KaoLRM: Repurposing Pre-trained Large Reconstruction Models for Parametric 3D Face Reconstruction

Qingtian Zhu[1*], Xu Cao[2], Zhixiang Wang[2], Yinqiang Zheng[1], Takafumi Taketomi[2]
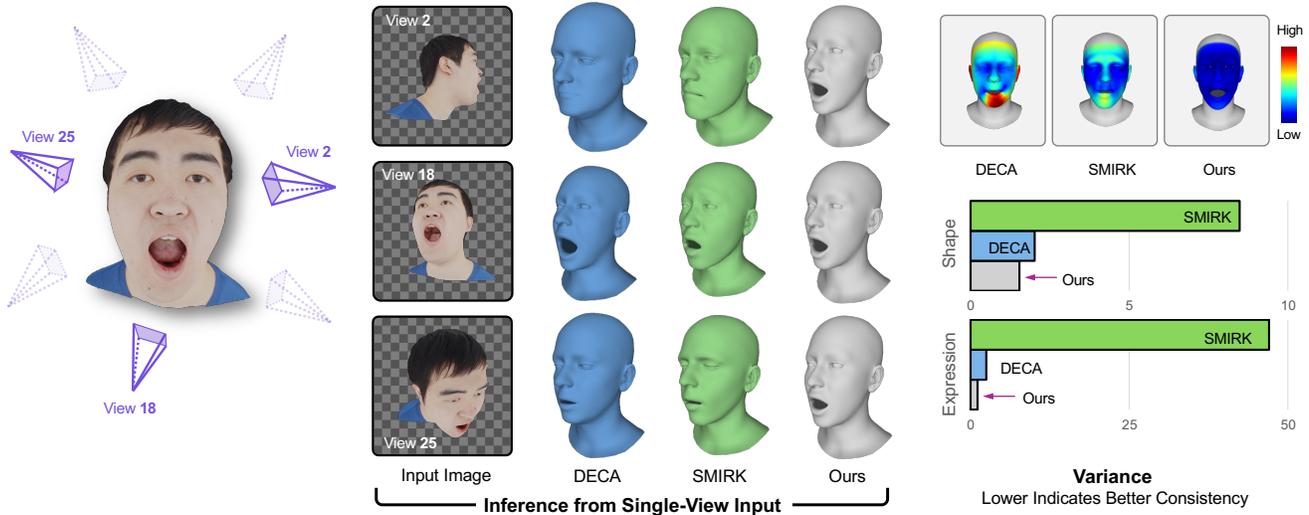
[1] The University of Tokyo   [2] CyberAgent

Figure 1. KaoLRM predicts FLAME parameters from a *single-view* facial image and yields more faithful and *consistent* predictions for the same subject captured from different viewpoints. We validate the consistency qualitatively through visual inspection of the reconstructed FLAME meshes and quantitatively by measuring the variance of the FLAME parameters across the views. The methods for reference are DECA [12] and SMIRK [38].

Abstract

We propose KaoLRM to re-target the learned prior of the Large Reconstruction Model (LRM) for parametric 3D face reconstruction from single-view images. Parametric 3D Morphable Models (3DMMs) have been widely used for facial reconstruction due to their compact and interpretable parameterization, yet existing 3DMM regressors often exhibit poor consistency across varying viewpoints. To address this, we harness the pre-trained 3D prior of LRM and incorporate FLAME-based 2D Gaussian Splatting into LRM's rendering pipeline. Specifically, KaoLRM projects LRM's pre-trained triplane features into the FLAME parameter space to recover geometry, and models appearance via 2D Gaussian primitives that are tightly coupled to the FLAME mesh. The rich prior enables the FLAME regressor to be aware of the 3D structure, leading to accu-rate and robust reconstructions under self-occlusions and diverse viewpoints. Experiments on both controlled and in-the-wild benchmarks demonstrate that KaoLRM achieves superior reconstruction accuracy and cross-view consistency, while existing methods remain sensitive to viewpoint variations. The code is released at https://github.com/CyberAgentAILab/KaoLRM.

1. Introduction

Inferring parametric 3D Morphable Models (3DMMs) from single-view facial images is a long-standing challenge in computer vision and graphics. Due to the compact and interpretable parametric representation, 3DMMs provide intuitive control over identity, expression, and pose, making them crucial for applications such as animation, virtual reality, and telepresence [47].

Training single-view 3DMM regressors with annotated parameters or explicit 3D supervision requires paired data,

---

[1]Work done during Q. Zhu's internship at CyberAgent AI Lab.

which are often expensive to obtain. To alleviate the reliance on such annotations, many existing methods [7, 12, 38] adopt self-supervised, analysis-by-synthesis training paradigms, where the reconstructed 3DMM is re-rendered from the input viewpoint and compared with the input images using a photometric loss. However, this paradigm lacks cross-view supervision, making the prediction sensitive to viewpoint changes at test time. Without cross-view constraints, the model relies primarily on 2D visual features to regress 3DMM parameters, rather than inferring the 3D structure in the image. Consequently, when viewpoint changes and key facial components become self-occluded, the reconstruction deteriorates rapidly, as shown in Fig. 1.

Recent advances in 3D foundation models [18, 48, 51] have marked a paradigm shift in 3D vision. Rather than relying heavily on inductive biases and task-specific architectures, these models exploit massive amounts of diverse 3D data and high-capacity neural networks to learn general-purpose priors for 3D reconstruction. For example, LRMs [18] reconstruct implicit 3D representations (*e.g.*, triplane-based radiance fields [5, 32]) with cross-view photometric supervisions that are robust to viewpoint changes and missing features. However, these implicit representations are not directly suitable for downstream applications that require structured, parameterized control, leaving a gap between general-purpose reconstruction models and task-specific parametric modeling.

