# Learning Personalized High Quality Volumetric Head Avatars from Monocular RGB Videos Supplementary Material

Ziqian Bai[1,2*] Feitong Tan[1] Zeng Huang[1] Kripasindhu Sarkar[1] Danhang Tang[1]
Di Qiu[1] Abhimitra Meka[1] Ruofei Du[1] Mingsong Dou[1] Sergio Orts-Escolano[1]
Rohit Pandey[1] Ping Tan[2] Thabo Beeler[1] Sean Fanello[1] Yinda Zhang[1]
[1] Google [2] Simon Fraser University

We provide additional information in this supplementary material, including Warp Field Formulation (Sec. A), Implementation Details (Sec. B), Examples of Data (Sec. C), as well as Image and Video Results (Fig. A, Sec. D, and the accompanying supplementary webpage). Please see our project webpage augmentedperception.github.io/monoavatar for more results.

## A. Warp Field Formulation

**Motivation**. Though the 3DMM fitting can reasonably track the head and expression motions, there are still adhoc motions that cannot be handled by the 3DMM, such as the hair movements and tracking errors, which lead to misalignments between the 3DMM mesh and images and cause the model to learn blurred appearances.

As described in Sec. 3.1, inspired from prior works on deformable NeRF [4, 10], we learn error-correction warp fields with small magnitudes during training to reduce the misalignments, enabling the model to learn sharper appearances. During testing, we discard the warp fields since they are overfit to training frames. Since the warp fields are small in magnitudes (encouraged by the loss function $\mathcal{L}_{mag}$ in Eq.3), they do not affect the inference heavily. As a result, the renderings are equally sharp, albeit with slightly missaligned finer details compared to the ground truth.

**Formulation**. We input the original query point $\boldsymbol{q}$ and a learnable per-frame latent code $\boldsymbol{e}_i$ ($i$ is frame index) into the error-correction MLPs $\mathcal{F}_{\mathcal{E}}$ to predict a rigid transformation. The rigid transformation contains a rotation $\boldsymbol{R} \in SO(3)$ , a rotation center $\boldsymbol{c}^{rot}$, and a translation $\boldsymbol{t}$. Finally, the rigid transformation is applied to the query point to obtain the warped point $\boldsymbol{q}'$. Formally, we have

$$\boldsymbol{R}, \boldsymbol{c}^{rot}, \boldsymbol{t} = \mathcal{F}_{\mathcal{E}}\left(\boldsymbol{q}, \boldsymbol{e}_i\right) \quad (1)$$

$$\boldsymbol{q}' = \boldsymbol{R}\left(\boldsymbol{q} + \boldsymbol{c}^{rot}\right) - \boldsymbol{c}^{rot} + \boldsymbol{t}, \quad (2)$$

where $\boldsymbol{R}$ is parameterized by a pure log-quaternion predicted by the MLPs. We denote the full transformation as $\boldsymbol{q}' = \mathcal{T}_i(\boldsymbol{q}) = \mathcal{F}_{\mathcal{E}}(\boldsymbol{q}, \boldsymbol{e}_i)$. Then, the warped point is used as the query point to decode the density and color as described in Sec. 3.1. Note that this warping field is only used during training and disabled during testing.

## B. Implementation Details

To improve the training convergence, we remove the background [3, 7] and align the head in 3D space by normalizing the 3DMM vertices with its neck pose. Similar to NeRF [6], our full model is hierarchical with the coarse and the fine networks, which are simultaneously optimized by a photometric reconstruction loss. To ensure stable training, we disable 3D warping field in the first 5k iterations, and enable it in the following iterations. For optimization, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size is set as 1024 rays and the learning rates are empirically set to: (1) $10^{-4}$ and exponentially decay to $10^{-5}$ after 400k for warp field networks. (2) $10^{-3}$ and exponentially decay to $10^{-4}$ after 400k for other networks. We train the model with total 400k iterations for each subject. We adapt coarse-to-fine positional encoding (as used in Nerfies [8]) on the coordinate input of the warp field networks for better training stability. More specifically, we start with 0 frequency bands and linearly increase to 6 after 80k iterations. For other modules, we adapt positional encoding as in NeRF [6] with 10 frequency bands on all coordinate inputs and 4 on camera views.

