Camera3DMM: Leveraging Perspective Camera for Estimating Parametric 3D Head Models

Vishwesh Vhavle
Mercedes-Benz R&D, IIT Jodhpur
India
vishwesh@iitj.ac.in

Hiteshi Jain
Mercedes-Benz R&D
India
hiteshi.jain@mercedes-benz.com

Avinash Sharma
IIT Jodhpur
India
avinashsharma@iitj.ac.in

Figure 1: *Camera3DMM* enables 3D reconstruction of human heads under perspective distortions. We visualize selfie image samples from the NoW dataset[Sanyal et al. 2019] on top, and our FLAME fitting results on bottom.

Abstract

3D human head modeling is often formulated under scaled-orthographic assumptions, which fail in close-range scenarios such as handheld mobile device captures, where perspective distortion dominates and leads to unstable and inconsistent reconstructions. We propose *Camera3DMM*, a novel perspective-aware 3D human head reconstruction framework that jointly estimates 3D facial geometry and camera parameters from a single image. To address the lack of perspective-rich training data, we leverage high-quality 3D RGB scans to render images with pseudo ground truth labels across diverse focal lengths and perspective distortions, thereby enabling explicit modeling of perspective variability. Trained on this data, *Camera3DMM* achieves stable and consistent reconstructions under varying intrinsics and demonstrates a 22% improvement in mesh quality over the best-performing baseline. These results establish *Camera3DMM* as a strong baseline for perspective-aware 3D face reconstruction, particularly in challenging close-range scenarios.

CCS Concepts

• **Computing methodologies** → **Mesh models**; *Computer vision.*

Keywords

3D Head Parametric Models, Perspective Camera, Tracking, Reconstruction

ACM Reference Format:
Vishwesh Vhavle, Hiteshi Jain, and Avinash Sharma. 2025. Camera3DMM: Leveraging Perspective Camera for Estimating Parametric 3D Head Models. In *SIGGRAPH Asia 2025 Technical Communications (SA Technical Communications '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3757376.3771420

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SA Technical Communications '25, Hong Kong, Hong Kong
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2136-6/2025/12
https://doi.org/10.1145/3757376.3771420

1 Introduction

With recent breakthroughs in 3D rendering techniques, 3D Human Head Modelling has become a well sought-after use case for digital content creation, telepresence, facial biometrics, AR/VR, and entertainment. The task of 3D reconstruction of coarse head geometry, sans fine grained high frequency geometrical details, from 2D images is sufficient for most downstream applications and is a fundamental problem, as the majority of data related to human heads exists in the form of images. The prevailing approach for modeling human heads is to use 3D Morphable parametric models FLAME [Li et al. 2017] and NPHM [Giebenhain et al. 2023]. In these models, the head is represented by parameterizing its person-specific shape, expressions, and poses, which enables articulation across joints typically located near the jaw, neck, etc.

A common approach to this problem is to regress the parametric model from images. In this line of work, the majority of existing methods[Danecek et al. 2022; Deng et al. 2019; Retsinas et al. 2024] assume a scaled-orthographic projection, which greatly simplifies optimization. However, in practical scenarios such as selfie images, video calls, or AR try-on applications (using consumer devices

like smartphones and laptops), the human face is often closer to the camera, where perspective distortion becomes dominant thus violating the scaled-orthographic assumption. This leads to inconsistent reconstruction of face geometry across frames as perspective distortions in images are considered due to false geometric variation in face geometry. Nevertheless, some work has explored fixed field-of-view approaches [Wang et al. 2024] or predicted dense facial representations to optimize the head model parameters for a full perspective projection setup [Giebenhain et al. 2025; Wood et al. 2022]. However, the prediction of the parameters of the full perspective camera remains largely underexplored in head model regression. The primary limitation is the scarcity of large-scale datasets containing 3D face reconstructions captured with diverse camera intrinsics. Most existing feed-forward head model regressors are trained and evaluated on images acquired under imaging conditions where scaled-orthographic projection provides adequate approximation. As a result, these methods cannot capture the complex variability introduced by different focal lengths, fields of view, and close-range distortions. This scarcity of accurate, diverse training data has slowed progress toward perspective-aware human head reconstruction. Recently, field-of-view prediction has shown promising results in human body reconstruction from single images [Patel and Black 2025]. This success is encouraging for extending such approaches to human head modeling. Recent advances in 3D rendering, combined with the democratization of high-quality 3D RGB scan datasets such as NPHM dataset [Giebenhain et al. 2023], have opened new possibilities for addressing this challenge.