To address this discrepancy, we propose **KaoLRM**[1] to adapt the pre-trained prior of LRM for parametric face reconstruction. Figure 2 illustrates our key motivation. Leveraging LRM's priors provides two main benefits. First, the rich 3D prior encoded in LRM captures 3D structure in images more effectively, improving the robustness to invisible facial components under viewpoint changes. Second, pre-training on a large-scale multi-view general object dataset reduces the reliance on large-scale multi-view facial datasets.

Specifically, we incorporate a FLAME decoder into LRM to project the pre-trained triplane features into the FLAME parameter space, enabling parametric facial geometry prediction. To support cross-view photometric supervision, we model appearance using surface-bounded 2D Gaussian primitives derived from the FLAME mesh, and introduce a binding loss to enforce tight coupling between geometry and appearance. This coupling prevents the model from overfitting to appearance while ignoring geometric accuracy. Experimental results demonstrate that KaoLRM achieves superior reconstruction accuracy and improved cross-view consistency compared to prior methods, validating the effectiveness of transferring LRM's pre-trained prior from general objects to parametric face models.

In summary, our contributions are threefold:

---

[1]"Kao" is derived from the Japanese word for "face".
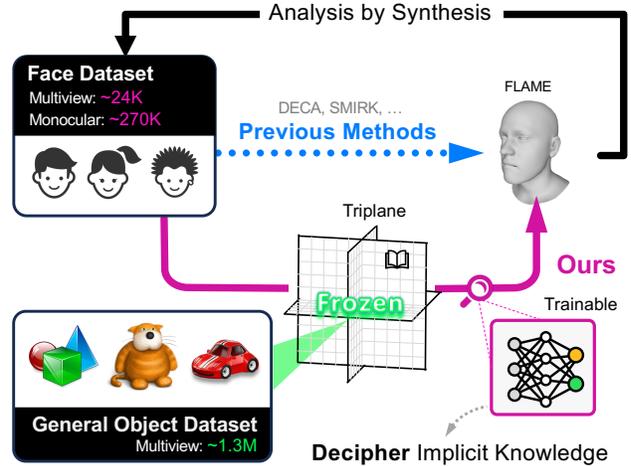

Figure 2. Motivation of KaoLRM. Instead of learning 3DMM regressors from scratch using multi-view face datasets that are expensive to scale, we leverage the learned 3D prior of LRM [18] (*e.g.*, tri-plane features), allowing KaoLRM to be trained with a moderate amount of data.

- We introduce KaoLRM, a framework that re-targets the pre-trained knowledge of LRM for parametric face reconstruction.
- We bridge the gap between LRM's implicit 3D representations and FLAME parameters by integrating a parametric analysis-by-synthesis pathway into LRM's rendering framework.
- KaoLRM outperforms previous methods regarding reconstruction quality and view robustness while maintaining competitive performance in in-the-wild scenarios.

2. Related Work

2.1. Large Reconstruction Model

As the prevailing trend in both neural network architectures and dataset sizes continues to move towards larger scales, one notable outcome in the specific domain of 3D reconstruction is the emergence of the Large Reconstruction Model (LRM) [18]. The underlying philosophy of LRM is to incorporate as little inductive bias as possible, *e.g.*, the camera geometry usually adopted for unprojection, so that the model can learn representations and mappings directly from data, potentially surpassing the constraints of traditional, hand-crafted geometric assumptions. As a result, LRM can regress the triplane-based radiance field directly from a single image in a completely feed-forward manner.

Following LRM, a number of methods have emerged that further investigate and expand upon different facets of this paradigm. LGM [42] and GS-LRM [59] shift the target of regression from radiance fields to pixel-aligned 3D Gaussians; PF-LRM [50] studies the power of LRM in

joint pose and shape prediction; MeshLRM [54] re-targets a NeRF LRM to directly generate meshes with higher quality; LRM-Zero [57] relieves the data reliance of LRM by employing synthetic data for training; LVSM [20] further extends the paradigm to novel view synthesis.

Specifically regarding LRM approaches to digital human avatars, we observe the emergence of methods such as Human-LRM [55] and FaceLift [30], both of which rely on multi-view diffusion models to generate pseudo multi-view inputs and employ LRM-based models to perform reconstruction. However, we observe that these methods produce generic 3D geometric representations, such as point clouds or meshes, which are not readily applicable to downstream tasks like animation, where parametric models such as SMPL [28] or FLAME [26] are typically preferred. In this paper, we opt to explore the possibility of re-targeting LRM to a compact 3D representation, namely a parametrized 3DMM, to achieve a more reasonable balance between generalizability and practicality.

2.2. 3DMM Regression

3DMM [11] refers to parametric and statistical models of 3D face shape (and optionally appearance) built from real 3D scans. The original 3DMM [2] is a typical example where Principal Component Analysis (PCA) is applied to a dataset of registered 3D faces to model variations in both shape and texture. Since then, many specialized 3DMMs have been developed, *e.g.*, BFM [35], LSFM [3], and FLAME [26]. It is also worth noting that in addition to conventional 3DMMs, some works [14, 31] use a learnable neural network to achieve decoding from latent to mesh, thereby enabling parametric modeling with neural networks. In this paper, we exclusively employ FLAME as the target parametric model.