### B.1. Network Architecture

As detailed in the main paper, the framework consists of three modules: a 3DMM-anchored NeRF, a expression-dependent feature predictor, and a warping field predictor.

[*] Work done while Ziqian Bai was an intern at Google.

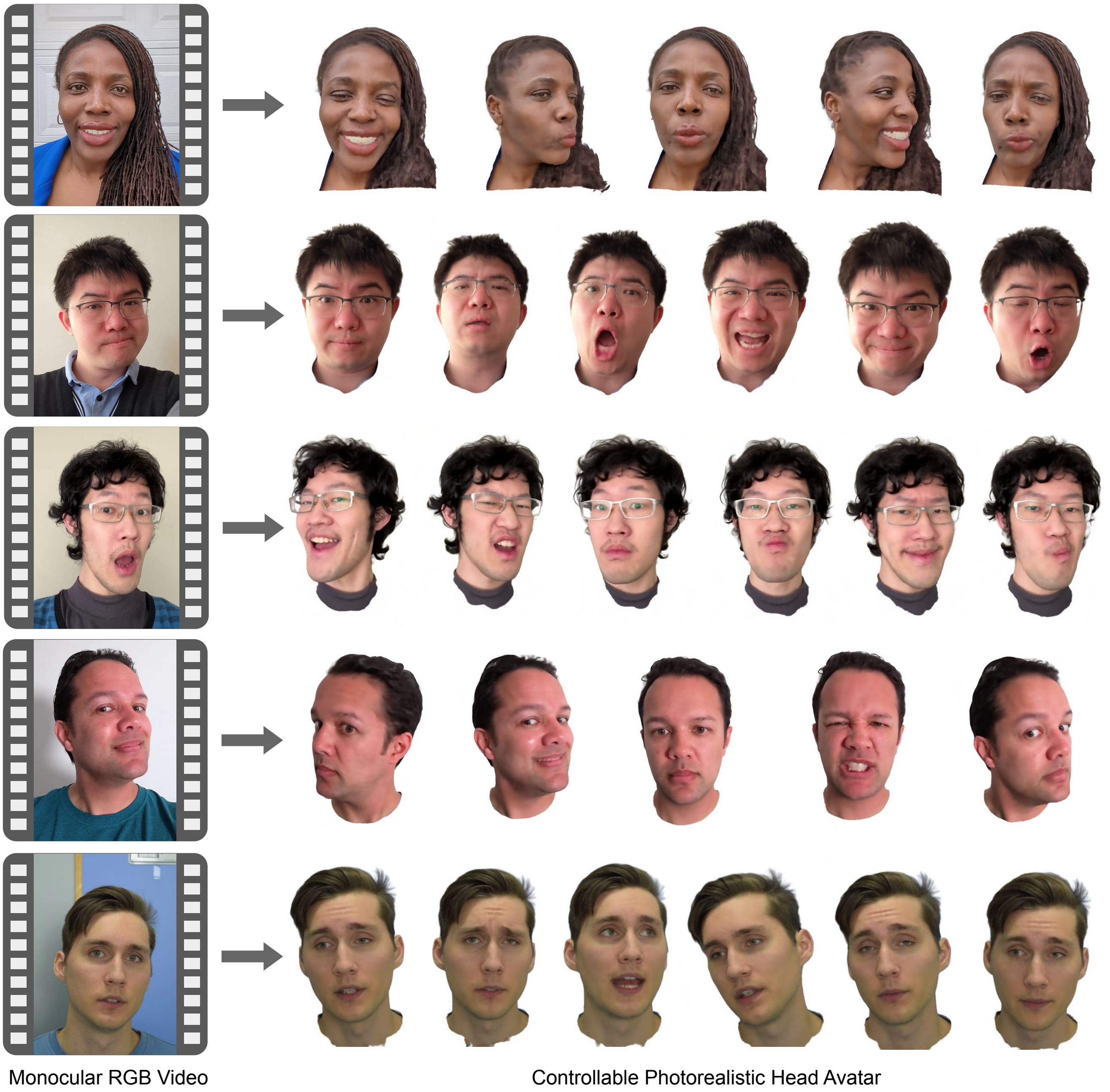

Figure A. We propose a method to build a 3D avatar representation of a person using just a single short monocular RGB video (e.g., 1-2 minutes), which can be rendered with user-defined expression and viewpoint. Note how our method captures extreme expressions and fine scale facial details. Please check our supplementary webpage for more video results, and discussions on the limitation.