To this end, we propose *Camera3DMM*, a novel method for 3D parametric human head reconstruction that incorporates camera parameter prediction directly from a single input image. Specifically, we train separate regressors (named prior models) for shape, expression, and pose parameters of the FLAME model, as well as for camera parameters. We leverage the NPHM dataset's RGB scans, estimating FLAME parameters for rigid registration to serve as pseudo ground truth labels. We then render these meshes with varying camera intrinsics, lighting conditions, and random backgrounds, creating a dataset of images, FLAME parameters, and camera parameters. This enables us to provide high-quality supervision for our prior networks. Finally, during inference, we further optimize our feed-forward estimates using sparse 2D keypoints and projected 3D landmarks from our FLAME estimates. We demonstrate that joint optimization with the estimated FLAME parameters from RGB scans and supervision of full perspective camera parameters leads to substantial quantitative and qualitative improvements over existing baselines, including a 22% relative improvement in mesh reconstruction quality compared to the best prior method on the NPHM dataset hold-out set.

2 Method

Our framework for 3D human head modeling from a single RGB image under a perspective camera model is composed of a data generation module and a two-phase modeling pipeline. In the first phase, four dedicated encoders are trained to estimate shape, expression, pose, and camera parameters. In the second phase, a lightweight linear optimization module refines these initial estimates to achieve higher reconstruction fidelity. In the following section, we provide an overview of these components along with the relevant preliminaries.

2.1 Camera Models

In this work, we consider two commonly used camera models for projecting a 3D head mesh $\mathcal{M} \in \mathbb{R}^{3\times n}$, where each column corresponds to a vertex in canonical 3D space, onto the 2D image plane.

Scaled-Orthographic Model: Under the scaled-orthographic assumption, the 2D projections $V \in \mathbb{R}^{2\times n}$ are obtained as

$$V = s\,\Pi(\mathcal{M}) + t, \tag{1}$$

where $s \in \mathbb{R}$ is isotropic scale, Π is the orthographic 3D-to-2D projection matrix, and $t \in \mathbb{R}^2$ is the 2D translation vector.

Full-Perspective Model: In the full-perspective case, the projection is expressed as

$$V = \Pi\big(K\big(R\mathcal{M} + t\big)\big), \tag{2}$$

where $R \in SO(3)$ and $t \in \mathbb{R}^3$ denote the 3D rotation and translation of the mesh, $K \in \mathbb{R}^{3\times 3}$ is the intrinsic camera matrix, and $\Pi(\cdot)$ is the perspective division operator that maps homogeneous 3D coordinates to 2D by normalizing with depth.

2.2 Dataset generation

Our approach leverages the Neural Parametric Head Models (NPHM) dataset [Giebenhain et al. 2023], which provides high-quality 3D head scans with FLAME topology meshes but lacks corresponding FLAME parameters. Following the rigid alignment optimization methodology proposed by the NPHM authors [Giebenhain et al. 2023], we employ a two-stage optimization process that yields shape parameters ($\beta \in \mathbb{R}^{300}$), expression parameters ($\psi \in \mathbb{R}^{100}$), and jaw pose parameters for each scan.

Figure 2: Sample Visualizations of our Rendering pipeline wih Flame meshes

In our dataset generation, we emphasize lower focal lengths and closer camera distances to introduce perspective-induced distortions, using eight configurations ranging from 600–2400 pixels focal length with distances of 0.3–1.2 meters. To train a more robust camera prior network, more images are rendered at shorter focal lengths where perspective distortions are more prevalent. Images are generated at 512×512 resolution, with random backgrounds from [Quattoni and Torralba 2009], and the projection center fixed at the image center.