Though there are attempts with direct 3D supervision [60, 62], considering the lack of high-quality 3D ground-truth annotations, most of the methods are trained in a self-supervised manner, by enforcing various types of consistency loss. MoFA [43, 44] introduces a model-based deep autoencoder that enables unsupervised learning via differentiable rendering. This analysis-by-synthesis paradigm was subsequently advanced to achieve fast reconstruction [45] and to learn the generative face model itself from videos [46]. Most of the subsequent 3DMM regressors focus on the design of the synthesis procedure including the choice of appearance modeling methods and the formulation of loss functions—particularly, the type of consistency enforced between the synthesized and observed images, such as photometric, perceptual, or semantic consistency. Following this line of research, RingNet [40] relies on the alignment of facial landmarks and the shape consistency between images of the same identity. DECA [12] further solidifies the analysis-by-synthesis

approach for FLAME by enforcing robust photometric consistency. Meanwhile, for real-time applications, 3DDFA-V2 [16] utilizes a lightweight backbone to regress 3DMM parameters and introduces a meta-joint optimization strategy and a 3D-aided short-video synthesis method. Recent works have integrated perceptual constraints to enhance the realism of facial dynamics in 3D reconstruction. For instance, EMOCA [7] exploits emotion-consistency during training to improve the expressiveness and robustness of the reconstructed avatars under varying emotions and facial expressions. In addition, SPECTRE [13] leverages a lipread loss to improve the reconstruction of mouth movements. To mitigate the impact of domain gap on optimization, SMIRK [38] adopts an analysis-by-neural-synthesis approach that employs a generative neural rendering module to replace the traditional differentiable rendering. SHeaP [41] learns FLAME mesh reconstruction and tracking via rigged head motions and use the combination of Gaussian prototypes based on face-aligned features to model the appearance.

We position our method KaoLRM within the analysis-by-synthesis paradigm, similar to DECA [12], EMOCA [7], and SMIRK [38]. Distinctively, KaoLRM leverages the pre-trained LRM prior [18] to obtain 3D-aware features for FLAME parameter estimation, and models appearance by converting LRM-generated triplane features into a Gaussian-based representation.

3. Method

KaoLRM infers the parameters of pre-defined FLAME [26] from a single image. The key insight is that the pre-trained prior from LRM [18] can be adapted to aid the prediction of compact parametric models. To this end, we integrate the analysis-by-synthesis pipeline for regressing FLAME into LRM's NeRF-targeting framework.

Fig. 5 shows the overall architecture of KaoLRM, which consists of three key components: (1) a pre-trained image-to-triplane transformer from LRM, (2) a 3D-aware decoder to predict FLAME parameters with the triplane tokens, and (3) a lightweight appearance decoder to convert the radiance-relevant features to surface-bounded 2D Gaussian primitives. KaoLRM makes use of the image-to-triplane transformer to predict the feature-level triplane representation $\mathcal{T}$ of the input image $\mathbf{I}$. Based on the triplane-based 3D representation, we predict FLAME parameters and model the local appearance with surface-bounded Gaussians $\{G\}$, so that we can render efficiently and thus become able to use 2D supervision in the image space to supervise the network.

3.1. Transforming Images as Triplanes

LRM [18] employs an image-to-triplane (I2T) transformer to lift 2D image features into the canonical 3D space as triplane representations. The I2T transformer is designed
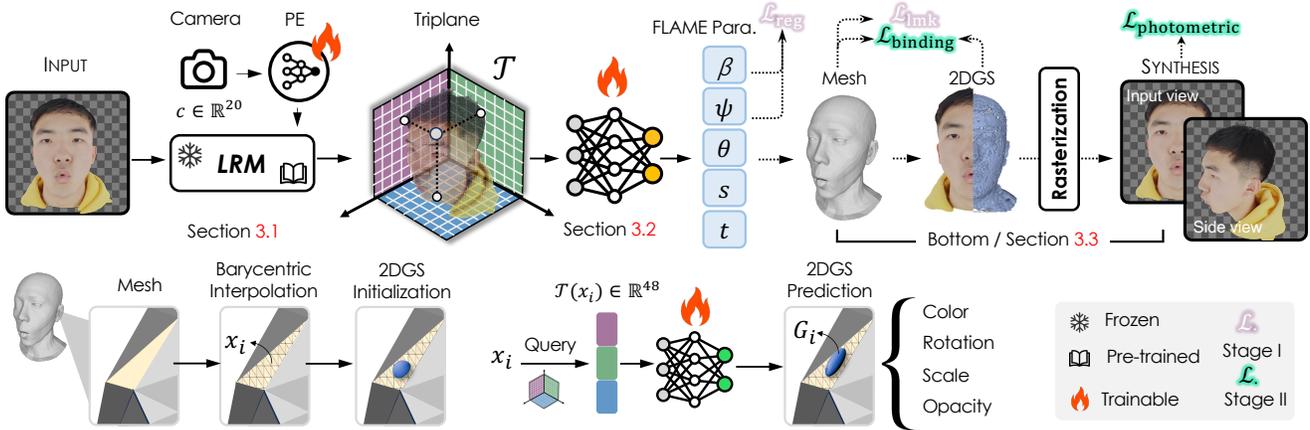
Figure 3. Overall architecture of KaoLRM. **Top**: KaoLRM leverages the pre-trained image-to-triplane transformer of LRM [18], and trains an additional network to predict FLAME parameters from the triplane features. The resulting FLAME mesh is then converted to 2D Gaussian primitives for synthesizing input-view or novel-view images. **Bottom**: After obtaining the FLAME mesh, we sample points via barycentric interpolation to initialize the centers of 2D Gaussians. At each Gaussian center, the corresponding triplane feature is queried and decoded into Gaussian attributes.

with minimal inductive bias and pre-trained on large-scale datasets, enabling it to capture rich geometric and appearance priors. It first extracts visual features by DINOv2 [33], and then camera intrinsic and extrinsic parameters $\mathbf{c} \in \mathbb{R}^{20}$ are flattened, concatenated, and embedded to modulate learnable positional embeddings (denoted as PE in Fig. 3). The final triplane tokens are upsampled to a higher spatial resolution and organized as the triplane representation used for synthesis.

