DPHMs: Diffusion Parametric Head Models for Depth-based Tracking – Supplementary Material –

Jiapeng Tang[1] Angela Dai[1] Yinyu Nie[1] Lev Markhasin[2] Justus Thies[3] Matthias Nießner[1]

[1] Technical University of Munich [2] Sony Semiconductor Solutions Europe
[3] Technical University of Darmstadt

In this supplementary material, we delve into additional details about the network architectures in Sec. 1. Subsequently, we elaborate on collecting the DPHM-Kinect dataset in Sec. 2. Following that, we provide a comprehensive explanation for the implementation of DPHMs for depth-based tracking in Sec. 3. Moving forward, we showcase the results of unconditional head generation in Sec. 4. Finally, we present supplementary comparisons against state-of-the-art head tracking methods in Sec. 5, detailed results of the robustness analysis in Sec. 6, along with some further discussions in Sec. 7.

1. Network Architectures

1.1. Modified NPHMs

In our DPHMs, we utilized a modified version of the Neural Parametric Head Models (NPHMs) [3, 4] to learn overparametrized latents from high-resolution head scans in the NPHMs dataset [3]. Specifically, we replaced the forward deformation network with the backward deformation network, enabling topology changes during expression tracking. The network architecture of the modified Neural Parametric Head Models is illustrated in Fig. 1.

We represent the human head geometry by a volumetric signed distance field decoded from two disentangled latents: the identity latent $\mathbf{z}^{id}$ and the expression latent $\mathbf{z}_{ex}$. The $\mathbf{z}^{id}$ is the concatenation of a global latent $\mathbf{z}^{id}_{glo}$ and local latents $\mathbf{z}^{id}_1, ..., \mathbf{z}^{id}_K$. $\mathbf{z}^{id}_{glo}$ is used to estimate $K = 39$ pre-defined anchor positions on a human head through a small $\mathrm{MLP}_{\mathrm{anc}}$. Each local identity latent $\mathbf{z}^{id}_k$ is used to describe the head geometries around the k_{th} anchor. To be specific, we define $K = 2K_{\mathrm{sym}} + K_{\mathrm{usym}}$ facial anchors, denoted as $\mathbf{a} \in \mathbb{R}^{K\times 3}$. K_{sym} anchors are on the left face, mirrored to form the other K_{sym} anchors. K_{usym} anchors are positioned in the middle of the face, shared by both the left and right sides. To predict the SDF value of a point $\mathbf{p} \in \mathbb{R}^3$ within the expression space, we concatenate it with the identity latent $\mathbf{z}^{id}$ and expression latent $\mathbf{z}^{ex}$. The resulting feature is then passed through the backward deformation decoder F_{ex}, which warps $\mathbf{p}$ to $\mathbf{p}' \in \mathbb{R}^5$ in the canonical space. It is important to note that $\mathbf{p}'$ includes two hyper-dimensions [8] to model topological changes under different expressions, as continuous deformation fields alone may struggle with such changes. Then, we feed $\mathbf{z}^{id}$ and $\mathbf{p}'$ into the canonical identity decoder F_{id} to obtain its SDF prediction. The F_{id} is implemented using an ensemble of smaller local Multi-Layer Perceptron (MLP)-based networks, each responsible for a local region centered around an anchor. For the k_{th} local region of facial anchor, we feed the corresponding local latent vector $\mathbf{z^{id}_k}$, along with the global latent vector $\mathbf{z^{id}_{glo}}$ into an SDF decoder MLP_{θ_k} with learnable weights θ_k:

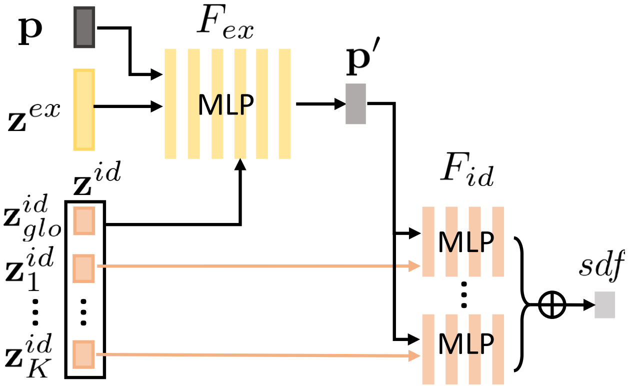

Figure 1. The network architecture of **our modified Neural Parametric Head Models** based on backward deformations.

$$f_k(\mathbf{p}, \mathbf{z^{id}_k}, \mathbf{z^{id}_{glo}}) = \mathrm{MLP}_{\theta_k}([\mathbf{p} - \mathbf{a}_k, \mathbf{z^{id}_k}, \mathbf{z^{glo}}]). \quad (1)$$

To exploit facial symmetry, we share the network parameters and mirror the coordinates for each pair (k, k^*) of symmetric regions:

$$f_{k^*}(\mathbf{p}, \mathbf{z^{id}_{k^*}}, \mathbf{z^{id}_{glo}}) = \mathrm{MLP}_{\theta_k}([\mathrm{flip}(\mathbf{p}-\mathbf{a}_k), \mathbf{z^{id}_k}, \mathbf{z^{glo}}]). \quad (2)$$

Finally, we can composite all local fields f_k into a global field:

$$F_{id}(\mathbf{p}) = \sum_{k=1}^{K} \omega_k(\mathbf{p}, \mathbf{a_k}) f_k(\mathbf{p}, \mathbf{z_k^{id}}, \mathbf{z_{glo}^{id}}). \tag{3}$$