### B.1.1 3DMM-anchored NeRF

As described in Sec. 3.1 of the main paper, we adopt the 3DMM-anchored neural radiance field (NeRF) to represent our head avatar. As shown in Fig. B, we attach 64-dimensional feature vectors on each vertex of the FLAME model [5], which are predicted from the U-Net described in Sec. 3.2. During inference, we first concatenate the normalized coordinates $\boldsymbol{v}_i^j - \boldsymbol{q}$ (positional encoded) of the vertex with it's corresponding attached features and pass them into the MLP0, which comprises 3 hidden layers with 128 neurons each and applies ReLU activation, to produce latent features. We then aggregate the latent features of the nearest 4 vertices by a inverse-distance based weighted sum. The aggregated feature is then decoded into density and color

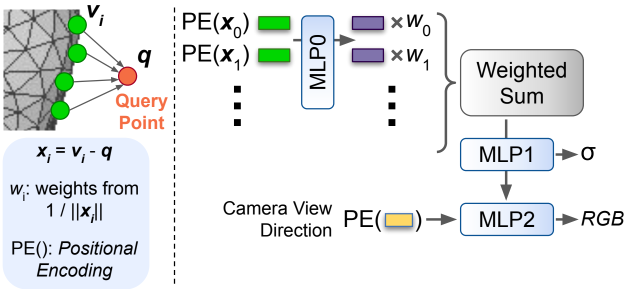

Figure B. Illustration of Avatar Representation. Given a query point, we find its k-Nearest-Neighbor (k-NN) vertices from the 3DMM. Then, we decode these vertices and features into a density and color with respect to the input camera view direction, via Multi-Layer-Perceptrons (MLPs) interleaved with inverse-distance based weighted sum.

with 2 branches. For density, the aggregated feature is decoded by MLP1 + a Fully Connected (FC) layer. For color, the aggregated feature is decoded by MLP1 + MLP2. MLP1 comprises 3 hidden layers with 128 neurons each and applies ReLU activation. MLP2 comprises 1 hidden layers with 64 neurons and 1 FC layer with 3 outputs. To handle view-dependent effects, we also pass the ray view direction (positional encoded) into the MLP2 to decode the RGB color.

### B.1.2 Expression-Dependent Features Predictor

Our expression-dependent features predictor is a 6-level residual U-Net. We use residual blocks to extract feature, and the feature channels of each level are set as 8, 16, 32, 64, 128, 256. In the decoder, residual blocks with transposed convolutions are applied to increase the spatial resolution. The leaky ReLU is applied after each convolutional layer with slope 0.2. The input of the predictor is a 3D deformation map in $256 \times 256$ resolution which stores the vertex displacements from the neural expression to the current facial expression in UV space.

### B.1.3 Warping Field Predictor

The error-correction MLPs $\mathcal{F}_{\mathcal{E}}$ is utilized to predict error-correction warping fields to reduce misalignments from 3DMM and improve the training. It consists of 5 hidden layers with 128 neurons each, followed by ReLU activation, then 3 branches of two-layers MLPs with 128 neurons are added at the end for regressing each output (as described in Sec. A: pure log-quaternion of the rotation (*i.e.*, SO(3)) $\boldsymbol{R}$, rotation center $\boldsymbol{c}^{rot}$, and translation $\boldsymbol{t}$).

## B.2. 3DMM Fitting Details

We have implemented the same optimization-based fitting algorithm as NHA [2] with the following differences: We 1) used MediaPipe for improved nose, eyes, and eyebrows landmarks; 2) re-initialized camera poses (by Perspective-n-Point) and expression parameters (to neutral) every 200 frames to prevent local optima; 3) increased optimization steps per frame to accommodate for more challenging expressions in our data. Note that we use the same fitting results across all methods for a fair comparison.