During training, we freeze the pre-trained I2T transformer and keep the camera embedder trainable. This design preserves the representations learned by the I2T transformer, while allowing the model to adapt to the viewpoint distribution of the facial dataset. In early training, facial geometry is still inaccurate, and unreliable supervision may corrupt the learned tri-plane features. Moreover, as KaoLRM fine-tunes LRM [18] on a new face dataset with viewpoints differing from LRM pre-training, freezing the camera embedder would limit effective adaptation.

3.2. Predicting FLAME with Triplane Tokens

LRM [18] decodes triplane tokens into a triplane-based radiance and occupancy field, from which a surface can be extracted by applying marching cubes [29]. This implies that the triplane tokens encapsulate sufficient geometric information, motivating KaoLRM to instead decode them into the parametric space of FLAME.

To effectively decode triplane tokens into FLAME parameters, we introduce a self-gating mechanism to selectively emphasize the token representation. As different facial geometries activate different subsets of triplane tokens, it is crucial to adaptively identify geometry-relevant tokens

for each input image. To this end, we treat the flattened tokens as a sequence and apply a lightweight gating module. A shallow MLP followed by a sigmoid activation assigns a learnable importance score to each token, which is used to modulate the token features by suppressing geometry-irrelevant responses.

We follow DECA [12] in predicting the same set of FLAME parameters: shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{100}$, expression parameters $\boldsymbol{\psi} \in \mathbb{R}^{50}$, and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{6}$, which include global head pose and jaw pose. Unlike DECA [12], which estimates 2D translation under a scaled orthographic camera model, we instead predict a global scaling factor $s \in \mathbb{R}_{+}$ and a translation vector $\mathbf{t} \in \mathbb{R}^{3}$ in world coordinates, enabling novel view synthesis beyond the input view. The FLAME mesh $\mathcal{M}$ is reconstructed as

$$\mathcal{M} = s \cdot V(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}) + \mathbf{t}, \qquad (1)$$

where V is the FLAME model detailed in supplementary.

3.3. Synthesizing Appearance by Gaussians

Since we explicitly predict FLAME parameters and reconstruct a FLAME mesh, we no longer require ray marching to repeatedly query triplane features, as used in the radiance field rendering of LRM [18]. We therefore replace volume rendering with a rasterization-based approach using Gaussian Splatting [23] to improve rendering efficiency. Among Gaussian-based representations, we adopt 2D Gaussian Splatting (2DGS) [19], as it enables surface-aligned appearance modeling and provides more accurate depth and normal estimates for geometric supervision.

A straightforward way to bind 2D Gaussian primitives to the FLAME mesh is to initialize Gaussian centers at mesh

vertices. However, as noted in GaussianAvatars [36], the $5,023$ vertices of a FLAME mesh are insufficient to capture the realistic appearance of the full head. To address this, we sample $8k$ points on the FLAME mesh via differentiable barycentric interpolation, ensuring dense coverage of the facial surface. For each 3D point $x \in \mathbb{R}^3$ sampled from the FLAME mesh M, we query the corresponding triplane feature $\mathcal{T}(x)$ and decode it with an MLP into Gaussian attributes, including opacity, scale, rotation, and color. We apply different activation functions to different Gaussian attributes. Following LGM [42], the scale parameters are activated using SoftPlus and further scaled by a factor of 0.1. Opacity is activated with a standard sigmoid function, while colors are predicted using a slightly saturated sigmoid $f(x) = (1 + 2\epsilon)/(1 + \exp(-x)) - \epsilon$ with $\epsilon = 0.001$, as used in Mip-NeRF [1]. The predicted quaternions are ℓ_2-normalized to ensure unit norm.

Once the 2D Gaussians are obtained, novel views can be synthesized by splatting them under arbitrary cameras intrinsics and extrinsics. This enables photometric supervision during training, effectively encouraging the predicted 3DMM to capture fine-grained facial expressions and details.

3.4. Loss Function

We integrate an analysis-by-synthesis pipeline for FLAME estimation into the pre-trained LRM. Since this integration introduces new supervision signals and optimization requirements, we adapt LRM's loss functions by modifying existing terms and introducing additional ones.

Photometric Loss. The photometric loss encourages the appearance consistency between the rendered and input images, and consists of three terms:

$$\mathcal{L}_{\text{photometric}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{D-SSIM}} + \mathcal{L}_{\text{VGG}}. \quad (2)$$

The pixel loss $\mathcal{L}_{\text{pixel}}$ encourages pixel-wise color consistency in facial regions, and is defined as

$$\mathcal{L}_{\text{pixel}} = \|(\mathbf{I}_{\text{target}} - \mathbf{I}_{\text{render}}) \odot (\lambda \mathbf{m} + (1 - \lambda)\mathbf{1})\|^2, \quad (3)$$

where $\odot$ denotes Hadamard product and $\mathbf{m}$ stands for the pixel-wise binary mask for facial regions. λ balances the influence of facial and non-facial regions. The perceptual loss [10, 21] and D-SSIM loss are computed for the entire image to encourage perceptual consistency.

