Learning Person-Specific Animatable Face Models from In-the-Wild Images via a Shared Base Model

Anonymous CVPR submission

Paper ID 10007

Abstract

Training a generic 3D face reconstruction model in a self-supervised manner using large-scale, in-the-wild 2D face image datasets enhances robustness to varying lighting conditions and occlusions while allowing the model to capture animatable wrinkle details across diverse facial expressions. However, a generic model often fails to adequately represent the unique characteristics of specific individuals. In this paper, we propose a method to train a generic base model and then transfer it to yield person-specific models by integrating lightweight adapters within the large-parameter ViT-MAE base model. These person-specific models excel at capturing individual facial shapes and detailed features while preserving the robustness and prior knowledge of detail variations from the base model. During training, we introduce a silhouette vertex re-projection loss to address boundary "landmark marching" issues on the 3D face caused by pose variations. Additionally, we employ an innovative teacher-student loss to leverage the inherent strengths of UNet in feature boundary localization for training our detail MAE. Quantitative and qualitative experiments demonstrate that our approach achieves state-of-the-art performance in face alignment, detail accuracy, and richness. The code will be released to the public upon the acceptance of this paper.

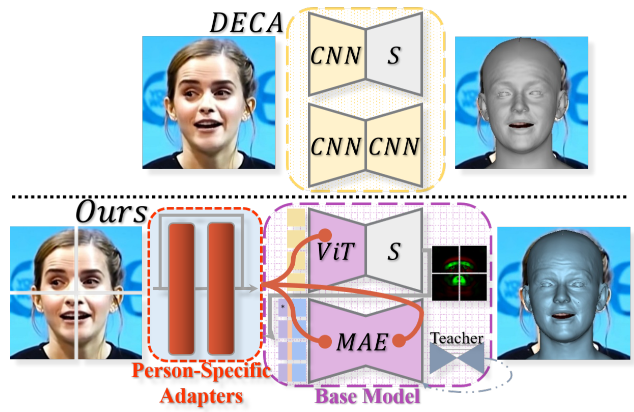

Figure 1. In contrast to previous work (e.g., DECA), we first develop a scalable, high-capacity base model (purple) and then transfer it to create person-specific models by integrating lightweight, person-specific adapters (red).

1. Introduction

The reconstruction of 3D faces from 2D images has garnered considerable attention recently [5, 8, 12, 18, 20, 42, 60, 62, 72, 76], with applications spanning diverse fields such as 3D avatar creation [1, 24, 30], face recognition [2, 3, 50], and face animation driven by speech [11, 17, 47, 48] or video [27, 47, 81]. Leveraging deep learning, the majority of recent methods [5, 6, 8, 12, 15, 18, 20, 21, 26, 28, 34, 38, 42, 49, 56–59, 62, 66–70, 72, 76, 79, 80] focus on reconstructing 3D faces from in-the-wild images, primarily employ a unified set of model weights across images of different individuals. While robust, these methods often underfit individual-specific features. However, in real-world scenarios, multiple images or a video of the same person are often available, allowing models to focus on reconstructing that specific individual more accurately—a context where current methods still have limitations.

In contrast to previous work, we first develop a scalable, high-capacity base model, trained in a self-supervised manner on extensive 2D images, and then transfer it to yield person-specific models. Our base model reconstructs feature-aligned 3D faces from in-the-wild images in real time, capturing expression-related wrinkle details (animatable features). When multiple images or a video of an individual are available, we can transfer the generic model by integrating a small set of person-specific parameters, achieving more precise real-time reconstruction (30 fps on an Nvidia GeForce RTX 3090) while maintaining robustness against occlusions and preserving priors of various facial details associated with expressions.

Vision Transformer (ViT) [16] and Masked Autoencoder (MAE) [22] are models with strong expressive capabilities,

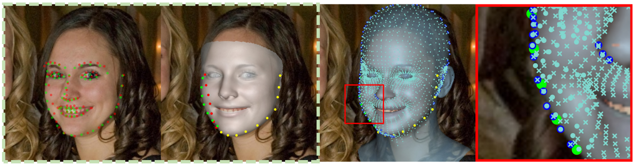

Figure 2. **Motivation of our silhouette vertex re-projection loss.** (a) 2D landmark ground truth [55] (green) and projected 3D landmarks provided by 3DDFA-v2 [20] (red). (b) The occluded boundary 3D landmarks (red) "move" to the face silhouette in the 2D annotation (green). (c) All vertices of the coarse model reconstructed by our method (light cyan, with crosses indicating normals pointing inward towards the image and dots indicating normals pointing outward), and the outer edge points of the 3D face (blue). (d) A zoom-in of the red box area from (c).

effectively gathering both local and global dependencies in visual data. They also excel in achieving robust performance and generalization on target tasks with limited training data by leveraging pre-training on large-scale datasets, aligning well with our design. Capitalizing on these advantages, We propose a ViT-MAE architecture for generic 3D face reconstruction, leveraging differentiable rendering [46] for self-supervised training. The model learns a parametric face model for coarse reconstruction, refined by a UV displacement map. This trained generic model serves as a base model, which can be further transferred to a person-specific model by integrating lightweight adapter modules [23] within the transformer layer. Our approach enables the single-image face reconstruction model to process multiple images or videos, fully leveraging the data to refine the reconstruction outcomes.

When a 3D face is projected onto an image, inner and boundary landmarks outlining facial features and the cheek are also projected. Face reconstruction methods [26, 56, 57, 66–68, 79] constrain the reconstructed 3D face by minimizing the error between the projected 3D landmarks and the annotated ground truth. However, in non-frontal views, some landmarks, especially occluded boundary ones, become invisible (red points in Fig. 2.(b)), making accurate annotation difficult. In 2D landmark annotation, boundary landmarks for the 3D cheek "shift" to align with the face silhouette, causing misalignment—known as "landmark marching" [82]. Our innovative silhouette vertex re-projection loss addresses this by aligning 2D silhouette landmarks with silhouette edge vertices on the 3D model based on current vertex normal distribution (Fig. 2.(c)&(d)). Additionally, using dense silhouette edge vertices as loss candidates enhances the model's sensitivity to normal errors, given the high variance in manual silhouette landmark annotations [55] and their tendency to spread along the boundary tangent [25]. Experimental results show that our method surpasses prior approaches in face alignment.