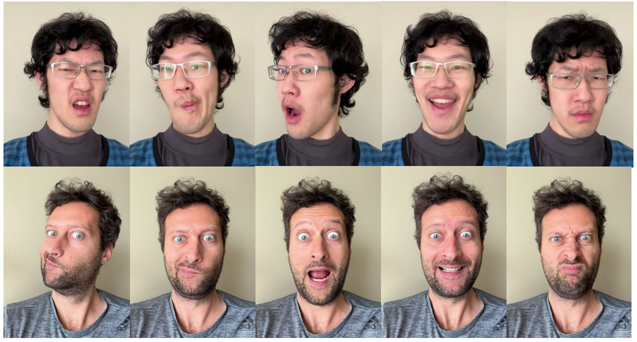

Figure C. Examples of reference expressions for video capture.

## B.3. Video Capture Protocol

We ask users to capture 1-2 min selfie videos with high resolution (over 500 × 500 pixels in the head) under well-lit conditions using phone/webcam, following instructions below (the same as in Sec.4.1). For the training clip, the users are asked to first keep a neutral expression and rotate their heads, then perform different expressions during the head rotation, with extreme expressions included. For the testing clip, the users are asked to perform freely without any constraints. We provide several reference expressions shown in Fig. C for users to follow, but users are not asked to strictly perform the same expressions.

## C. Examples of Data

| Our Data | | | | NerFACE Data |
|---|---|---|---|---|
| *Subject0* | *Subject1* | *Subject2* | *Subject3* | *Subject4* |
| 0.657 | 0.610 | 0.589 | 0.796 | 0.426 |

Table A. The standard deviations of fitted 3DMM expression codes, averaged across code dimensions, on different subjects.

We include data examples (Fig. D) to show that our data has a large expression coverage, thus is more challenging than talking head style data used by prior works [1]. We also compare the standard deviations of fitted 3DMM expression codes (averaged across code dimensions) on our data and NerFACE data. As shown in Tab. A, our data has significantly larger standard deviations, which indicates more diverse expression coverage in our data.

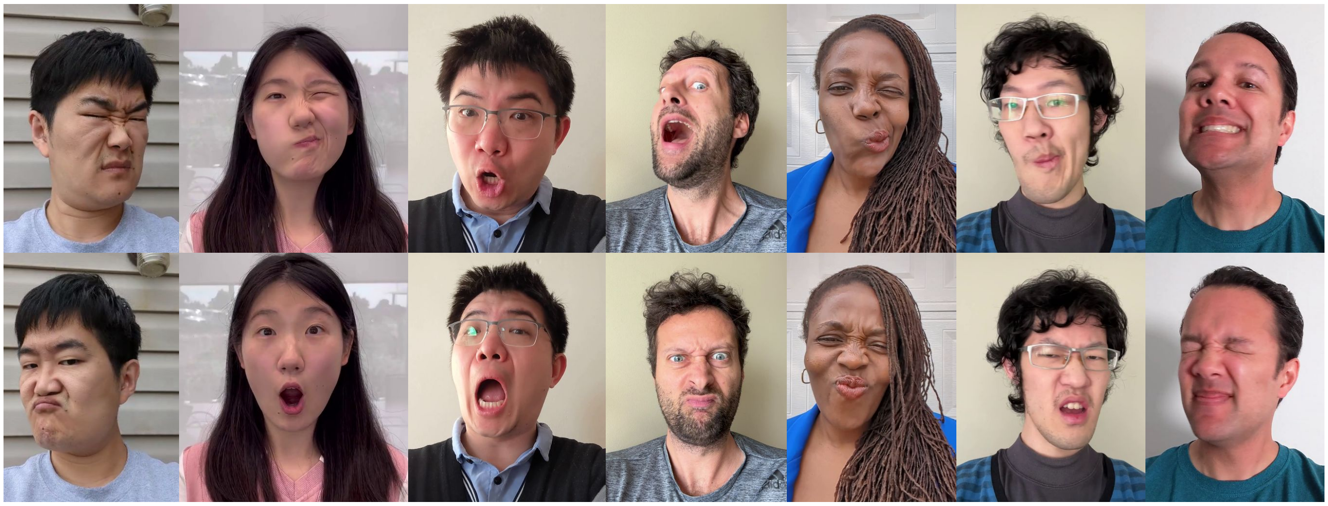

Figure D. Examples of our captured training data, which includes various large expressions.