Following prior work [Danecek et al. 2022; Deng et al. 2019; Retsinas et al. 2024] we represent head orientation using a rotation parameter $\theta_{\text{rot}} \in \mathbb{R}^3$, defined in the canonical coordinate system of

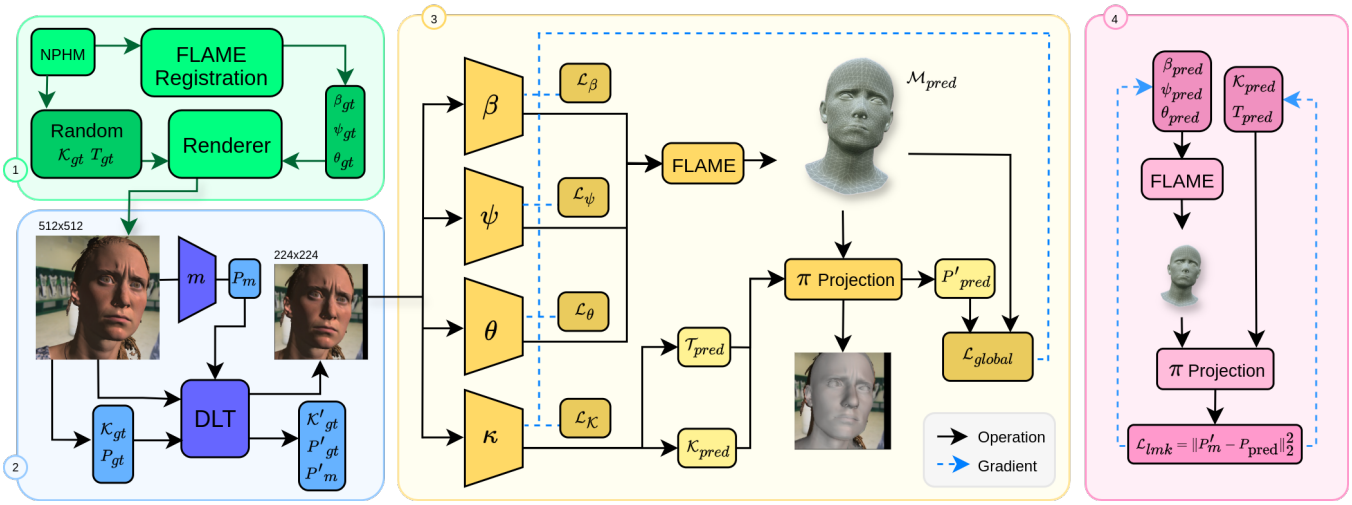

Figure 3: Overview of the proposed method for 3D head reconstruction: (1) Data Generation, (2) Data Transform, (3) Network Training, and (4) Inference with Optimization.

FLAME. To introduce pose variation, we apply random head rotations within ±0.25 radians. The camera is constrained to translation along its Z-axis with a fixed rotation R. This setup simplifies the parameter space by capturing all head orientation changes through θ_{rot}. Furthermore, we maintain a proportional relationship between the focal length and camera distance to preserve perspective distortion while maximizing face resolution.

Training Dataset: Our training dataset comprises 140,000 images derived from 22 scans of 378 subjects. The data encompasses a wide range of ethnicities, shapes, and expressions, with aligned FLAME parameters and ground-truth intrinsics for each sample. Figure 2 shows samples of our rendered dataset.

Evaluation Data: Due to the lack of datasets with ground-truth camera intrinsics and 3DMM parameters, we evaluate *Camera3DMM* on a held-out test set. This set contains 16,500 images of 20 subjects not seen during training.

2.3 Camera3DMM Training

We employ MobileNetV3 networks for four prediction tasks. The shape and expression networks use large MobileNetV3 encoders to predict FLAME shape parameters ($\beta \in \mathbb{R}^{300}$) and expression parameters ($\psi \in \mathbb{R}^{100}$), respectively. The pose and camera networks use small MobileNetV3 encoders, where the pose network outputs FLAME-space rotation (θ_{rot}) and jaw pose parameters (θ_{jaw}), and the camera network predicts intrinsic parameters ($\mathcal{K}'$) for the cropped and resized 224x224 image, which includes focal lengths (f_x, f_y) and principal point coordinates (c_x, c_y), as well as extrinsic translation parameter (t_z).