We define facial details as functions of identity and "tension" [44] within facial geometry, which vary with expressions. The local correspondence in UV space between the detailed displacement map and the unwrapped image texture makes a UNet [51] well-suited for learning it, with skip connections aiding in the precise localization of feature boundaries. However, the skip connections indiscriminately convey all facial details to the displacement map, hindering the modeling of the detail changes due to facial deformation, and also making it challenging to fill in the "holes" caused by occlusion. We employ a masked autoencoder (MAE) [22] for detail generation, using a consistency loss [18] to decouple identity and "tension"-related details. During training, a UNet acts as a "teacher" to guide the MAE to learn detailed feature boundaries accurately. Integrating the strengths of both UNet and MAE in detail recovery and animation, our model outperforms previous works in wrinkle diversity and accuracy without relying on 3D data. Our ablation study underscores the effectiveness.

In summary, our main contributions are:

Personalized Face Model. To the best of our knowledge, we are the first to construct a person-specific 3D face reconstruction model by transferring a large-scale generic model. The person-specific adapters within the generic ViT-MAE enhance the reconstruction from that person's images on both coarse and fine scales.

Silhouette Vertex Re-projection Loss. We introduce an innovative loss function that aligns the annotated 2D silhouette landmarks with the 3D facial boundary edges, providing more robust facial contour constraints.

Animatable Details and Teacher-Student Architecture. We innovatively employ a teacher-student architecture that leverages UNet's inherent strength in feature boundary localization alongside MAE's animation and robustness capabilities, enabling the MAE, which is ultimately used for reconstruction and animation, to capture both identity-related and tension-related details effectively.

2. Related Work

Monocular Coarse Reconstruction. Landmark loss is widely used in self-supervised approaches [26, 37, 56, 57, 66–68, 79] to ensure the alignment of projected 3D landmarks with 2D landmarks detected by face alignment techniques. However, the "landmark marching" phenomenon remains a persistent challenge. [56, 68] use dynamic boundary landmarks determined by head pose to maintain correct correspondence but overlook the influences of facial shape and expression. [26] defines silhouette landmarks on horizontal mesh lines, requiring manual redefinition for new topologies. [37] renders a projection area to identify occluded boundary landmarks, though it is time-intensive.

Animatable Detail Reconstruction. Adding facial details greatly improves model authenticity and expressiveness. While many studies have achieved high-quality detail

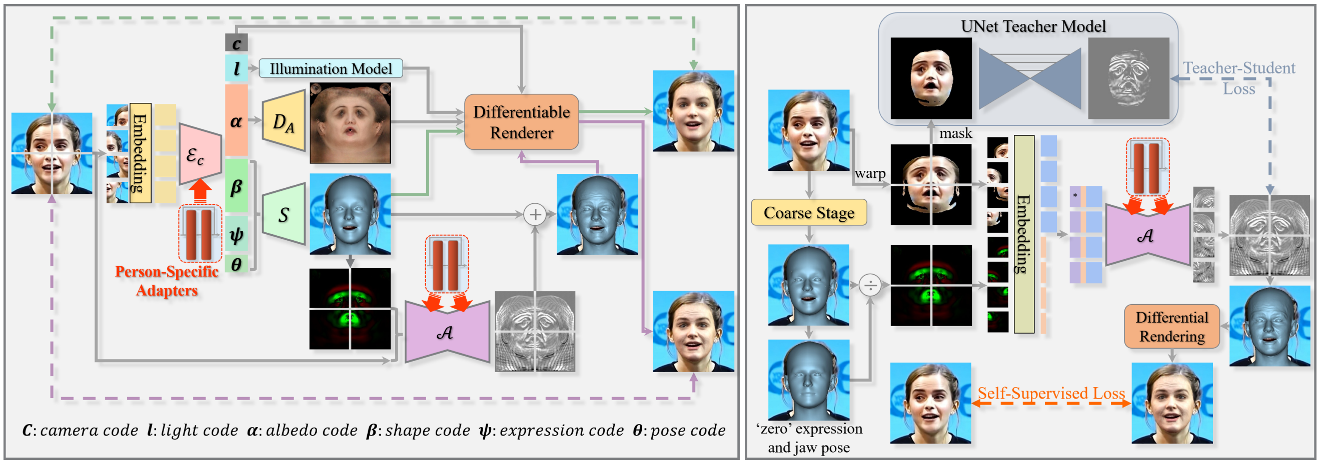

Figure 3. **Illustration of our overall architecture. Left box:** End-to-end learning framework of our coarse and detail stage. Given an image, we first regress the FLAME parameters to obtain a coarse facial shape (following the green arrows) with the help of our proposed silhouette vertex re-projection loss. We then use an MAE to reconstruct animatable facial details from the warped image texture and current facial tension [44] (following the purple arrows). During person-specific transfer, we integrate lightweight, trainable adapters (red) while freezing the base model weights. **Right box:** The training pipeline for animatable details. We first train a UNet teacher model that can effectively recover visible details from the current perspective. Under the guidance of the UNet, we train our animatable detail MAE $\mathcal{A}$ using a teacher-student paradigm.

generation [8, 21, 26, 34, 38, 49, 58, 59, 69, 72, 80], animatable details remain underexplored. Such details are essential for lifelike avatars that respond naturally to expressions. Methods like [31, 78] generate impressive animatable details from textured meshes or neutral faces but struggle to differentiate static from dynamic features. While [14] and [32] generate animatable details, neither method supports 3D face reconstruction from 2D images. [36] uses StyleGAN2 to animate details from facial images or 3D meshes but lacks robustness in varied lighting. Closest to our approach, [5] and [18] reconstruct 3D faces with animatable details from single images; however, [5] is limited by its dependency on synthetic data for supervised training, and [18] achieves reasonable results but faces challenges in detail accuracy and the realism of expression-related variations.

Multiple Images or Monocular Video. When multiple images or videos of a subject are available, we aim to fully utilize this data. Optimization-based methods [41, 53, 54, 65] often face limitations such as slow inference, difficulty adapting to unseen facial areas, and insufficient geometric detail. Learning-based methods [15, 21, 38, 45, 60, 77] leverage deep networks to integrate different viewpoints but struggle to capture intricate facial details. While [21] and [38] can reconstruct high-detail textured geometry from single videos, they fail to model unique expression-specific features or wrinkles absent in the video. Critically, none of these methods produce animatable facial details.