# D. More Results

In this section, we provide more qualitative comparisons of our method with state-of-the-art techniques and ablations against our design choices. We also demonstrate the robustness of our method in challenging cases where the generated avatar is driven under significantly different conditions than the original training sequence.

## D.1. Multi-subject Comparison with SOTA

Fig. E shows a comparison of our method against state-of-the-art techniques across several subjects for non-neutral expressions. Note that our technique is able to faithfully model and synthesize these challenging expressions across the range of subjects, while preserving fine scale details such as wrinkles and hair, mouth and lip motion, and eye gaze direction, without introducing any significant artifacts. While NerFACE [1] is able to capture the general expression and gaze, it introduces artifacts for example in Subject 3 and produces blurry details on skin and hair due to the limitation of using a single global MLP to model the full appearance. IMAvatar [11] and NHA [2] struggle with capturing volumetric effects in the hair and out-of-model objects such as glasses due to the underlying surface based geometry representation and result in artifacts along the boundaries. FOMM [9] fails to produce these challenging expressions due to it's inherent 2D representation.

## D.2. Design Ablation Analysis

In Fig. F we visualize the close-up result produced by various design choice ablations of our method as detailed in Sec 4.4 of the main paper. These ablations show different ways of predicting per-vertex features on the 3DMM mesh which are spatially interpolated to obtain the volumetric radiance field of the avatar. “Static Features” learns fixed per-vertex features on the 3DMM mesh over the course of training. Since the features are not conditioned on the expression parameters, it struggles to properly model non-neutral expressions. “3DMM codes” concatenates the expression and pose codes to the static features. This results in reproducing the expression better but still results in local artifacts. “3DMM codes MLP” improves the model capacity by conditioning an MLP based VAE on the 3DMM codes that decodes to vertex features. While this improves the local artifacts, it still produces blurry result due to the global representation. “Ours-C” uses a convolutional decoder to produce UV space features from 3DMM codes. This significantly improves the level of high-frequency spatial details in the synthesized image. Finally, “Ours-D” poses the problem as an image-translation task in the UV space by using a convolutional encoder-decoder architecture to directly translate the geometry deformations of the 3DMM to UV space features. This generates local features that achieve the most faithful reconstruction of the expression along with better preserved spatial details.

## D.3. Demonstrating Robustness and Applicability

In Fig. G, we demonstrate the avatars being driven by the same subject at a different time and place than the original training sequence. Note that the subject's hair style, scene lighting, and accessories such as glasses are different. Our technique is able to faithfully reproduce the pose and expressions even under the novel conditions, demonstrating robustness and practical applicability. Please see the full sequence of this challenging avatar driving in novel conditions in the accompanying supplementary webpage.