Geometric Binding Loss. To encourage a tight coupling between the FLAME mesh and 2D Gaussian primitives, we employ a binding loss that minimizes the depth and normal difference between 2DGS and FLAME:

$$\mathcal{L}_{\text{binding}} = \|(\mathbf{D}_{\text{mesh}} - \mathbf{D}_{\text{gs}}) \odot \mathbf{m}\| \\ + \|(\mathbf{N}_{\text{mesh}} - \mathbf{N}_{\text{gs}}) \odot \mathbf{m}\|, \quad (4)$$

where $\mathbf{D}_{\text{mesh}}$ and $\mathbf{N}_{\text{mesh}}$ denote differentiably rendered depth and normal maps from the FLAME mesh, while $\mathbf{D}_{\text{gs}}$ and $\mathbf{N}_{\text{gs}}$ are alpha-blended ones according to Gaussian attributes. Because 2D Gaussian primitives are highly expressive, facial appearance can be reconstructed without faithfully capturing the underlying geometry. The binding loss is therefore essential to prevent the model from overfitting to appearance while neglecting geometric accuracy.

Geometric Regularization Terms. As a common practice in FLAME regression, we also compute the reprojection error $\mathcal{L}_{\text{lmk}}$ of the 68 landmarks. We further apply regularization terms $\|\boldsymbol{\beta}\|^2$ and $\|\boldsymbol{\psi}\|^2$, to prevent FLAME parameters from leaning to bizarre shapes and expressions (*e.g.*, extremely long ears or a protruding forehead), as is a common practice [12, 26].

Staged Training. Based on empirical observations, we use a two-stage training strategy for KaoLRM. The first stage (stage I) trains KaoLRM only with landmark supervision $\mathcal{L}_{\text{lmk}}$ and regularization on shape and expression parameters $\mathcal{L}_{\text{reg}}$, facilitating the learning of head localization and orientation. This stage reduces the optimization difficulty and degrees of freedom, providing a stable initialization for subsequent geometric refinement. In the second stage (stage II), photometric and binding losses are further introduced to refine the reconstruction, encouraging both visual fidelity and geometric consistency. All loss terms are applied per view over all source views, except for the view-independent shape and expression regularization terms.

4. Experiments

4.1. Datasets

We follow a rendering protocol similar to that of Objaverse [8], as is used in the pre-training of OpenLRM [17], and apply it to textured 3D human head datasets, including FaceScape [61], Multiface [56], FaceVerse [49], and Headspace [6]. The only modification is that, rather than uniformly sampling viewpoints over the full sphere, we restrict sampling to the frontal hemisphere to avoid uninformative rear-view renderings. For each 3D asset, we normalize the mesh to fit within the bounding box of $[-1, +1]^3$ and render 32 random views, with all cameras sharing identical intrinsics and pointing toward the origin. To further enhance the performance in in-the-wild scenarios, we finetune KaoLRM on two in-the-wild monocular portrait datasets, FFHQ [22] and CelebA [27]. We evaluate the performance on the test sets of FaceVerse [49] and FFHQ [22], and benchmark shape accuracy on the NoW dataset [40].

Methods	Chamfer Distance			Shape Var(β)		Expr. Var(ψ)	
	mean	median	std	10	full	10	full
DECA [12]	3.17	3.13	0.331	1.91	2.02	2.32	2.48
EMOCA [7]	3.15	3.11	0.359	1.91	2.01	4.61	4.67
SMIRK [38]	3.20	3.18	0.329	9.47	8.48	45.94	47.04
MICA [62]	2.63	2.58	0.381	27.41	28.52	–	–
Ours	**2.68**	**2.67**	**0.157**	**1.46**	**1.54**	**1.01**	**1.10**

Table 1. Quantitative comparison on FaceVerse test dataset [49]. We report Chamfer distances ($\times 10^{-2}$) and cross-view variances of FLAME parameters predicted from single-view inputs under different viewpoints; lower is better for both metrics. "Full" denotes the variance computed over the full parameter dimensions (100 for shape and 50 for expression), while "10" refers to the first 10 dimensions of each.

4.2. Implementation Details

The implementation is based on the open-source code of OpenLRM [17], which is implemented by PyTorch, with mixed precision powered by Accelerate [15]. The differentiable renderer for FLAME depth and normal maps is implemented by PyTorch3D [37]. We train our model using 4 NVIDIA A100 80GB SXM GPUs, with input images resized to 224×224 and rendered images at a resolution of 192×192. Training on multi-view datasets takes approximately 4 days, while fine-tuning on monocular portrait datasets takes approximately 6 days. The weighting factor λ is set to 0.7 to place greater emphasis on facial regions.

For the multi-view datasets, we render 4 source views per subject in one feedforward pass, with a batch size of 16 subjects per GPU. The facial masks $\mathbf{m}$ used in stage II are obtained by projecting the face-wise semantic masks of FLAME [26] onto the image plane. We generate ground-truth 3D landmarks for the datasets not providing. To ensure multi-view consistency of landmarks, we first detect 2D landmarks by FAN [4], triangulate them by the 5 most frontal views to minimize the reprojection error, and then project the 3D landmarks to different views. Note that the triangulated landmarks could still be error-prone so the loss weight should be carefully tuned.