3. Method

Our base model and transferred person-specific models share a two-stage self-supervised training framework. This section details the coarse and detailed stages, followed by the transfer process to obtain a person-specific model.

3.1. Preliminary

3D Geometry Model. FLAME [33] is a statistical 3D head model that, given identity shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$, facial expression parameters $\boldsymbol{\psi} \in \mathbb{R}^{|\psi|}$, and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{3k+3}$ for rotations around $k = 4$ joints (neck, jaw, and eyeballs) and global rotation, outputs a mesh $S(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ with $n_v = 5023$ vertices and $n_f = 9976$ faces.

Appearance Model. We use a texture statistical model that aligns the Basel Face Model's albedo space [40] with FLAME's UV layout [18], which outputs a FLAME texture map $A(\boldsymbol{\alpha}) \in \mathbb{R}^{d \times d \times 3}$, where $d = 256$, given texture parameters $\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{\alpha}|}$.

Camera Model. The camera model is parameterized by $\mathbf{c} = (s, \mathbf{t})$. An orthographic projection transformation is used to project 3D mesh vertices into the image space, formulated as $\mathbf{u} = s\Pi(\mathbf{v}) + \mathbf{t}$, where $\mathbf{v}$ is a vertex in the 3D mesh, $\Pi \in \mathbb{R}^{2\times 3}$ is the orthographic projection matrix, $s \in \mathbb{R}$ is the isotropic scale, and $\mathbf{t} \in \mathbb{R}^2$ is the 2D translation.

Illumination Model. The spherical harmonics illumination model [43] is adopted to estimate the illumination conditions in the input image. The shaded face texture is computed as $U = U(A, \mathbf{l}, N)$ with $U_{i,j} = A_{i,j} \odot \mathcal{H}_{i,j} : = A_{i,j} \odot \sum_{k=1}^{9} \mathbf{l}_k H_k(N_{i,j})$, where $\odot$ denotes the Hadamard product, $A_{i,j}$, $N_{i,j}$, and $U_{i,j}$ represent the albedo, surface normal, and shaded texture in UV coordinates respectively, $H : \mathbb{R}^3 \rightarrow \mathbb{R}$ are the Spherical Harmonics (SH) basis functions, and $\mathbf{l} \in \mathbb{R}^9$ is the SH coefficient.

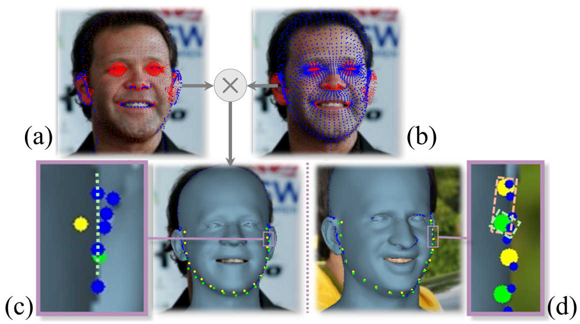

Figure 4. **Our silhouette vertex re-projection loss construction method and its superiority.** (a) Edge points (blue) selected from all model vertices (red) based on vertex normals. (b) Candidate region (blue) for boundary edge vertices excludes the nose and the ear regions. (c) & (d) Two typical scenarios.

3.2. Generic Coarse Reconstruction

We start by employing self-supervised learning to develop a base model that robustly reconstructs any in-the-wild image with fine details. Our base model first learns a coarse reconstruction. A vision transformer [16] $\mathcal{E}_c$ serves as the coarse encoder. It splits a 2D face image I into 16×16 patches, embeds each patch linearly and adds positional embeddings to form a sequence of tokens. A learnable "classification token" ($\mathbf{z}_0^0 = \mathbf{x}_{\texttt{class}}$) is prepended to this sequence, which is subsequently passed through L transformer blocks. The output classification token $\mathbf{z}_L^0$ is passed through an MLP with one hidden layer to generate a latent code comprising FLAME parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\psi}$, along with albedo parameters $\boldsymbol{\alpha}$, camera parameters $\mathbf{c}$, and lighting parameters $\mathbf{l}$. With FLAME and albedo parameters, we generate a textured 3D mesh, which, combined with camera and lighting parameters, enables differentiable rendering to produce the reconstructed facial image I_r (Fig. 3, left, green arrows).

Silhouette Vertex Re-Projection Loss. The first 17 landmarks $\mathbf{k}_i \in \mathbb{R}^2, i \in 1, \ldots, 17$ from the annotated ground truth are 2D silhouette landmarks that should align with the outer edge of the 3D face from the current view. Our silhouette vertex re-projection loss naturally enforces this alignment. Compared to FLAME's dynamic landmark marching approach [33], which may misalign landmarks under variations in facial shape and expression, our method, illustrated in Fig. 4.(c)&(d), achieves more accurate matching in two typical scenarios: (1) when the dynamic boundary landmark (yellow) provided by the FLAME model does not accurately locate on the model's edge (green dashed line); and (2) when the FLAME algorithm necessitates matching the ground truth landmark (green) with the yellow point as indicated by the orange box, whereas our method matches it to the blue edge point shown in the green box, providing a more reasonable alignment.

We define the 'zero-pose' boundary landmarks of the FLAME model $\widetilde{\mathbf{k}}_i$ as those selected by FLAME when given a zero-pose input. If a landmark $\widetilde{\mathbf{k}}_i$ is non-occluded, it is constrained using the vanilla landmark re-projection; if occluded, we match it to the nearest edge point in V_e and calculate the $\mathbb{L}1$ loss. Occlusion is determined by vertex normals: if the z-direction points outward, it is non-occluded; otherwise, it is occluded. The determination of edge point set V_e is as follows: if the normals of two vertices at the ends of an edge have opposite signs in the z-direction, it is considered that both of these vertices are located on the edge of the 3D model, we then project them onto the image plane (Fig. 4.(a)). Furthermore, as shown in Fig. 4.(b), we exclude the edges of the nose and ear regions using pre-segmented vertex labels $\mathbf{Z} \in \{0,1\}^{5023}$. We refer to the cleaned edge points as the 3D "silhouette vertices".