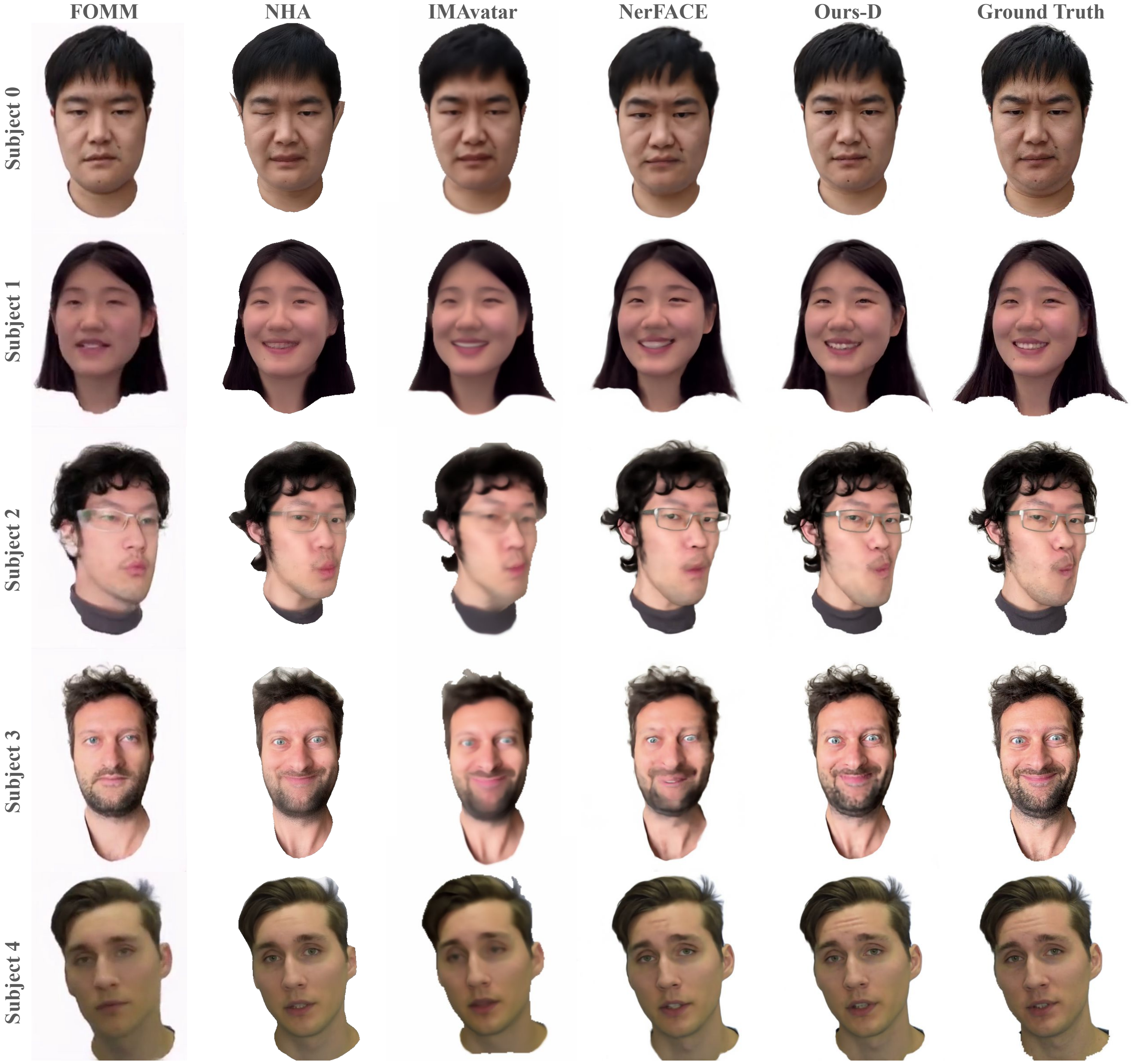

Figure E. Qualitative Comparison to prior state-of-the-art monocular head avatars. Note how our approach more faithfully reconstructs the ground truth expressions while preserving most of the high frequency details. Please refer to Sec. 4.2 in main paper for more discussions.

### D.4. Video Results

In the accompanying supplementary webpage, we demonstrate full-sequence results for following cases:

- Driving the avatar using a test clip that is captured in the same conditions as the training data (*i.e.*, same subject, same capturing condition).
- Driving the avatar by the same subject under novel conditions of lighting, appearance, and accessories (*i.e.*, the same subject under different capturing conditions).

