Supplementary Material for Camera3DMM

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

ACM Reference Format:
Vishwesh Vhavle, Hiteshi Jain, and Avinash Sharma. 2025. Supplementary Material for Camera3DMM. 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

1 FLAME Model

We employ the FLAME model [2], a 3DMM that parameterizes shape, expression, and pose in a unified framework. FLAME represents facial shape variations through $\beta \in \mathbb{R}^{300}$, which encodes identity-dependent geometry, and expression-specific deformations through $\psi \in \mathbb{R}^{100}$. Pose articulation is modeled with joint rotations, where $\theta_{\text{jaw}} \in \mathbb{R}^3$, $\theta_{\text{eyes}} \in \mathbb{R}^6$, and $\theta_{\text{neck}} \in \mathbb{R}^3$ denote the jaw, eyes, and neck rotations, respectively. In this work, we do not explicitly model the eye and neck poses.

2 Data Preparation

Our approach leverages the Neural Parametric Head Models (NPHM) dataset [1], which provides high-quality 3D head scans with FLAME topology meshes but lacks corresponding FLAME parameters. To address this, we develop a gradient-based optimization pipeline to fit FLAME parameters to the FLAME meshes.

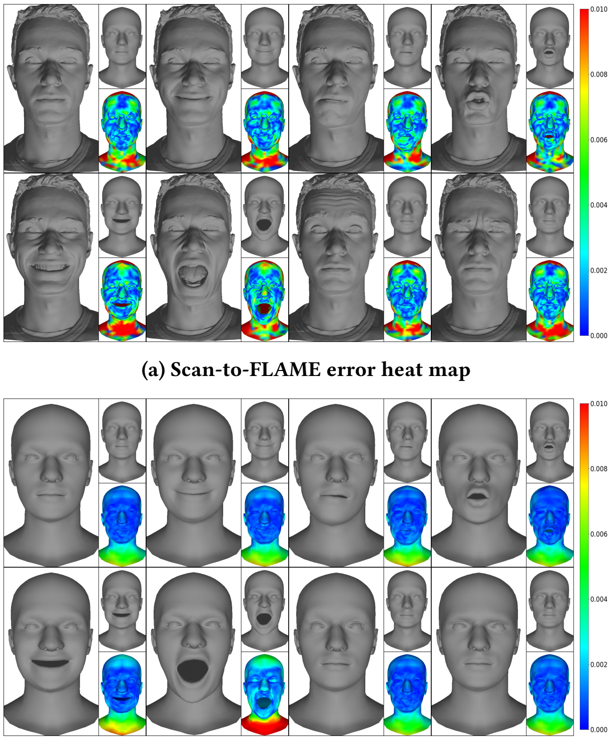
3 FLAME Parameter Estimation

Following the methodology proposed by the NPHM authors [1], we employ a two-stage optimization process. First, we identify shape parameters across all scans of each subject to maintain identity consistency. Subsequently, we estimate expression and jaw parameters for each individual scan. This process yields shape parameters ($\beta \in \mathbb{R}^{300}$), expression parameters ($\psi \in \mathbb{R}^{100}$), and jaw pose parameters for each scan.

The optimization minimizes the vertex-to-vertex distance between the NPHM provided FLAME mesh and the FLAME reconstruction through parameters $(\beta, \psi, \theta_{jaw})$ along with regularisation ($\mathcal{R}$):

$$\mathcal{L}_{fit} = ||\mathcal{M}_{FLAME} - \mathbf{FLAME}(\beta, \psi, \theta_{jaw})||_2^2 + \mathcal{R}. \tag{1}$$

where $\mathbf{FLAME}(\beta, \psi, \theta_{jaw})||_2^2$ denotes the FLAME mesh vertices parameterized by shape, expression, and jaw pose, and $\mathcal{M}_{FLAME}$ represents the NPHM vertices.



(a) Scan-to-FLAME error heat map

(b) FLAME-to-FLAME error heat map

Figure 1: Visualization of 8 random scans of a subject with estimated FLAME meshes and error heat maps.

4 Training Setup and Hyperparameters

The model is trained on a single NVIDIA A6000 GPU using the setup and hyperparameters listed in Table 1. Hyperparameters and loss weights are fixed across all datasets, and selected based on validation performance to ensure fairness in comparison.

