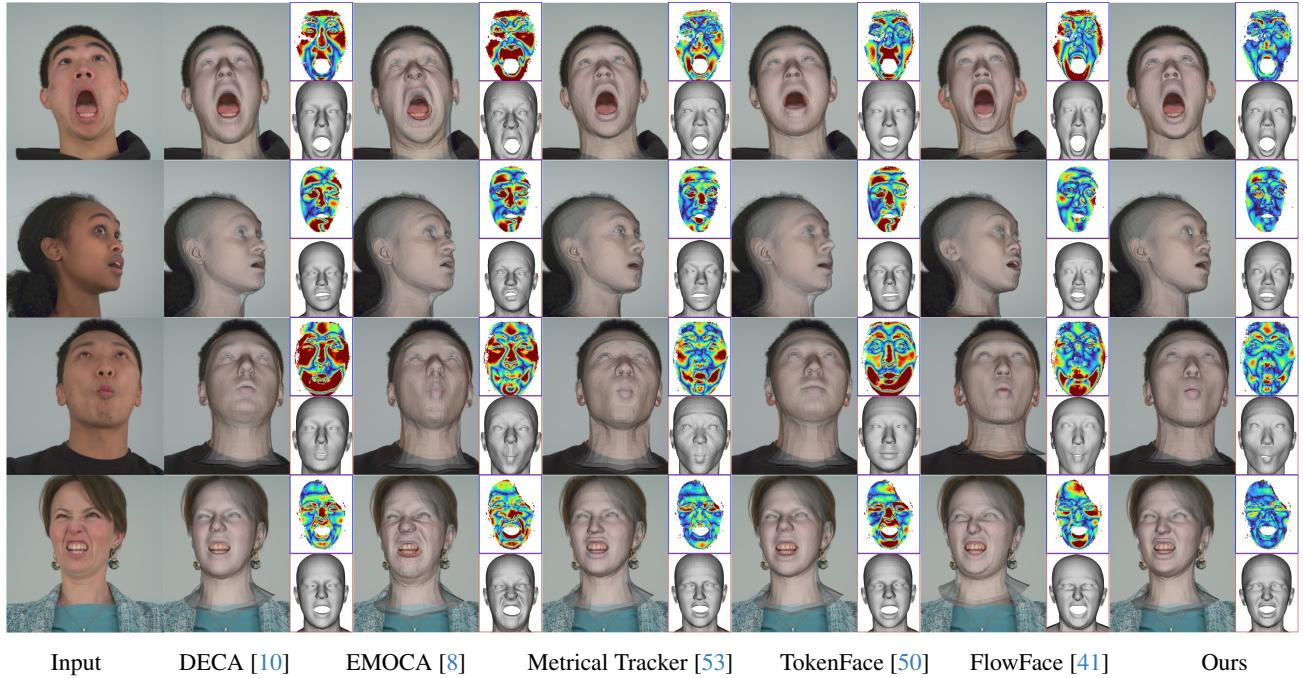


Figure 4. **Qualitative Comparison (Posed):** We show overlays of the reconstructed meshes to judge the reconstruction alignment. Insets with a blue border depict L_2 -Chamfer distance as an error map, rendered from a frontal camera. Red insets show the reconstructed mesh from the same camera. We encourage the reviewers to watch our supplementary material for additional visualizations.

We then compute three metrics: (i) uni-directional Chamfer distance (L1 and L2) from GT points to the nearest mesh surface, (ii) cosine similarity (NC) of predicted mesh normals and GT pointcloud normals, (iii) Recall thresholded at 2.5mm ($R^{2.5}$) which is the percentage of GT points whose nearest mesh surface is 2.5mm or closer.

5. Experimental Results

5.1. Implementation Details

Prior Learning We train Pixel3DMM using the Adam [23] optimizer, a batch size of 40, and 2 A6000 GPUs, which takes 3 days until convergence. We use a learning rate of 1×10^{-4} for the prediction head and 1×10^{-5} for the DINO backbone. For simplicity we choose a light-weight network head. Using a DPT [33] head instead resolves the last remaining patch artifacts of the ViT-Base backbone but drastically increases runtime without improving down-stream reconstruction performance. Similarly, we find that replacing ViT-Base with Sapiens-300M [22] backbone (the smallest available Sapiens model) incurs high computational costs without reconstruction benefits. We use 10% of the subjects as validation set, and exclude all the subjects from our benchmark from the training set.