To drive our avatar, we first obtain camera and 3DMM parameters from the driving video via per-frame 3DMM fitting, then apply these 3DMM parameters to our avatars and render from frontal or novel camera views. Note in the videos that our method produces high-quality controllable avatars that capture identity, pose, and expression specific idiosyncrasies. The avatar can be rendered in 3D from any desired viewpoint. Since the training data is captured only

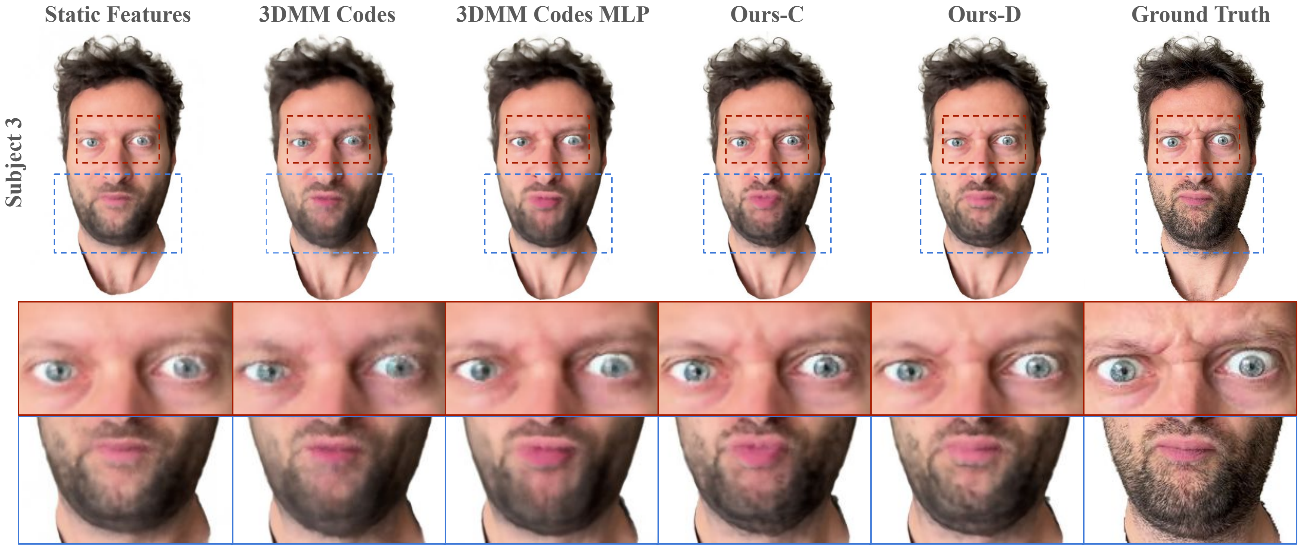

Figure F. Comparison between different designs for local vertex feature learning. See Sec. 4.4 in main paper for more details. “Static feature” struggles to capture personalized expressions. “3DMM Codes” improves the personalization but suffers from overall blurriness. “3DMM Codes MLP” further improves the sharpness, but still cannot present the details. Overall, our convolution-based methods lead to superior renderings on areas such as eyes, facial hairs, and frown wrinkles.

from frontal views, more extreme side views sometimes result in artifacts at the back of the head, which is an expected limitation of our method. Other common challenging cases are people with long hair and wearable. Our method is still able to generate plausible results though indeed shows relatively more artifacts.

Though small temporal jitters are also shown the videos, we observed that the jitters are significantly mitigated when the avatar is driven using synthetically smoothed 3DMM motions. This suggests that the jitters are mainly due to errors in 3DMM fitting. Improved 3DMMs and fitting algorithms in the future would resolve this issue. Future research could also explore the mitigation of temporal jitters from a neural rendering perspective.

### D.5. Visualize 3DMM and Final Geometry

We provide the visualizations and comparisons of the 3DMM mesh and the learned final geometry in Fig. H. Our method is able to reasonably capture the out-of-3DMM geometry such as glasses and hairs.

## References

[1] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. pages 8649–8658, 2021. 3, 4

[2] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural Head Avatars From Monocular RGB Videos. pages 18653–18664, 2022. 3, 4

[3] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The Relightables: Volumetric Performance Capture of Humans With Realistic Relighting. *ACM Transactions on Graphics*, Nov. 2019. 1

[4] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural Human Radiance Field From a Single Video. 2022. 1

[5] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a Model of Facial Shape and Expression From 4D Scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[6] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[7] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 1

[8] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. pages 5865–5874, 2021. 1

[9] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First Order Motion Model for