5 Inference-Time Optimization

To further refine predictions at test time, we employ an optional gradient-based optimization procedure that fine-tunes FLAME parameters using the initial network predictions as initialization. This inference optimizer iteratively adjusts shape, expression, pose, and jaw parameters to minimize reprojection error between projected landmarks from the predicted mesh and detected 2D MediaPipe landmarks, while maintaining proximity to the network's initial estimates through regularization.

The optimization minimizes the following objective:

$$\mathcal{L}_{lmk} = ||\mathbf{P}'_m - \mathbf{P}'_{pred}||_2^2 + \mathcal{R}_{params}, \tag{2}$$

SA Technical Communications '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2136-6/2025/12
https://doi.org/10.1145/3757376.3771420

Table 1: Training hyperparameters and data augmentations.

Hyperparameter	**Value**
Optimizer	Adam
Learning rates	Shape/Expr/Pose/Camera: 1×10^{-4}
Schedulers	CosineAnnealingWarmRestarts ($T_0 \approx 0.1$ of total steps; $\eta_{\min} = 0.1 \cdot \text{LR}$ for shape, $0.01 \cdot \text{LR}$ for others)
Weight decay	0
Global loss mix-in	Landmarks (MP/FAN): 30, Vertex: 10000, Global: 0.35
Shape losses	MSE: 1.0; Reg: 0.01
Expression losses	MSE: 0.1; Reg: 0.01; Jaw MSE: 1000; Jaw Reg: 1.0
Pose losses	Pose MSE: 10
Camera losses	Focal: 1×10^{-5} (penalty multiplier 3); Center: 10^{-3}; Translation: 10
Input resolution	224 × 224
Data augmentations	RandomBrightnessContrast (p=0.5), RandomGamma (p=0.5), ColorJitter (brightness=0.05, contrast=0.05, saturation=0.05, hue=0.05, p=0.25), CLAHE (p=0.255), RGBShift (p=0.25), Blur (p=0.1), GaussNoise (p=0.5)

Table 2: Inference-time optimization hyperparameters.

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	1×10^{-2}
Learning rate decay	Factor: 0.5, Patience: 30 iterations
Maximum iterations	40
Loss weights	
Landmark (MediaPipe)	1.0
Vertex consistency	0.1
Landmark optimization weight	1000.0
Optimized parameters	
Shape (β)	Yes
Expression (ψ)	Yes
Pose	Yes
Jaw (θ_{jaw})	Yes

where $||\mathbf{P}'_m - \mathbf{P}'_{pred}||_2^2$ denotes landmark reprojection error, and $\mathcal{R}_{params}$ provides regularization on parameter deviations from network predictions.

The hyperparameters for the inference optimizer are detailed in Table 2. This optimization is performed independently for each input image and typically converges within 40 iterations, providing refined geometry particularly beneficial for challenging cases with extreme poses or occlusions.

6 Evaluation Metrics

We evaluate our approach using two complementary metrics: reprojection error on landmarks and mesh reconstruction error. For landmarks, we follow the widely used MediaPipe indexing convention, but instead of using detections from the MediaPipe library, we compute reprojection errors between the predicted FLAME mesh landmarks and the ground-truth FLAME landmarks at the same indices. Specifically, given predicted 2D landmark positions $\hat{\mathbf{p}}_i$ and ground-truth 2D positions $\mathbf{p}_i$ for N indexed landmarks, the mean reprojection error is defined as

$$E_{\text{lm}} = \frac{1}{N} \sum_{i=1}^{N} ||\hat{\mathbf{p}}_i - \mathbf{p}_i||_2, \tag{3}$$

and we report the mean $\mu(E_{\text{lm}})$ and standard deviation $\sigma(E_{\text{lm}})$ across the evaluation set.

To assess mesh quality, we compute vertex-to-vertex reconstruction error between the predicted mesh $\hat{\mathbf{v}}_j$ and ground-truth mesh $\mathbf{v}_j$, with M vertices:

$$E_{\text{mesh}} = \frac{1}{M} \sum_{j=1}^{M} ||\hat{\mathbf{v}}_j - \mathbf{v}_j||_2. \tag{4}$$

These two metrics together capture both the 2D consistency of reprojections and the 3D fidelity of the reconstructed mesh geometry.

7 Limitations and Future Work

While our method demonstrates effective 3D geometry reconstruction from monocular RGB input, several avenues remain for future exploration. For example, explicit modeling of teeth, and detailed eyelid deformation would provide more complete facial representations. Additionally, incorporating temporal regularization techniques could better exploit inter-frame coherence in video sequences, leading to more stable and temporally consistent reconstructions.