The silhouette vertex re-projection loss is defined as follows:

$$\mathcal{L}_{sil} = \sum_{i, N_{z,i} > 0} d(\mathbf{k}_i, \widetilde{\mathbf{k}}_i) + \sum_{i, N_{z,i} < 0} \min_{\mathbf{v}_e \in \widetilde{V_e}} d(\mathbf{k}_i, \mathbf{v}_e), \quad (1)$$

where $N_{z,i} = (N_{\widetilde{\mathbf{k}}_i})_z$, $\widetilde{\mathbf{k}}_i, i \in 1, \cdots, 17$ represents the 3D landmark coordinates that match the 2D landmark $\mathbf{k}_i$ in the zero-pose, N represents the vertex normal, $(\bullet)_z$ denotes the z-component. $\widetilde{V_e} = V_e \odot \widetilde{Z} \in \mathbb{R}^{5023 \times 2}$, $\widetilde{Z} \in \{0,1\}^{5023 \times 2}$. If $Z_i = 0$, then $\widetilde{\mathbf{Z}}_i = (0,0)^T$; otherwise, $\widetilde{\mathbf{Z}}_i = (1,1)^T$.

Overall Losses for Coarse Reconstruction. The base model learns coarse reconstruction using the total loss $\mathcal{L}_C$:

$$\mathcal{L}_C = \mathcal{L}_{sil} + \mathcal{L}_{inL} + \mathcal{L}_{spL} + \mathcal{L}_{pho} + \mathcal{L}_{per} + \mathcal{L}_{reg} + \mathcal{L}_{sc}, \quad (2)$$

where $\mathcal{L}_{inL}$ is the landmark loss for static inner landmarks, $\mathcal{L}_{spL}$ is the special landmark pairs loss that is calculated on a set of landmark pairs (e.g., upper/lower eyelid or lip landmarks) to constrain features like eye and mouth opening in a translation-invariant manner. $\mathcal{L}_{pho}$ is the photometric loss commonly used in self-supervised methods [15, 18]. $\mathcal{L}_{per} = \mathcal{L}_{id} + \mathcal{L}_{emo} + \mathcal{L}_{lr}$ combines three perceptual losses to ensure high-level identity [18] and emotion consistency [12], as well as accurate lip movements [19]. $\mathcal{L}_{reg}$ includes regularization losses for $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and $\boldsymbol{\alpha}$. Finally, $\mathcal{L}_{sc}$ is the shape consistency loss [18]

3.3. Generic Detail Reconstruction

Our base model subsequently learns a displacement map $D \in [-0.01, 0.01]^{d \times d}$ to refine the FLAME geometry. We use an MAE [22] to capture animatable facial details (Fig. 3, right), defined as $D = \mathcal{A}(I, T_{UV}, \mathbf{x}_{\mathtt{d}})$, where I is the input image, T_{UV} is the current face tension map, and $\mathbf{x}_{\mathtt{d}}$ is a learnable, patch-wise detail latent that is shared across all individuals. Unlike the vanilla MAE, we do not apply masking (i.e., use a zero masking ratio) because our task differs significantly from standard MAE applications, requiring the autoencoder to leverage all available information to accurately recover facial details aligned with the input images.

Facial Tension. Facial tension quantifies the vertex-wise compression or expansion on the 3D mesh caused by deformation from a neutral expression with a closed mouth to the current expression [5, 44]. We propose a new method for calculating the tension at vertex v_i:

$$t_{v_i}(S, S') = \frac{1}{E_i} \sum_{E_i}^{j=1} \frac{e^{k\|e'_j\|} - e^{k\|e_j\|}}{e^{k\|e'_j\|} + e^{k\|e_j\|}}, \quad (3)$$

where $e_1, \cdots, e_{E_i}$ are the E_i edges connected to v_i in $S = S(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$, and $e'_1, \cdots, e'_{E_i}$ are the E_i edges connected to v'_i in $S' = S(\boldsymbol{\beta}, \mathbf{0}, \mathbf{0})$. Here, $\|\bullet\|$ denotes the edge length, and k is a fixed scaling factor. Compression in the vertex neighborhood results in positive tension, while stretching yields negative tension. Our tension metric $t_{v_i}(S, S')$ satisfies: 1) Antisymmetry: $t_{v_i}(S', S) = -t_{v_i}(S, S')$; and 2) Boundedness: $\forall S, S', -1 < t_{v_i}(S, S') < 1$, making this tension calculation more suitable as input to a neural network. The tension of S can be represented as $\{t_{v_i}(S, S')\}$. Using the mapping relationship of UV coordinates from the FLAME mesh S, we derive the tension map $T_{UV}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ in UV space.

Teacher-Student Strategy. Unwrapping the input image to UV space using the reconstructed coarse FLAME geometry creates a local correspondence between the image texture and the UV space displacement map. Thus, a UNet [51] is more suitable for estimating the displacement map from the UV image texture than an encoder-decoder or autoencoder, as its skip connections efficiently utilize local input information. However, the skip connections make it difficult to animate facial details according to facial deformations, and non-frontal poses or occlusions may cause incompleteness in the unwrapped texture, leading to missing details in the reconstruction. Autoencoder structures, on the other hand, have unique advantages in terms of animation and robustness to occlusion. Therefore, we train a shallow UNet detail reconstruction network as a teacher model, using its estimates as pseudo ground truth to guide our autoencoder in detail reconstruction. This teacher model offers more direct supervision compared to shape-from-shading photometric loss, as 3D facial details and 2D shading on rendered images do not have a one-to-one mapping—multiple ways exist to add details that result in the same shadows in the rendered output.

UNet Teacher Training. We train the UNet teacher network $\mathcal{D}_{UNet}$ for detail estimation by minimizing:

$$\mathcal{L}_{UNet} = \mathcal{L}_{phoD} + \mathcal{L}_{mrf} + \mathcal{L}_{smo} + \mathcal{L}_{regD}. \quad (4)$$

where $\mathcal{L}_{pho}$ is the photometric loss for detail rendering [18]. $\mathcal{L}_{mrf}$ is an ID-MRF loss [74], computed on the $conv3\_2$ and $conv4\_2$ layers of VGG19 [63], encouraging the model to capture high-frequency geometric details. $\mathcal{L}_{smo}$ is a smoothness loss prevents overly sharp or high-frequency artifacts in the reconstructed details. $\mathcal{L}_{regD}$ is the detail regularization loss regularizes the estimated displacements to reduce noise and artifacts. For more details, please refer to the supplementary materials.