For fine-tuning on monocular datasets, we render a single input view per subject and set the batch size to 32 per GPU. We employ off-the-shelf segmentation methods [24, 25] to obtain facial masks, and use FAN [4] to detect 68 facial landmarks [39]. To resolve the inherent scale–depth ambiguity in monocular reconstruction, (*e.g.*, an object appearing larger or closer), we fix the z-component of the predicted translation vector to zero during training.

4.3. Evaluation Protocol

To demonstrate the effectiveness of our method, we conduct a twofold evaluation: the absolute accuracy of the reconstructed geometry, and the relative consistency of FLAME
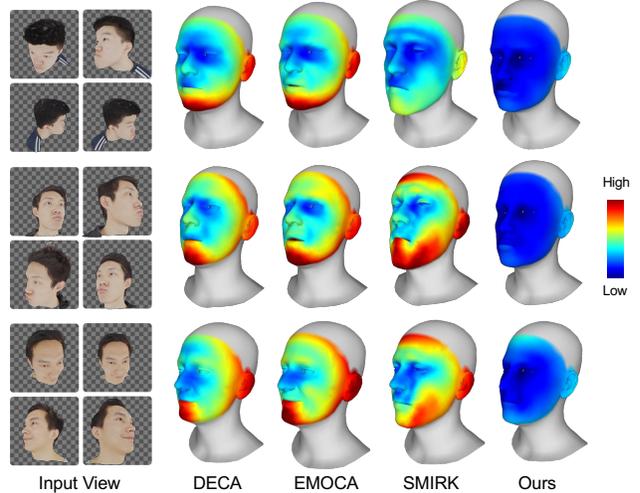

Figure 4. Per-vertex variances in local geometries. The heatmap is plotted on the mean face of multi-view predictions after keypoint alignment. Ours achieves more consistent predictions across viewpoints.

predictions across different viewpoints.

For quantitative comparisons on FaceVerse [49], we follow the default evaluation protocol of NoW [40], computing the Chamfer distance between the predicted mesh and the ground truth over cropped facial regions after performing Umeyama alignment on the estimated keypoints. For geometric accuracy on the NoW benchmark [40], we adopt the default non-metrical evaluation protocol. Following the benchmark setting, expression parameters and poses are set to zero, as only the neutral face geometry is evaluated.

As noted in [53], a good 3DMM regressor should produce consistent predictions across multi-view inputs. Since KaoLRM leverages LRM's prior for 3D-aware FLAME regression, it is expected to achieve strong consistency and robustness. To verify this, we additionally evaluate the variance of the predicted FLAME parameters under different camera views. Let N denote the number of views of the same subject. The variance of the shape parameters is

$$\mathrm{Var}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{\beta}_i - \overline{\boldsymbol{\beta}}\right\|_2^2, \text{ with } \overline{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_i. \quad (5)$$

The variance for the expression parameters $\mathrm{Var}(\boldsymbol{\psi})$ is defined in the same manner.

Baselines. We compare KaoLRM with **(1)** analysis-by-synthesis methods trained with 2D image supervision, including DECA [12], EMOCA (v2) [7], and SMIRK [38], and **(2)** methods trained with direct FLAME parameter supervision, represented by MICA [62].
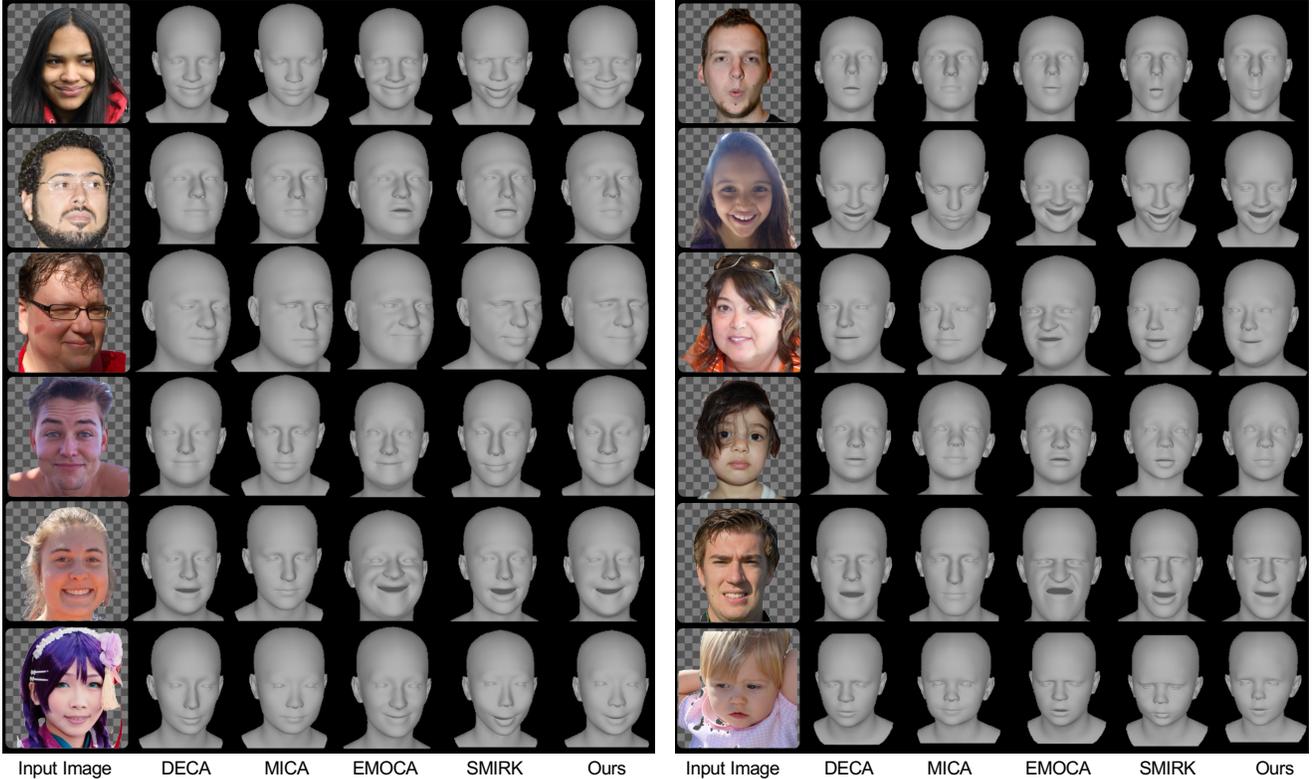
Figure 5. Qualitative comparison on the test set of FFHQ dataset [22]. We align the results of different methods via an estimated transformation of the keypoints. Ours reconstructs shapes more accurately and recovers expressions more faithfully.