8 Conclusion

We have presented *Camera3DMM*, a novel approach for 3D head modeling that explicitly models full perspective camera parameters alongside traditional 3DMM parameters. Overall, our work demonstrates that careful consideration of camera modeling and targeted synthetic data rendering can significantly improve 3D head modeling in practical scenarios with high perspective distortions.

References

[1] 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)*.

[2] 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

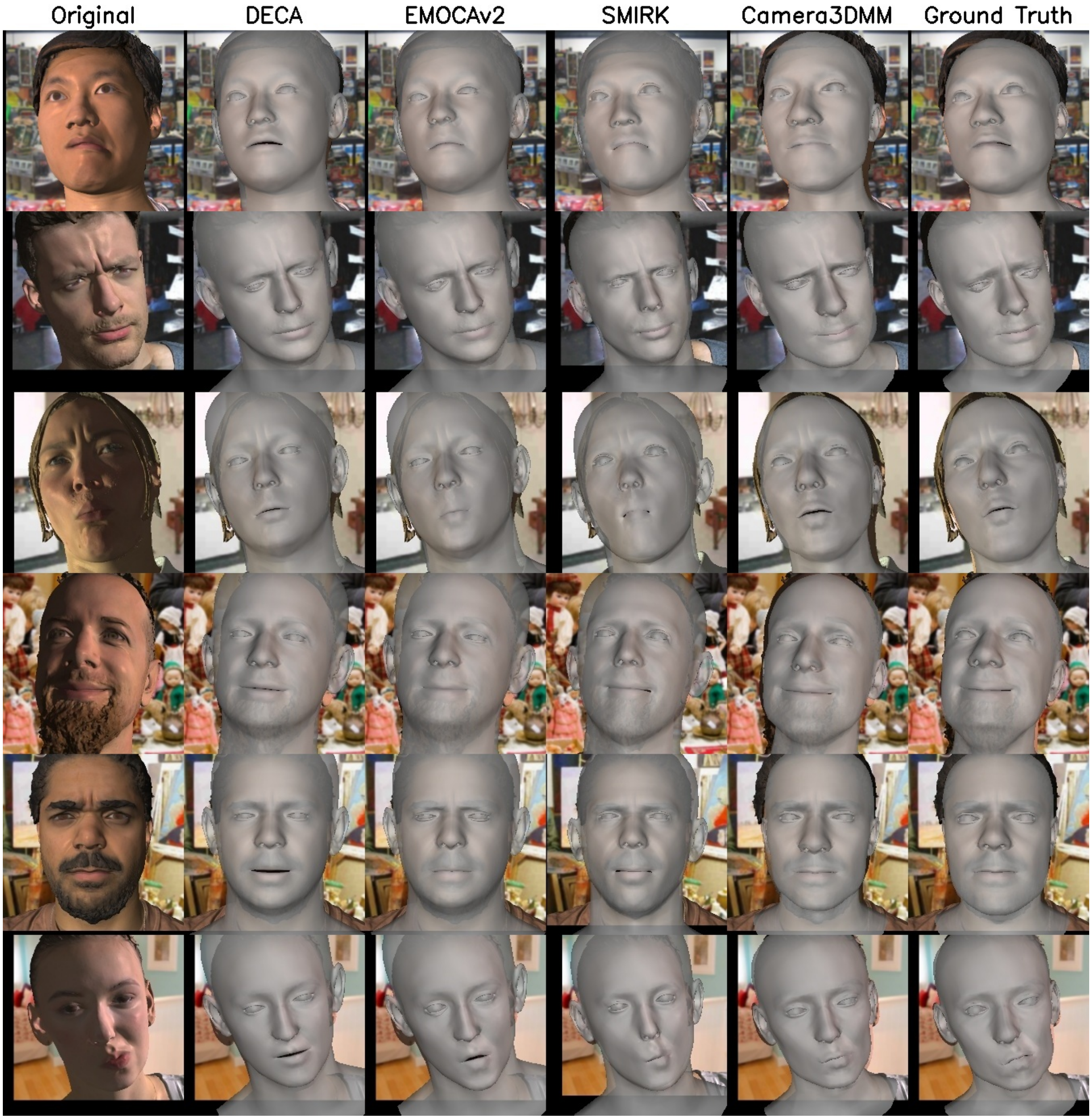


Figure 2: Qualitative Comparisons of FLAME meshes predicted by DECA, EMOCAv2, SMIRK, and Camera3DMM(Ours) against ground truth.