Teacher-Student Loss. With the UNet teacher network trained, we can employ the teacher-student loss to aid the training of our detail MAE $\mathcal{A}$. The teacher-student loss measures the similarity between the displacement given by the pretrained teacher network UNet and the student network $\mathcal{A}$. The photometric loss and SSIM [75] loss are used:

$$\mathcal{L}_{Tchr} = \mathcal{L}_{phoD}(D, D_{UNet}) + \mathcal{L}_{ssim}(D, D_{UNet}). \quad (5)$$

Overall Losses for Animatable Details. We train $\mathcal{A}$ on video datasets [35, 71], where each mini-batch contains images from different frames of the same video. This allows the model to learn animatable facial details that vary with changes in facial tension. In total, we optimize:

$$\mathcal{L}_{animD} = \mathcal{L}_{UNet} + \mathcal{L}_{Tchr} + \mathcal{L}_{sym} + \mathcal{L}_{dc}, \quad (6)$$

where $\mathcal{L}_{UNet} = \mathcal{L}_{phoD} + \mathcal{L}_{mrf} + \mathcal{L}_{smo} + \mathcal{L}_{regD}$ includes the losses used for training the UNet teacher model, which are also employed during the training of the student model $\mathcal{A}$. $\mathcal{L}_{sym}$ is the soft symmetry loss employed to enhance the model's robustness in occlusion regions and reduce boundary artifacts. $\mathcal{L}_{dc}$ is the detail consistency loss:

$$\mathcal{L}_{dc} = \mathcal{L}_{animD}(I, \mathcal{D}_{UNet}(I), \mathcal{D}(I', T_{UV})), \quad (7)$$

where, ensures that for images of the same individual, swapping the input images makes no difference, as they should convey the same identity information. Details from I' and T_{UV} should be consistent with image I and the pseudo ground truth $\mathcal{D}_{UNet}(I)$ from the teacher UNet.

3.4. Person-Specific Transfer

Given multiple images or videos of a person, we can transfer the base model to yield a person-specific model. The geometry of the person-specific model better aligns with the shapes and boundaries of the inputs, capturing the unique facial details of the individual. The person-specific model retains the base model's robustness to pose and occlusion, as well as valuable priors regarding dynamic details.

We achieve this transfer by incorporating lightweight modules, δ_{PS}, known as "adapters" [23], between layers of the base model (Fig. 3). The base model parameters remain fixed, and only the adapter parameters are trained. Due to the residual structure of the adapters, the modifications to the base model are incremental. Positioned within the transformer blocks, adapters δ_{PS} are added after the feed-forward layer, preceding layer normalization. Additionally, the learnable patch-wise detail latent $\mathbf{x_d}$ and the layer normalization parameters are also trained. More details are provided in the supplementary materials.

Table 1. Reconstruction error across different datasets.

Method	300-W [55]			300-VW [61]			FaceScape [78]	
	boundary↓	inner↓	overall↓	boundary↓	inner↓	overall↓	inner error↓	overall error↓
DECA [18]	0.0576	0.0402	0.0446	0.0563	0.0426	0.0460	0.0477 (±0.0097)	0.0496 (±0.0103)
EMOCA [12]	0.0579	0.0467	0.0495	-	-	-	-	-
EMOCA-v2 [12, 19]	0.0590	0.0377	0.0430	0.0553	0.0507	0.0519	0.0585 (±0.0111)	0.0600 (±0.0109)
SynergyNet [76]	-	0.0545	0.0545	-	0.0651	0.0651	0.0589 (±0.0117)	0.0785 (±0.0230)
3DDFA-v2 [20]	-	0.0470	0.0470	-	0.0524	0.0524	0.0514 (±0.0104)	0.0700 (±0.0218)
Ours-base (w/o $\mathcal{L}_{sil}$)†	0.0616	0.0401	0.0455	0.0660	0.0554	0.0580	-	-
Ours-base	**0.0559**	**0.0348**	**0.0401**	0.0585	0.0456	0.0488	**0.0404 (±0.0126)**	**0.0429 (±0.0128)**
Ours-base w/ δ_{PS}	-	-	-	**0.0359**	**0.0217**	**0.0252**	**0.0253 (±0.0045)**	**0.0305 (±0.0060)**

* A new version of EMOCA that incorporates perceptual lip reading loss [19] and produces better lip and eye alignment compared to the original model [12].

† Our base model trained with DECA's landmark re-projection loss [18], without the proposed silhouette vertex re-projection loss (Eqn. 1).

Person-specific transfer is also divided into a coarse stage and a detailed stage. Given multiple images or video frames of an individual, the model is trained using the loss function $\mathcal{L}_C$ (Eqn. 2) in the coarse stage and teacher supervision loss $\mathcal{L}_{Tchr}$ (Eqn. 5) in the detail stage. On an NVIDIA GeForce RTX 3090, several minutes of training yields significantly improved reconstruction results. Specially, since the occluded areas are not optimized by the losses during transfer, they retain the priors from the base model, allowing for reasonable generation of detailed facial features based on the extensive face details learned by the base model. Moreover, due to the strong identity consistency of the coarse shape of the person-specific model across different frames, variations in shape between frames are more attributed to expression changes. This facilitates better decoupling of identity and expression parameters in the facial statistical model, leading to more precise expression estimation and dynamic, expression-related details specific to the individual.

4. Experiments

4.1. Implementation Details

Datasets. We train our model using the publicly available datasets BUPT-Balancedface [73], Celeb-DF (v2) [35], and MEAD [71]. For each image, facial landmarks are automatically annotated using HRNet [64], and unreliable images are filtered based on the estimated per-landmark heatmaps. We utilize a facial skin region segmentation method following [7] to obtain a mask of the facial skin area.

Implementation Details. We implement our model in PyTorch [39], using the differentiable rasterizer from Pytorch3D [46] for rendering. We employ the Adam [29] optimizer with a learning rate of $1e$-4 for the base model and $1e$-3 for the person-specific transfer. Input images are cropped and aligned with RetinaFace [13], and resized to 256×256. Additional details on data augmentation, hyperparameter settings (e.g., loss balancing weights), and explanations of the losses are in the supplementary materials.