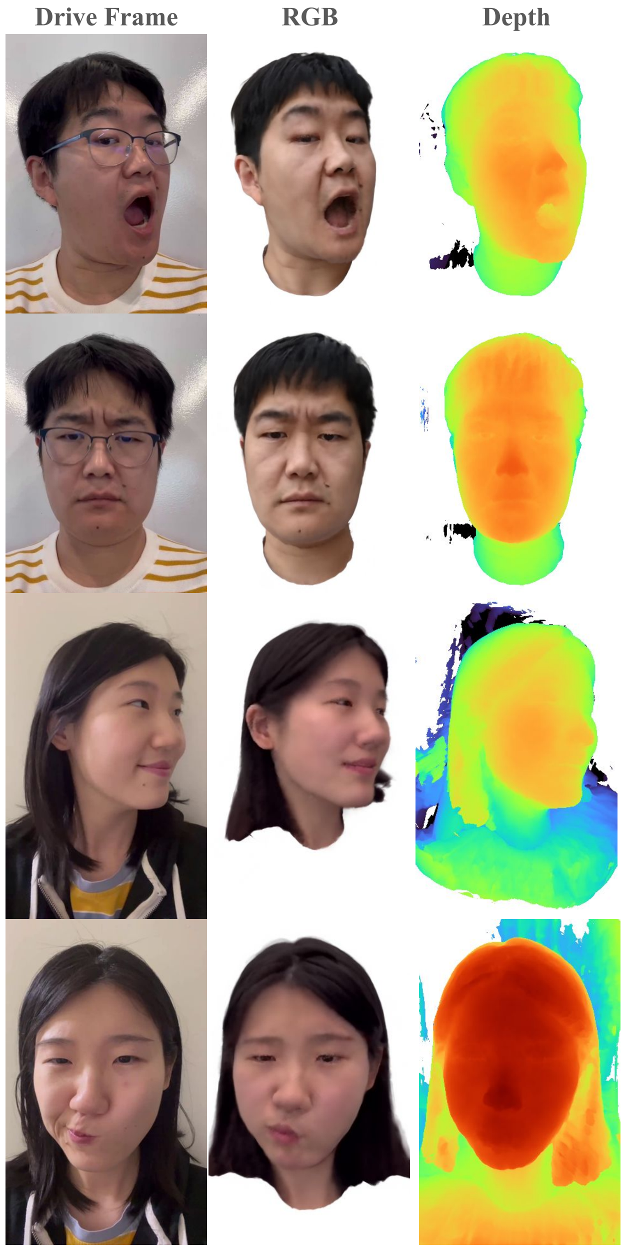

Figure G. Results on driving the learned avatar by the same subject under different capturing conditions. Our method produces faithful expressions and good geometry.

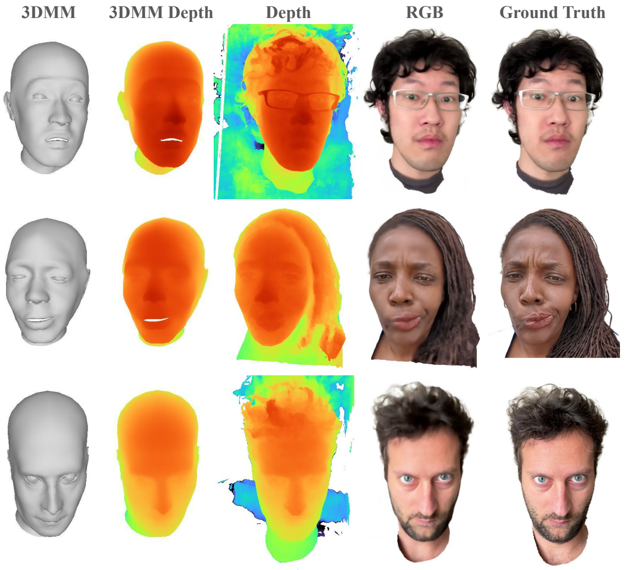

Figure H. Visualization of 3DMM and final geometry. Our method can reasonably capture out-of-3DMM geometry.

Image Animation. 32, 2019. 4

[10] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1

[11] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Imavatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 4