FLAME Fitting We use the Adam optimizer with $\text{lr}_{\text{id}} = 0.001$ and $\text{lr}_{\text{ex}} = 0.003$. We set $\lambda_{uv} = 2000$, $\lambda_n = 200$, $\lambda_{\text{id}} = 0.15$ and $\lambda_{\text{ex}} = 0.01$. We perform 500 optimization steps which takes 30 seconds in our unoptimized implementation.

5.2. Baselines

Feed-Forward FLAME Regressors The first category of approaches we compare against are feed-forward neural networks trained to predict FLAME parameters. We choose DECA [10] and EMOCA [8] as baselines which are trained in a self-supervised fashion on 2D data only. Additionally, we compare against MICA [53], which is trained solely on 3D data and only predicts identity parameters \mathbf{z}_{id} , and TokenFace [50] which trained on a mixture of 2D and 3D data. Since TokenFace is not publicly available, the authors ran their method on the images that we provided.

Optimization-Based Approaches We compare against MetricalTracker [53], which optimizes against two sets of facial landmark predictions [4, 5] and a photometric term. Additionally, we compare against FlowFace [41], a recent method that predicts flow from the uv-space into image space, in order to predict 2D image-space vertex positions. Similar to Pixel3DMM, FlowFace also uses a dense 2D vertex loss, but predicts them in a quite different manner. Note that all methods in this category rely on MICA estimates to

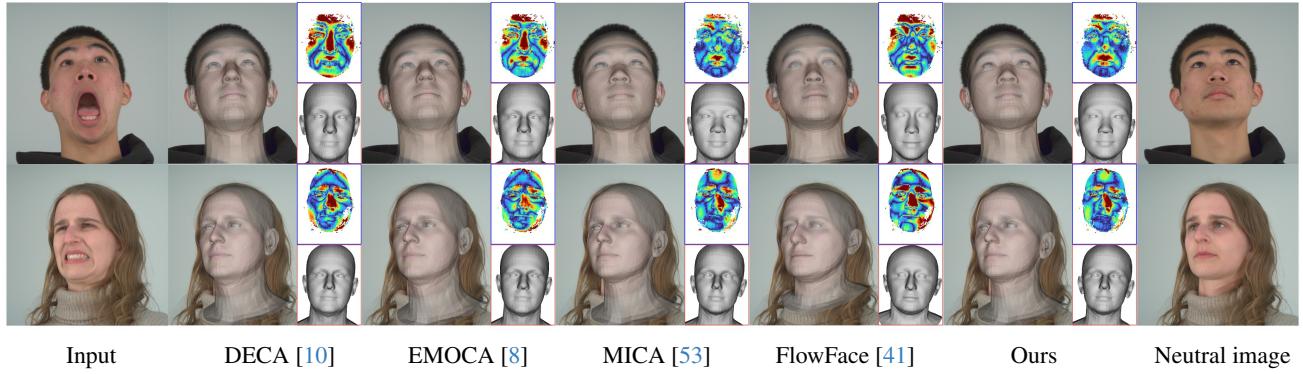


Figure 5. **Qualitative Comparison (Neutral):** Alignment of the neutral prediction against the neutral image and scan of a person.

	Neutral				Posed			
	L1↓	L2↓	NC↑	R ^{2.5} ↓	L1↓	L2↓	NC↑	R ^{2.5} ↓
MICA [53]	1.68	1.14	0.883	0.910	-	-	-	-
TokenFace [50]	-	-	-	-	2.62	1.78	0.865	0.768
DECA [10]	2.07	1.40	0.876	0.845	2.38	1.61	0.870	0.798
EMOCAv2[8]	2.21	1.49	0.873	0.824	2.63	1.78	0.860	0.758
Metr. Tracker	-	-	-	-	2.03	1.37	0.878	0.857
FlowFace [41]	1.93	1.31	0.878	0.870	1.96	1.33	0.879	0.879
Ours	1.66	1.12	0.883	0.912	1.66	1.11	0.884	0.916

Table 2. **Quantitative Comparison on Our Benchmark.**

initialize \mathbf{z}_{id} . Since FlowFace is not publicly available, as of yet, the authors ran their method on our benchmark images. In the future, we hope that our proposed benchmark will be adopted as a standard by the community to encourage further quantitative comparisons across methods.