Our training objective combines multiple loss terms, each corresponding to a specific prediction task. We define a global loss term that supervises all tasks and task-specific objectives for shape, expression, pose, and camera parameters.

Global Loss. We define a global supervision loss that enforces consistency across landmark and vertex predictions:

$$\mathcal{L}_{\text{global}} = \|\mathbf{P}'_{gt} - \mathbf{P}'_{pred}\|_2^2 + \|\mathcal{M}_{gt} - \mathcal{M}_{pred}\|_2^2, \tag{3}$$

where $\mathbf{P}'$ denotes a combination of FAN landmarks [Bulat and Tzimiropoulos 2017] and MediaPipe landmarks [Lugaresi et al. 2019] in cropped image space, and $\mathcal{M}$ denotes FLAME mesh vertices.

Shape Loss. The shape loss supervises the prediction of FLAME shape coefficients β:

$$\mathcal{L}_\beta = \|\beta_{gt} - \beta_{pred}\|_2^2 + \mathcal{L}_{\text{global}} + \lambda_\beta \|\beta_{pred}\|_2^2. \tag{4}$$

Expression Loss. The expression loss supervises FLAME expression coefficients ψ:

$$\mathcal{L}_\psi = \|\psi_{gt} - \psi_{pred}\|_2^2 + \mathcal{L}_{\text{global}} + \lambda_\psi \|\psi_{pred}\|_2^2. \tag{5}$$

Pose Loss. The pose loss supervises head rotation θ_{rot} and jaw pose θ_{jaw}:

$$\mathcal{L}_\theta = \|\theta^{rot}_{gt} - \theta^{rot}_{pred}\|_2^2 + \|\theta^{jaw}_{gt} - \theta^{jaw}_{pred}\|_2^2 + \mathcal{L}_{\text{global}} + \lambda_\theta \|\theta_{pred}\|_2^2. \tag{6}$$

Camera Loss. The camera loss supervises intrinsic and extrinsic parameters. Following Patel et al. [Patel and Black 2025], we design an *asymmetric focal length loss* $\mathcal{L}_f$ that penalizes *underestimation* of focal length more heavily than overestimation:

$$\mathcal{L}_f = \begin{cases} 3\|f_{gt} - f_{pred}\|_2^2 & \text{if } f_{pred} < f_{gt}, \\ \|f_{gt} - f_{pred}\|_2^2 & \text{if } f_{pred} \geq f_{gt}, \end{cases} \tag{7}$$

The full camera loss is then

$$\begin{aligned} \mathcal{L}_\mathcal{K} = \mathcal{L}_f + \|t_z^{gt} - t_z^{pred}\|_2^2 + \|\mathbf{c}_{gt} - \mathbf{c}_{pred}\|_2^2 \\ + \mathcal{L}_{\text{global}} + \lambda_\mathcal{K} \|\mathcal{K}_{pred}\|_2^2, \end{aligned} \tag{8}$$

where $f = (f_x, f_y)$ is the focal length and $\mathbf{c} = (c_x, c_y)$ is the principal point.

2.4 Inference with Optimization

During inference, we optimize all predicted parameters using the landmark loss $\mathcal{L}_{\text{lmk}}$, defined between the MediaPipe landmarks $\mathbf{P}'_m$ and the projected landmarks from the predicted mesh $\boldsymbol{\mathcal{M}}_{\text{pred}}$.

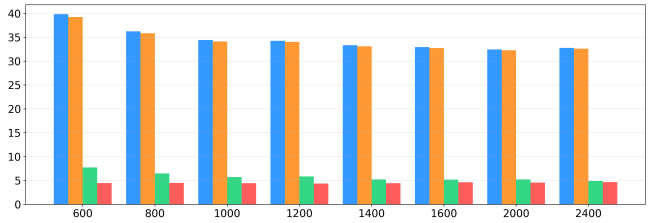

Figure 4: Vertex error across focal length values for different methods: DECA (blue), EMOCAv2 (orange), SMIRK (green), and Camera3DMM (red).