4.4. Results

FaceVerse Dataset. The quantitative results of the absolute geometric accuracy and the variance of multi-view predictions on FaceVerse [49] are shown in Tab. 1. Since the FLAME parameters, β and ψ, are obtained via Principal Component Analysis (PCA), the contribution of each parameter decreases from lower to higher indices. To provide more detailed insight, we additionally report variations of the first 10 components. Overall, KaoLRM outperforms the baselines trained with 2D supervision in both metrics.

Table 1 reveals that SMIRK [38] exhibits considerable sensitivity to viewpoint variations, particularly evident from the high variance of its predicted expression parameters. Similarly, MICA [62] shows noticeable sensitivity to view changes in shape prediction. As MICA relies on linear probing over ArcFace features [9], it primarily maps 2D facial appearance features to the FLAME parameter space, making it more sensitive to visible appearance cues. This observation is consistent with our hypothesis that prior methods largely depend on 2D visual features for FLAME prediction. Figure 4 further visualizes geometric variance by color coding per-vertex variances, where KaoLRM exhibits improved cross-view consistency.

FFHQ Dataset. As multi-view datasets are typically collected under highly controlled conditions, they often lack ethnic diversity and contain a limited number of samples. To improve the performance, we further fine-tune KaoLRM on monocular portrait datasets. Since FFHQ dataset does not provide ground-truth 3D meshes, we conduct a qualitative comparison. As shown in Fig. 5, KaoLRM demonstrates perceptually accurate reconstruction of overall facial shape while faithfully recovering facial expressions. In contrast, DECA [12] and SMIRK [38] often fail to capture accurate facial shape, while EMOCA [7] and SMIRK [38] tend to predict exaggerated expressions.

NoW Benchmark. We further evaluate the shape accuracy of all methods on the validation set of the NoW benchmark [40]. While the benchmark nominally assesses only the shape parameters β, the monocular estimation of FLAME shape and expression is intrinsically ill-posed. Consequently, the NoW evaluation also serves as an indirect measure of a method's ability to decouple shape from expression. The quantitative results are summarized in Tab. 2, where Chamfer distances are reported across different challenge categories. Notably, KaoLRM is trained on substantially fewer samples and with less identity diversity than

Methods	Neutral		Expressions		Occlusions		Selfie		Overall	
	mean	median	mean	median	mean	median	mean	median	mean	median
FLAME-neutral [26]	1.36	1.70	1.38	1.73	1.39	1.77	1.35	1.72	1.37	1.73
DECA [12]	1.20	1.52	1.20	1.50	1.34	1.66	1.28	1.62	1.24	1.56
EMOCA [7]	1.19	1.47	1.16	1.45	1.28	1.55	1.25	1.56	1.21	1.49
SMIRK [38]	1.00	1.24	1.02	1.25	**1.03**	**1.24**	1.04	1.27	1.02	1.25
MICA [62]	0.96	1.19	0.95	1.17	0.96	1.18	1.00	1.24	0.96	1.19
Ours	**0.98**	**1.22**	**0.98**	**1.20**	1.03	1.26	**1.01**	**1.24**	**0.99**	**1.23**

Table 2. Challenge-wise and overall prediction errors, reported as Chamfer distances in mm on the NoW validation set [40].

competing approaches. For example, DECA [12] is trained on $\sim$2 million in-the-wild portrait images, while our model is trained on only $\sim$0.3 million. This validates the effectiveness of leveraging the pre-trained prior of LRM learned from general objects for parametric facial reconstruction.

4.5. Ablation Study

We conduct an ablation study to evaluate the impact of training-related design choices, as summarized in Tab. 3. Each component is found to contribute positively to geometric reconstruction quality.

First, removing the binding loss results in the poorest reconstruction quality, as evidenced by the largest Chamfer distance, highlighting the necessity of enforcing a tight coupling between the FLAME mesh and the 2D Gaussian primitives. Notably, removing the binding loss also yields the most consistent cross-view FLAME predictions. This is because, without explicit geometric constraints, the highly expressive 2D Gaussian representation can absorb cross-view appearance variations, reducing the need for the model to explain photometric differences through changes in FLAME geometry. As a result, the predicted FLAME parameters vary less across viewpoints, despite degraded geometric accuracy.