5.3. Our Benchmark

Posed Face Reconstruction We present quantitative and qualitative results for the posed reconstruction task (see Sec. 4.1) in Tab. 2 and Fig. 4, respectively. Quantitatively, Pixel3DMM outperforms all baselines by a large margin. In general, the feed-forward predictors (DECA, EMOCAv2, TokenFace) perform significantly worse than the optimization based approaches (MetricalTracker, FlowFace and Ours). Visually, DECA and TokenFace seem to underfit facial expressions, while EMOCAv2 exaggerates them. Compared to our approach, FlowFace sometimes exhibits performance drops for extreme facial expressions.

Neutral Face Reconstruction Results on the neutral reconstruction task (see Sec. 4.1) are provided in Fig. 5 and Tab. 2. First of all, we can observe that the significantly better posed reconstruction metrics of FlowFace and Pixel3DMM do not immediately translate to the neutral reconstruction. We attribute this to the ambiguities between identity and expression in the optimization process. Note

Method	NoW [38]			FaceScape [52]		
	Median↓	Mean↓	Std↓	CD↓	MNE↓	CR↑
Dense [47]	1.02	1.28	1.08	-	-	-
PRNet [45]	-	-	-	3.56	0.126	0.896
3DDFAv2 [16]	-	-	-	3.60	0.096	0.931
DECA [10]	1.09	1.38	1.18	4.69	0.108	0.995
MICA [53]	0.90	1.11	0.92	-	-	-
FlowFace [41]	0.87	1.07	0.88	2.21	0.083	-
TokenFace [50]	0.76	0.95	0.82	3.70	0.101	0.938
Ours	0.87	1.07	0.89	1.76	0.077	0.980

Table 3. **NoW [38] and FaceScape [52] Benchmark.**

that both FlowFace and Pixel3DMM rely on MICA predictions to initialize identity parameters \mathbf{z}_{id} . While FlowFace ends up with worse neutral reconstructions, our approach is able to improve upon MICA by a small margin. Nevertheless, we highlight the importance of using MICA to help disambiguate between \mathbf{z}_{id} and \mathbf{z}_{ex} , as ablated in Sec. 5.7. Note, that TokenFace is missing from the neutral evaluation, since the authors only provided posed meshes.

5.4. Results on Existing Benchmarks

FaceScape Benchmark [52] The FaceScape benchmark only evaluates the posed reconstruction task. The relative performance across methods matches with results on our benchmark, see Tab. 3. Our method outperforms all baselines by a large margin w.r.t. chamfer distance (CD) and mean normal error (MNE), and has a slightly worse completeness rate (CR) than DECA, see [52] for more details.

NoW Benchmark [38] On the NoW benchmark, which only evaluates the neutral reconstruction task, we achieve the same metrics as FlowFace, which is the best-performing optimization based approach, but perform worse than TokenFace. Note, however, that on FaceScape and our benchmark, we significantly outperform TokenFace. Similarly to