3 Results

Table 1: Method Comparison

Method	Landmarks (↓)	Vertex (↓)
DECA	1.173 ± 36.422	34.8 ± 23.8
EMOCAv2	1.165 ± 36.466	34.5 ± 23.5
SMIRK	1.040 ± 31.890	5.9 ± 7.3
Ours (Ortho)	0.261 ± 0.256	5.8 ± 6.0
Ours (w/o Optim)	0.367 ± 0.290	5.7 ± 5.9
Ours	**0.073 ± 0.039**	**4.5 ± 4.6**

As shown in Table 1, our method outperforms DECA [Deng et al. 2019], EMOCA-v2 [Danecek et al. 2022], and SMIRK [Retsinas et al. 2024], achieving the lowest mean losses for both landmarks and vertex error. Compared to the best-performing prior method SMIRK, we improve vertex error that actually defines the 3D head modeling performance by about **22%**.

As shown in Table 1, the ablation results clarify the source of improvements in our approach. Replacing our perspective projection with an orthographic one leads to a clear drop in accuracy. Using perspective projection without the final optimization step yields a modest improvement, but the largest gains are achieved when both perspective modeling and optimization are combined. This demonstrates that the improvements are not solely due to training with a perspective-aware dataset, but arise from the complete pipeline that explicitly models perspective and refines predictions through optimization.

Figure 4 illustrates the comparative performance of our method against existing baselines across varying focal lengths. While our approach consistently achieves lower vertex errors than DECA [Deng et al. 2019], EMOCA [Danecek et al. 2022], and SMIRK [Retsinas et al. 2024], the most pronounced improvements are observed at shorter focal lengths, where perspective distortions are strongest and scaled orthographic assumptions break down. In this challenging regime, our method provides a clear margin of advantage, highlighting its robustness to geometric distortions that degrade the performance of scaled-orthographic-based models. This behavior directly validates the core motivation of our work: by explicitly incorporating perspective modeling, we are able to handle cases where classical scaled orthographic formulations fail, thereby addressing a fundamental limitation in prior approaches.

Qualitative comparisons in Figure 5 further highlight the effectiveness of our approach across different focal lengths. The FLAME

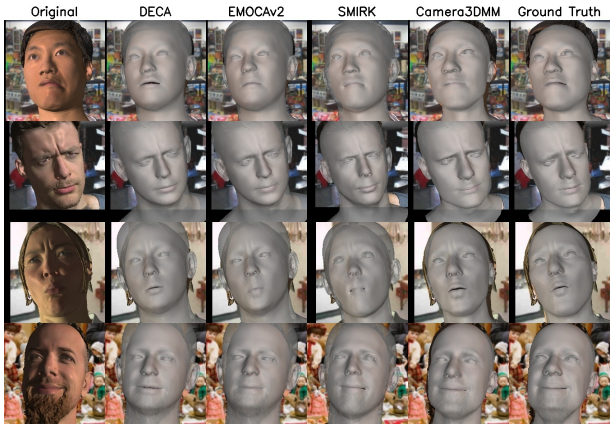

Figure 5: FLAME meshes predicted by DECA, EMOCAv2, SMIRK, and Camera3DMM(Ours) against ground truth.

renderings obtained with our method exhibit more accurate facial alignment and fewer artifacts than those produced by the baselines.

References

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Radek Danecek, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 20311–20322.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *IEEE Computer Vision and Pattern Recognition Workshops*.

Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2023. Learning Neural Parametric Head Models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2025. Pixel3DMM: Versatile Screen-Space Priors for Single-Image 3D Face Reconstruction. doi:10.48550/arXiv.2505.00615 arXiv:2505.00615.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. doi:10.48550/arXiv.1906.08172 arXiv:1906.08172.

Priyanka Patel and Michael J. Black. 2025. CameraHMR: Aligning People with Perspective. In *International Conference on 3D Vision (3DV)*.

Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 413–420. doi:10.1109/CVPR.2009.5206537 ISSN: 1063-6919.

George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 2024. 3D Facial Expressions through Analysis-by-Neural-Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. 2019. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 7763–7772.

Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, and Zhen Lei. 2024. 3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1672–1682.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, and others. 2022. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*. Springer, 160–177.