4.2. Quantitative Comparison

We compare the accuracy of our models in face alignment with publicly available facial reconstruction methods, namely 3DDFA-v2 [20], SynergyNet [76], DECA [18] and EMOCA [12]. To comprehensively demonstrate the superiority of our method in coarse shape, encompassing facial contours and features, we conduct evaluations across monocular image reconstruction (300-W dataset [55]), monocular video reconstruction (300-VW dataset [61]), and multi-view image reconstruction (FaceScape dataset [78]). Note that we do not evaluate our method on 3D benchmarks, as mainstream self-supervised face reconstruction approaches typically assume an orthographic camera model, whereas 3D dataset photos are often taken from close distances and exhibit significant perspective distortion. Thus, directly comparing orthographic projection-based reconstructions with ground truth would be unfair.

300-W Dataset. We employed the 300-W dataset to assess the precision of our base model in single-image face alignment. As shown in Tab.1 and Fig. 5, on 1424 cleaned test images, our method achieves a lower RMSE error [55] than previous works for both boundary and inner landmarks. This is attributed to our novel silhouette vertex re-projection loss, which establishes more precise correspondences for the ground-truth 2D silhouette landmarks while naturally mitigating the relatively large variance in manual landmark annotations along the silhouette tangent.

300-VW Dataset. The 300-VW dataset provides a comprehensive benchmark for landmark tracking in long-term 'in-the-wild' facial videos. Due to the semi-automatic annotation utilized in the 300-VW Challenge [9], discrepancies exist between the annotated landmarks and their actual facial positions. Consequently, testing on the 300-VW dataset serves primarily as a reference for evaluating face reconstruction accuracy in videos with continuously changing poses. Methods exhibiting similar test errors should be considered comparable. As depicted in Tab.1, our base model delivers results on par with DECA, EMOCA-v2, and 3DDFA-v2 [20], while significantly surpassing SynergyNet [76]. Additionally, our person-specific models substan-

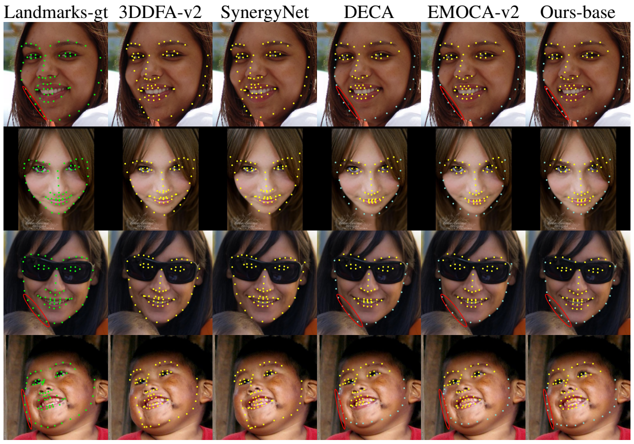

Figure 5. **Face alignment on 300-W [55].** From left to right: Ground truth 2D landmarks, projected 3D landmarks estimated by 3DDFA-v2 [20] and SynergyNet [76], and '2D landmarks' provided by DECA [18], EMOCA-v2 [12], and our base model.

tially outperform both our base model and previous works.

FaceScape Dataset. FaceScape [78] is a large-scale detailed 3D face dataset with multi-view images, camera parameters, 3D face scans, and parametric models with their registration parameters. To address the significant perspective distortion in the images, we used FaceScape's parametric models to extract 3D facial landmarks and projected them onto the image plane using the camera parameters, establishing them as the ground truth for image facial landmarks. As only 43 landmarks from FaceScape's parametric model are applicable (numbered $17 \sim 59$ in the 68-landmark annotation), we completed the set (17 facial boundary and 8 inner mouth circle landmarks) with the annotation from HRNet [64]. As shown in Tab.1, our base model outperforms previous works in both inner and overall landmark accuracy. The transferred person-specific models further reduce the error across different views for each identity significantly.

4.3. Qualitative Comparison

Given an in-the-wild image, our base model reconstructs a 3D face with animatable details. Given multiple images or a video of an individual, we transfer the base model by inserting trainable person-specific adapters. Our person-specific model achieves higher fidelity reconstruction of images from that individual. We conduct visualized comparisons with previous work in terms of self-supervised coarse shape reconstruction [12, 18, 20, 76], detail reconstruction [12, 18, 72, 78], and detail animation [12, 18]. The input images are taken from the FaceForensics++ dataset [52], where the images and identities were never encountered during the training of the base model.

Coarse Shape Reconstruction. Fig. 6 qualitatively compares the results of our base and person-specific models with state-of-the-art coarse reconstruction methods [12, 18, 20, 76] that are publicly available. Compared to these methods, our base model exhibits higher accuracy in fitting the outer contour, pose, and facial feature representation. Our person-specific model further enhances these advantages.

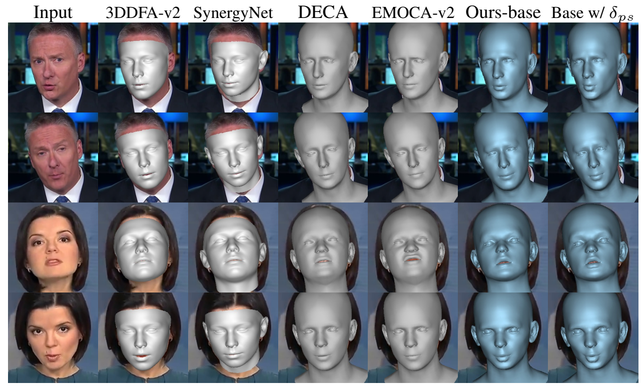

Figure 6. **Comparison on coarse shape reconstruction.** From left to right: Input image, 3DDFA-v2 [20], SynergyNet [76], DECA [18], EMOCA-v2 [12], our base model, and our transferred person-specific models.

Detailed Reconstruction. Fig. 7 visually compares our work to existing detailed reconstruction methods [12, 18, 72, 78]. Several methods [72, 78] optimize for the current image, which limits inference speed and robustness to pose and occlusion. Previous generic models that reconstruct animatable geometric details [12, 18] struggle with the fidelity of person-specific facial details, as seen in Fig. 7. Compared to the base model (penultimate column), the transferred person-specific models (last column) exhibit improved accuracy in wrinkle details.