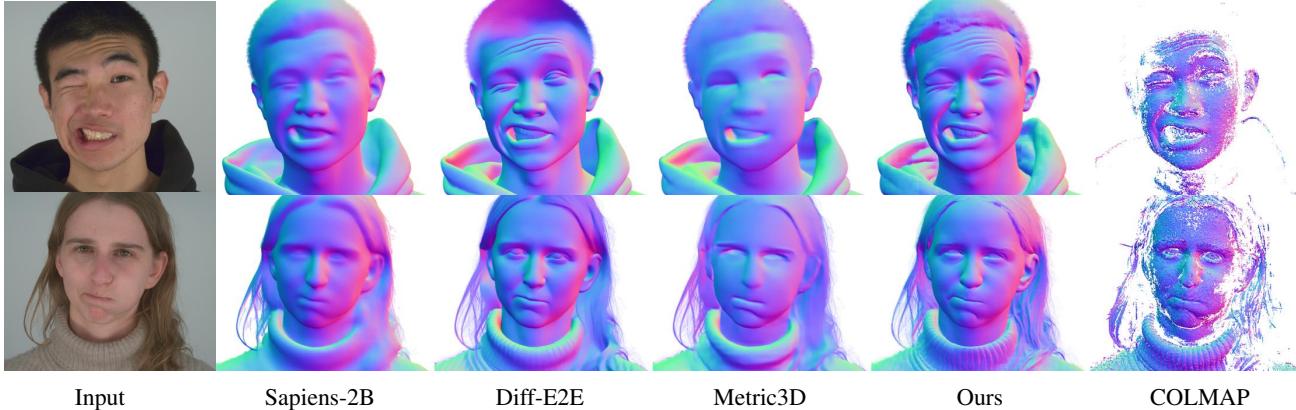


Figure 6. **Surface Normal Estimation:** Qualitative comparison to state-of-the-art surface normal estimators. From left to right we show the single input image and the predictions of Metric3D [19], Sapiens-2B [22], Diff-E2E [28], our result and COLMAP [39] normals.

	Metric3D[19]	Sapiens-2B[22]	Diff-E2E[28]	Ours
Normal Sim. \uparrow	0.900	0.913	0.913	0.931

Table 4. **Normal Estimation:** We report the cosine similarity of predicted normals against 16-view COLMAP [39] estimates. The results are averaged over all images from our benchmark.

the results on our benchmark, Pixel3DMM can only improve a small amount on top of the MICA predictions. We hypothesize that our prior significantly helps posed reconstructions, but struggles to guide the optimization to properly disentangle between \mathbf{z}_{id} and \mathbf{z}_{ex} .

5.5. In-the-Wild Results

In Fig. 1, we demonstrate the robustness of our prior networks and fitting algorithm on challenging in-the-wild examples, including strong appearance variation, various background contexts and surroundings, lighting/shadow effects, and occlusions such as glasses, head wear and hands. Ultimately, this demonstrates that our approach successfully generalizes, even beyond the training data distribution. We hope that this will inspire more work in a similar direction, especially since all data is available and 2 48GB GPUs are sufficient for training.

For tracking results on in-the-wild monocular videos we refer the reader to our supplementary video.

5.6. Surface Normal Estimation

In Tab. 4 and Fig. 6, we show quantitative and qualitative comparisons against recent state-of-the-art normal estimation methods [19, 22, 28]. Our network estimates more detailed and accurate normals than the baselines.

	Neutral			Posed		
	L1 \downarrow	L2 \downarrow	R $^{2.5}\downarrow$	L1 \downarrow	L2 \downarrow	R $^{2.5}\downarrow$
Lmks.	1.68	1.14	0.911	2.02	1.37	0.857
Lmks. + Pho.	1.69	1.14	0.908	2.05	1.38	0.854
Ours, only \mathcal{U}	1.66	1.11	0.913	1.72	1.16	0.906
Ours, only \mathcal{N}	1.69	1.12	0.907	1.70	1.14	0.910
Ours, only Sapiens	1.72	1.16	0.902	1.81	1.23	0.890
Ours	1.66	1.12	0.912	1.66	1.11	0.916
Ours, no MICA	1.90	1.29	0.872	1.74	1.17	0.901

Table 5. **Fitting Algorithm Ablations:** We compare different compositions of our optimization energy E , see Eq. (8).

5.7. Ablation Experiments

We conduct extensive ablations on different compositions of our optimization energy E in Tab. 5. We start by using the simplest energy, with only the landmark loss from MetricalTracker, and our regularization term. Next we add a photometric term, as in MetricalTracker. As shown in Tab. 5, these configurations achieve significantly worse posed reconstructions. Next, we investigate the effect of only using the predictions from \mathcal{N} and \mathcal{U} , respectively. Compared to our full model these variants showcase lower posed reconstruction scores. We also compare our normal predictor \mathcal{N} against Sapiens-2B [22], which confirms that our improved normal predictions translate to better reconstructions. Finally, we ablate the effect of using the MICA prior. Without MICA’s predictions of \mathbf{z}_{id} especially the neutral reconstruction metrics drop, indicating its importance for disentanglement between identity and expression.