Second, photometric supervision is essential for high-fidelity geometric reconstruction. Without the photometric supervision, KaoLRM relies primarily on the landmark loss, which alone is insufficient to recover detailed facial geometry. Finally, we attempt to train KaoLRM without initializing LRM with pre-trained parameters. In this setting, training fails to converge, likely because the available multi-view facial images are insufficient to train LRM from scratch. As capturing large-scale multi-view facial data is laborious, pre-training LRM on general objects becomes crucial for reducing the data requirements of KaoLRM.

4.6. Limitations

Compared to the closed-source Rodin [52] dataset comprising $100k$ head objects, the amount of training 3D assets available to us remains relatively limited. In addition, Gaussian-based representations typically require a high degree of redundancy to achieve photorealistic rendering. In

Settings	Chamfer Distance			Variance	
	mean	median	std	β	ψ
w/o regularization	2.75	2.74	0.170	2.08	0.41
w/o photometric loss	2.99	2.96	0.200	2.07	0.41
w/o binding loss	3.05	3.05	0.206	**0.75**	**0.27**
w/o pre-training			−		
Full	**2.68**	**2.67**	**0.157**	1.54	1.10

Table 3. Ablation study on the test set of FaceVerse dataset [49]. Variance is computed over the full parameter dimensions. Under the setting of "*w/o* pre-training", the network fails to produce meaningful outputs, and we leave it blank.

contrast, our method employs significantly fewer Gaussians than approaches optimized purely for rendering quality [36], as perfect visual fidelity is not our primary objective. Furthermore, we observe that our method is less effective at capturing eyelid and eyeball movements. This limitation is not reflected in the Chamfer distance, as open and closed eyes exhibit similar local geometry. We attribute this issue to the fact that, due to privacy concerns, eye regions of facial images are often removed or masked out in many 3D head datasets, resulting in limited photometric supervision for these areas.

5. Conclusion

We have presented KaoLRM, a framework that adapts the pre-trained prior of LRM to parametric face reconstruction in an analysis-by-synthesis fashion. By projecting pre-trained triplane features into the FLAME parameter space and coupling the FLAME mesh with surface-bounded 2D Gaussian primitives for appearance modeling, KaoLRM effectively bridges implicit 3D representations with interpretable parametric models. Experiments demonstrate that our method achieves competitive reconstruction accuracy and improved cross-view consistency compared to existing approaches, even when trained with fewer samples and less identity diversity. Our results indicate that the multi-view knowledge and generic 3D representations learned by LRM on general objects can facilitate parametric face estimation when multi-view training data is limited.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 5

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 3

[3] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016. 3

[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. 6

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 2

[6] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2020. 5, 2

[7] 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. 2, 3, 6, 7, 8, 1

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 5

[9] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 741–757. Springer, 2020. 7

[10] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016. 5

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

[12] 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. 1, 2, 3, 4, 5, 6, 7, 8

[13] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5755, 2023. 3

[14] 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. 3

[15] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022. 6

[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. 3

[17] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 5, 6, 2

[18] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4

[19] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 4, 1, 2

[20] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[22] 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. 5, 7, 1, 2, 3

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 4, 1

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 6

[25] Karina Kvanchiani, Elizaveta Petrova, Karen Efremyan, Alexander Sautin, and Alexander Kapitanov. Easyportrait– face parsing and portrait segmentation dataset. *arXiv preprint arXiv:2304.13509*, 2023. 6

[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 Trans. Graph.*, 36(6):194–1, 2017. 3, 5, 6, 8, 1

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5, 1, 2

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 3

[29] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. 4

[30] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. Facelift: Single image to 3d head with view generation and gs-lrm. *arXiv preprint arXiv:2412.17812*, 2024. 3

[31] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 3

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2

[33] 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. *Transactions on Machine Learning Research Journal*, 2024. 4

[34] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *International Conference on 3D Vision 2025*, 2025. 1

[35] Pascal Paysan, Marcel Lüthi, Thomas Albrecht, Anita Lerch, Brian Amberg, Francesco Santini, and Thomas Vetter. Face reconstruction from skull shapes and physical attributes. In *Pattern Recognition: 31st DAGM Symposium, Jena, Germany, September 9-11, 2009. Proceedings 31*, pages 232–241. Springer, 2009. 3

[36] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 5, 8

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

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2490–2501, 2024. 1, 2, 3, 6, 7, 8

[39] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013. 6

[40] 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. 3, 5, 6, 7, 8, 2

[41] Liam Schoneveld, Zhe Chen, Davide Davoli, Jiapeng Tang, Saimon Terazawa, Ko Nishino, and Matthias Nießner. Sheap: Self-supervised head geometry predictor learned via 2d gaussians. *arXiv preprint arXiv:2504.12292*, 2025. 3

[42] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2, 5

[43] 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. 3

[44] 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. 3

[45] 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. 3

[46] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 3

[47] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Transactions on Graphics (TOG)*, 37(2):1–15, 2018. 1

[48] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[49] 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. 5, 6, 7, 8, 2

[50] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[51] 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. 2

[52] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 8

[53] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. Sunstage: Portrait reconstruction and re-lighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20792–20802, 2023. 6

[54] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 3

[55] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*, 2024. 3

[56] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 5, 2

[57] Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Soren Pirk, Arie Kaufman, Xin Sun, and Hao Tan. Lrm-zero: Training large reconstruction models with synthesized data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[58] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 2

[59] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large recon-struction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2

[60] 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. 3

[61] 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*, 45(12): 14528–14545, 2023. 5, 2

[62] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 3, 6, 7, 8, 1