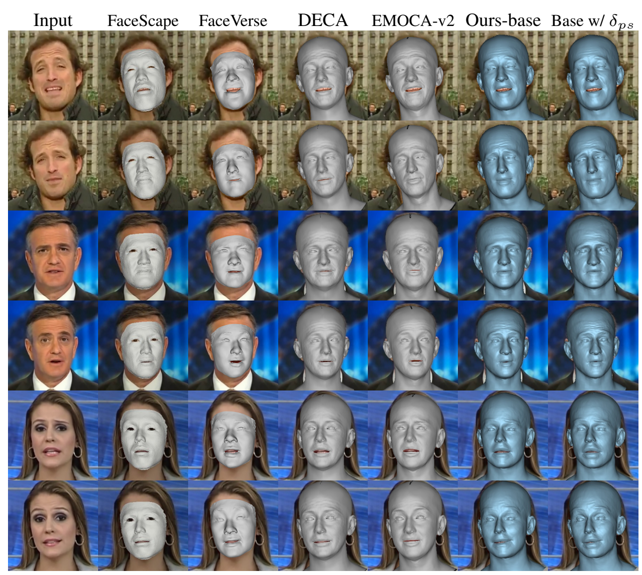

Figure 7. **Comparison on detail shape reconstruction.** From left to right: Input image, FaceScape [78], FaceVerse [72], DECA [18], EMOCA-v2 [12], our base model, and our person-specific models.

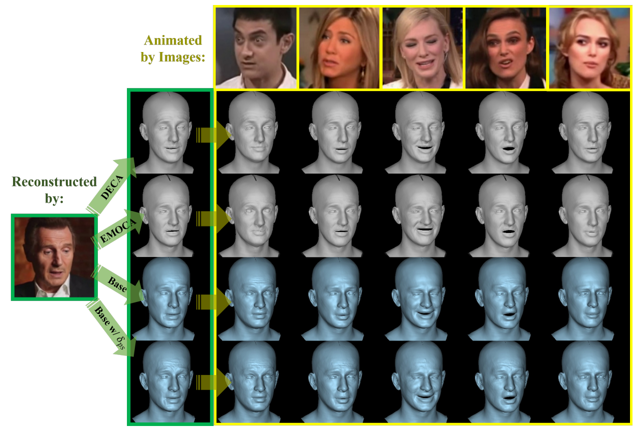

Figure 8. **Comparison on face animation.** Given a source image, DECA [18] (row 2), EMOCA-v2 [12] (row 3), and our base (row 4) and person-specific (row 5) models can respectively generate detailed 3D faces (green boxes). With a driving image (yellow boxes), these models can drive the face to exhibit corresponding expressions.

Detailed Face Animation. Fig. 8 demonstrates the animation quality of our models compared to the state-of-the-art detail animation models [12, 18]. Our base model has learned a rich prior of expression-related details, surpassing existing works in realism and accuracy. Meanwhile, our transferred model captures person-specific details more finely while inheriting the animatability and robustness to in-the-wild driving images from the base model. This enhances the model's accuracy in reconstructing the specific individual, with more detailed and enriched features.

4.4. Ablation Studies

Silhouette Vertex Re-Projection Loss. We trained a network, Ours-base (w/o $\mathcal{L}_{sil}$), without the proposed silhouette vertex re-projection loss $\mathcal{L}_{sil}$, using FLAME's landmark marching algorithm [33] to apply the landmark re-projection loss across all 68 landmarks, as done by DECA [18]. The penultimate row in Table 1 shows the evaluation results of Ours-base (w/o $\mathcal{L}_{sil}$) on the 300-W [55] and 300-VW [61] dataset. Ours-base (w/o $\mathcal{L}_{sil}$) performs slightly worse than DECA on the 300-W dataset in terms of overall error, which might be attributed to the different choice of training data (DECA uses VGGFace2 [4] and VoxCeleb2 [10]). In contrast, using $\mathcal{L}_{sil}$ improves edge and interior landmark errors by 9.1% and 13.2%, respectively, on 300-W. Fig. 9 visually shows the contribution of $\mathcal{L}_{sil}$, in terms of boundary fitting accuracy.

Teacher-Student Loss. We present an ablation study on the proposed teacher-student strategy for training the detail network. Fig. 9 demonstrates the contribution of the teacher-student strategy to the facial detail reconstruction. The network trained without the teacher supervision loss $\mathcal{L}_{Tchr}$ (Equation 5) (with other settings unchanged) generates facial details with numerous unrealistic artifacts. This occurs because the shape and the rendered RGB image do not have a one-to-one correspondence, resulting in fewer constraints on the optimization direction when training the network using shape-from-shading loss, making it challenging to achieve acceptable results.

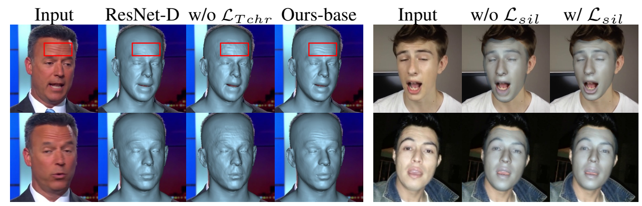

Figure 9. **Ablation studies.** Left: Compared to MAE (Ours-base), using convolutional (ResNet) and deconvolutional networks (ResNet-D) struggles to capture expression-dependent details. When training MAE without incorporating the teacher supervision loss $\mathcal{L}_{Tchr}$ (Eqn. 5), it results in inaccurate wrinkles and artifacts. Right: Without $\mathcal{L}_{sil}$, the facial boundary does not properly align with the input image.

Network Architecture. We compared the effectiveness of using convolutional (ResNet) and deconvolutional networks versus a MAE for detail reconstruction, both employing the teacher-student loss. For the former, we adopted the same network architecture and dynamic detail driving method as DECA [18]. Fig. 9 demonstrates that the MAE captures expression-related details more effectively. This is due to the superior long-range dependency capture and feature extraction capabilities of the MAE architecture we employed. Additionally, the transformer structure allows us to insert adapter layers [23], enabling an incremental person-specific transfer to retain the generalization capabilities of the base model on face animation and occlusion.