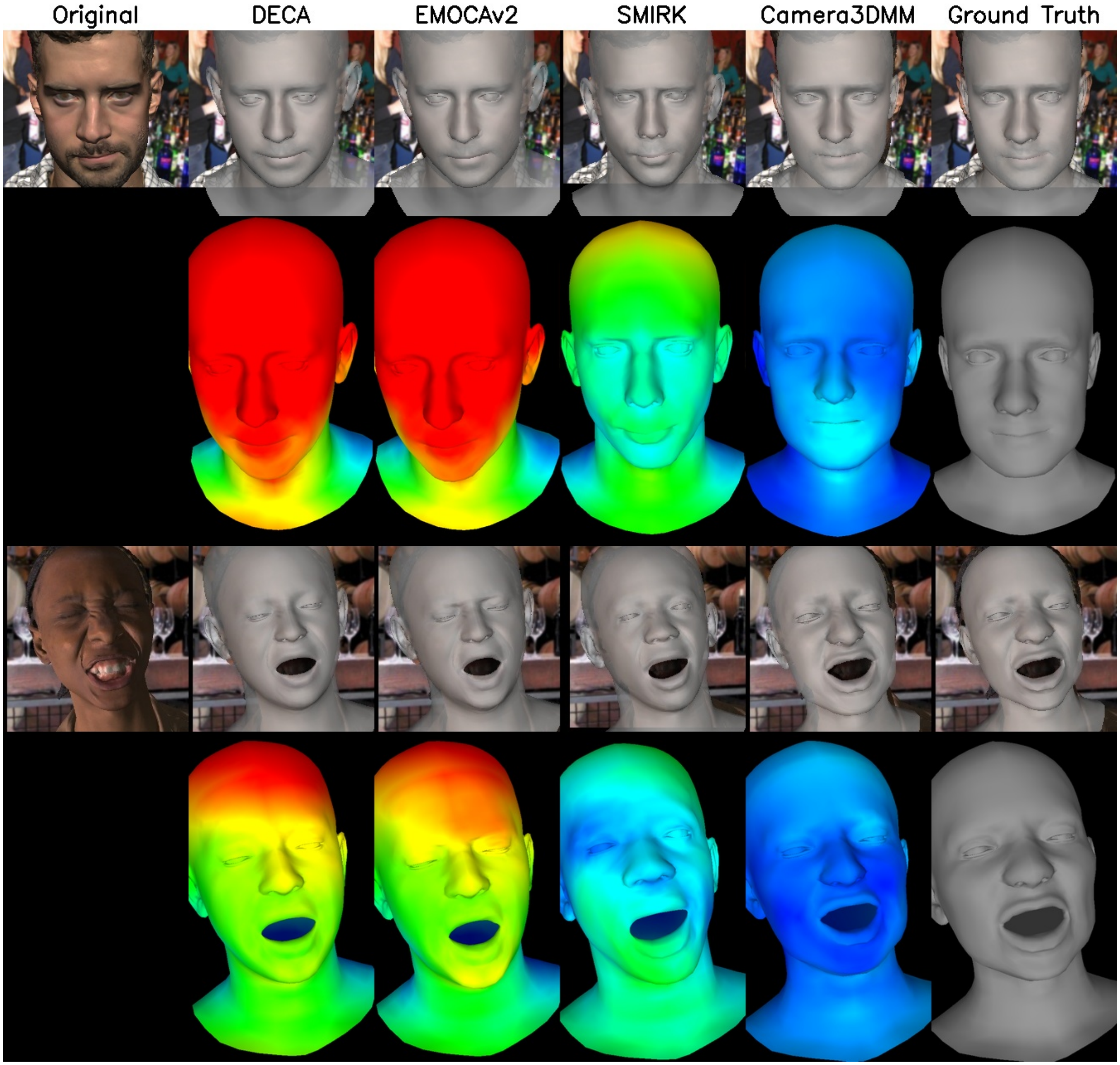


Figure 3: Qualitative Comparisons along with corresponding Vertex error of FLAME meshes predicted by DECA, EMOCAv2, SMIRK, and Camera3DMM(Ours) against ground truth.