6. Limitations and Future Work

While we demonstrate the effectiveness of our approach for single image 3D reconstruction, several limitations remain. While our optimization energy could be easily extended to incorporate observations from multiple viewpoints, our prior models cannot currently exploit multiview information. Future extensions of our architecture could include multiview inputs similar to DUS3R [44], or video inputs similar to RollingDepth [21]. Next, for training large-scale 3DMM conditioned generative models like 3D GANS [40] or diffusion models [25, 32, 42], e.g. on the LAION-Face dataset [51], fast reconstruction speed would be desirable. One potential avenue could be the distillation of our per-pixel predictors into a feed-forward 3DMM predictor. Finally, our experiments showcase, that optimization based approaches cannot flawlessly disambiguate identity and expression parameters. Therefore, specifically crafted priors for disambiguation are required.

7. Conclusion

In this paper, we trained pixel-aligned geometric prior networks, by leveraging pre-trained, generalized foundational features on publicly available 3D face datasets, which we registered into a uniform format. Our trained networks successfully generalize beyond the diversity of the training data, and we experimentally show that our normal predictor significantly outperforms all available normal estimators. We designed a 3DMM fitting algorithm on top of our prior predictions, which results in state of the art single image 3D reconstruction. Finally, we introduce a new benchmark, which features diverse and extreme expressions and allows, for the first time, to simultaneously evaluate neutral and posed geometry.

Acknowledgements

This work was funded by Synthesia and supported by the ERC Consolidator Grant Gen3D (101171131), the German Research Foundation (DFG) Research Unit “Learning and Simulation in Visual Computing”. Additionally, we would like to thank Angela Dai for the video voice-over.

References

- [1] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019. [2](#)
- [2] 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*, pages 187–194, 1999. [2](#)
- [3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. [2](#)
- [4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017. [2, 6](#)
- [5] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013. [2, 6](#)
- [6] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. [2, 3, 5](#)
- [7] Timothy F. Cootes, Gareth J. Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. [2](#)
- [8] Radek Danček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. [2, 5, 6, 7](#)
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [1](#)
- [10] 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, 6, 7](#)
- [11] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786. IEEE, 2018. [2, 3, 5](#)
- [12] Panagiotis P. Filntsis, George Retsinas, Foivos Paraperas-Papantonio, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos, 2022. [2](#)
- [13] 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](#)
- [14] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. [2](#)
- [15] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. *arXiv preprint arXiv:2006.10962*, 2020. [2](#)

- [16] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 7
- [17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2
- [18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022. 3
- [19] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 8
- [20] 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. 1
- [21] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. *arXiv preprint arXiv:2411.19189*, 2024. 9
- [22] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 4, 6, 8
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 6
- [24] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2, 3, 5
- [25] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5481–5492, 2024. 9
- [26] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 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. 2, 3, 4
- [27] Connor Lin, Koki Nagano, Jan Kautz, Eric Chan, Umar Iqbal, Leonidas Guibas, Gordon Wetzstein, and Sameh Khamis. Single-shot implicit morphable faces with consistent texture parameterization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 2
- [28] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 8
- [29] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shouo-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, drivable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 2, 3
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3
- [31] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009. 2
- [32] Malte Prinzler, Egor Zakharov, Vanessa Sklyarova, Berna Kabadayi, and Justus Thies. Joker: Conditional 3d head synthesis with extreme facial expressions. *arXiv preprint arXiv:2410.16395*, 2024. 9
- [33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 6
- [34] George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4
- [36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3
- [37] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: the first facial landmark localization challenge, 2013. 2
- [38] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. 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)*, pages 7763–7772, 2019. 2, 3, 5, 7
- [39] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 8
- [40] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars.

- In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20991–21002, 2023. 9
- [41] Felix Taubner, Prashant Raina, Mathieu Tuli, Eu Wern Teh, Chul Lee, and Jimmiao Huang. 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)*, pages 1227–1237, 2024. 2, 6, 7
- [42] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. *arXiv preprint arXiv:2412.12093*, 2024. 9
- [43] 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
- [44] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 9
- [45] Yue Wang and Justin M Solomon. Prnet: Self-supervised learning for partial-to-partial registration. *Advances in neural information processing systems*, 32, 2019. 7
- [46] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2
- [47] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 2, 7
- [48] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 2
- [49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 4
- [50] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. Accurate 3d face reconstruction with facial component tokens. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9033–9042, 2023. 2, 3, 6, 7
- [51] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dong-dong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 4, 5, 9
- [52] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2, 3, 5, 7
- [53] Wojciech Zienonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 2, 4, 6, 7