5. Conclusion

We propose constructing person-specific 3D face reconstruction models by integrating lightweight adapters into a large-scale ViT-MAE base model. During the coarse reconstruction stage, a novel silhouette vertex re-projection loss is introduced to address the issue of "landmark marching", thereby correcting the misalignment of boundary facial landmarks and achieving state-of-the-art performance. In the detailed stage, a teacher-student loss is employed to resolve the ambiguities inherent in the self-supervised shape-from-shading approach, allowing the detail MAE to effectively capture rich and accurate features. When provided with multiple images or videos of an individual, we can further transfer the base model to a person-specific model, improving reconstruction accuracy and enabling more effective decoupling of identity and expression details. Our advantages in facial boundary and detail alignment, combined with the ability to animate details through facial movements, make our approach highly suitable for face animation, wrinkle transfer, and downstream applications such as face reenactment and virtual avatar creation.

References

[1] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics (TOG)*, 41(1):3:1–3:21, 2021. 1

[2] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3d morphable model. In *Proceedings of fifth IEEE international conference on automatic face gesture recognition*, pages 202–207. IEEE, 2002. 1

[3] Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, and Chen Change Loy. Pose-robust face recognition via deep residual equivariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2018. 1

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 8

[5] Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9087–9098, 2023. 1, 3, 5

[6] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129. IEEE, 2018. 1

[7] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing (VCIP)*, pages 1–4. IEEE, 2017. 6

[8] Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020. 1, 3

[9] Grigoris G Chrysos, Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape. Offline deformable face tracking in arbitrary videos. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–9, 2015. 6

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 8

[11] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 1

[12] Radek Daněček, 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. 1, 4, 6, 7, 8

[13] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 6

[14] Qixin Deng, Luming Ma, Aobo Jin, Huikun Bi, Binh Huy Le, and Zhigang Deng. Plausible 3d face wrinkle generation using variational autoencoders. *IEEE Transactions on Visualization and Computer Graphics*, 28(9):3113–3125, 2021. 3

[15] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 285–295, 2019. 1, 3, 4

[16] 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, 4

[17] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 1

[18] 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): 88:1–88:13, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[19] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022. 4, 6

[20] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision*, pages 152–168. Springer, 2020. 1, 2, 6, 7

[21] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1294–1307, 2018. 1, 3

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 4

[23] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Confer-

ence on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 5, 8

[24] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017. 1

[25] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2021. 2

[26] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27 (10):4756–4770, 2018. 1, 2, 3

[27] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM transactions on graphics (TOG)*, 37(4): 1–14, 2018. 1

[28] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4625–4634, 2018. 1

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2014. 6

[30] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 760–769, 2020. 1

[31] Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. Dynamic facial asset and rig generation from a single scan. *ACM Transactions on Graphics (TOG)*, 39(6):215:1–215:18, 2020. 3

[32] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3407–3416, 2020. 3

[33] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 3, 4, 8

[34] Yue Li, Liqian Ma, Haoqiang Fan, and Kenny Mitchell. Feature-preserving detailed 3d face reconstruction from a single image. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1:1–1:9, 2018. 1, 3

[35] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 5, 6

[36] Jingwang Ling, Zhibo Wang, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Structure-aware editable morphable model for 3d facial detail animation and manipulation. In *Proceedings of the European Conference on Computer Vision*, pages 249–267. Springer, 2022. 3

[37] Peng Liu, Yao Yu, Yu Zhou, and Sidan Du. Single view 3d face reconstruction with landmark updating. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 403–408. IEEE, 2019. 2

[38] Luming Ma and Zhigang Deng. Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 11:1–11:10, 2019. 1, 3

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[40] 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. 3

[41] Marcel Piotraschke and Volker Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3418–3427, 2016. 3

[42] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4142–4160, 2020. 1

[43] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 3

[44] Chirag Raman, Charlie Hewitt, Erroll Wood, and Tadas Baltrušaitis. Mesh-tension driven expression-based wrinkles for synthetic faces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3515–3525, 2023. 2, 3, 5

[45] Eduard Ramon, Janna Escur, and Xavier Giró-i Nieto. Multiview 3d face reconstruction in the wild using siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 3096–3100, 2019. 3

[46] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020. 2, 6

[47] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 13759–13768, 2021. 1

[48] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 1

[49] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5553–5562, 2017. 1, 3

[50] Sami Romdhani, Volker Blanz, and Thomas Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*, pages 3–19. Springer, 2002. 1

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 5

[52] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 7

[53] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3d face reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2606–2615, 2015. 3

[54] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2016. 3

[55] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 2, 6, 7, 8

[56] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 1, 2

[57] Nedko Savov, Minh Ngô, Sezer Karaoglu, Hamdi Dibeklioglu, and Theo Gevers. Pose and expression robust age estimation via 3d face reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1270–1278, 2019. 2

[58] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1585–1594, 2017. 3

[59] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces 'in the wild'. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6296–6305. IEEE, 2018. 1, 3

[60] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Proceedings of the European Conference on Computer Vision*, pages 53–70. Springer, 2020. 1, 3

[61] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015. 6, 8

[62] Zhixin Shu, Duygu Ceylan, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, and Dimitris Samaras. Learning monocular face reconstruction using multi-view supervision. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 241–248. IEEE, 2020. 1

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[64] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 6, 7

[65] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *Proceedings of the European Conference on Computer Vision*, pages 796–812. Springer, 2014. 3

[66] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 1, 2

[67] Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):357–370, 2018.

[68] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 2

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

[70] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1493–1502, 2017. 1

[71] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proceedings of the European Conference on Computer Vision*, pages 700–717. Springer, 2020. 5, 6

[72] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. 1, 3, 7

[73] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 692–702, 2019. 6

[74] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *Advances in neural information processing systems*, 31, 2018. 5

[75] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[76] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *International Conference on 3D Vision (3DV)*, pages 453–463. IEEE, 2021. 1, 6, 7

[77] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019. 3

[78] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2020. 3, 6, 7

[79] Jae Shin Yoon, Takaaki Shiratori, Shoou-I Yu, and Hyun Soo Park. Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4601–4609, 2019. 1, 2

[80] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019. 1, 3

[81] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 1

[82] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 787–796, 2015. 2