The blending weights are calculated by a Gaussian kernel based on the Euclidean distance between the query point $\mathbf{p}$ and $\mathbf{a_k}$.

$$\omega_k^{'}(\mathbf{p}, \mathbf{a_k}) = \begin{cases} \exp\left\{\frac{-\|\mathbf{p}-\mathbf{a_k}\|_2}{2\sigma}\right\}, & k > 0 \\ c, & k = 0 \end{cases}$$

$$\omega_k(\mathbf{p}, \mathbf{a_k}) = \frac{\omega_k^{'}(\mathbf{p}, \mathbf{a_k})}{\sum_{k'} \omega_k^{'}(\mathbf{p}, \mathbf{a_k})} \tag{4}$$

We use a fixed isotropic kernel with standard deviation σ and a constant response c for f_0. The global identity latent $\mathbf{z_{glo}^{id}}$ has the dimension $d_{glo} = 64$, and each local identity latent $\mathbf{z_k^{id}}$ is of dimension $d_{loc} = 32$. Therefore, the total dimension of $\mathbf{z^{id}}$ is $(K+1) * d_{loc} + d_{glo} = 1344$. The backward deformation decoder F_{ex} is implemented by a six-layer MLP with a hidden dimension of 512. Each MLP_{θ_k} for local SDF field prediction has four layers with a hidden size of 200. The anchor prediction MLP has three layers with a hidden size of 128. To blend the ensemble of local SDF fields, we use $\sigma = 0.1$ and $c = \exp^{-0.2/\sigma^2}$.

1.2. Denoiser networks.

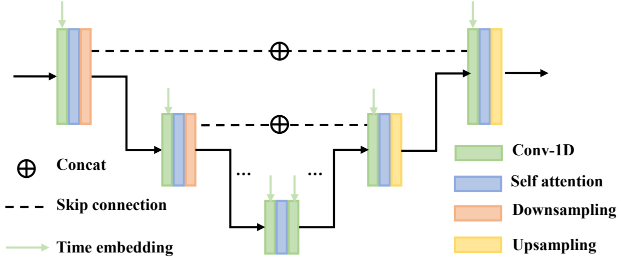

Figure 2. The denoiser network of our identity and expression latent diffusion models.

The identity and expression diffusion models are constructed using UNet-1D [9] architecture with incorporated attention layers [10], following DDPM [5]. We analogously treat the identity and expression latents as sequences of 1D scalar features, with the only difference being the sequence length, which is equal to the latent feature dimension. Fig. 2 illustrates the detailed network architecture, which has an encoder of 4 downsampling blocks and a decoder of 4 upsampling blocks. The encoder progressively increases the feature dimension from 1 to 64, 128, 256, and 512, while simultaneously reducing the sequence length by 2. The decoder follows the opposite pattern, reducing the feature dimension and doubling the sequence length.

2. DPHM-Kinect dataset

The monocular RGBD sequences of our DPHM-Kinect dataset are collected from different skin tones and ethnicities. After obtaining consent from each attendee, we recorded five types of head motion sequences: ' smile and laugh,' 'eyeblinks,' 'fast-talking,' 'random facial expressions,' and 'mouth movements.' The recording framerate is 16fps. Each motion sequence lasts 6-10s, thus containing 96-160 frames. During data capture, they sit in front of the Kinect Azure sensor at a 15-40cm distance. The examples of these motion types from different attendees are depicted in Fig. 3.

3. Implementation Details

Proprocessing. We begin by eliminating background pixels from depth maps using a threshold of $d_{\text{max}} = 0.6m$. Subsequently, bilateral smoothing is applied to the depth maps. Following this, surface normals are computed using the cross-product of derivatives along the x and y directions. Next, the depth and normal maps are lifted into 3D space, resulting in oriented partial point clouds. Finally, we crop out the points within the head region as the input.

Rigid registration. Prior to non-rigid tracking, we need to obtain the rigid transformation parameters that convert the provided scan from the camera coordinate system to that of DPHMs. Since the coordinate system of DPHMs aligns with the FLAME space [7], we opt to perform FLAME fitting. This includes the optimization of identity, expression, and pose parameters, as well as scale, rotation, and translation parameters.

Non-rigid tracking. With the rigid alignment serving as initialization, our method optimizes the identity and expression latents for depth-based head tracking, simultaneously fine-tuning the rigid parameters. Our non-rigid tracking comprises three phases: 'identity fitting', 'expression fitting', and 'joint finetuning'. In the first phase of identity fitting, we optimize the identity latent $\mathbf{z}^{id}$ and expression latent $\mathbf{z}_1^{ex}$ for the first frame. In the expression fitting, we fix the identity latent and optimize the expression latent frame by frame. The optimization result $\mathbf{z}_{i-1}^{ex}$ of the last frame is used as initialization for the expression latent of the next frame $\mathbf{z}_i^{ex}$. In the joint fine-tuning stage, we finetune the identity latent $\mathbf{z}^{id}$, all expression latents $\mathbf{z}_{1:N}^{ex}$, as well as the rigid transformation parameters for better alignments.

4. Unconditional Generation of DPHMs

4.1. Identity Generation

We present the randomly sampled results of our unconditional identity diffusion in Figure 4. Our approach demonstrates the ability to generate high-quality head avatars with diverse hairstyles.

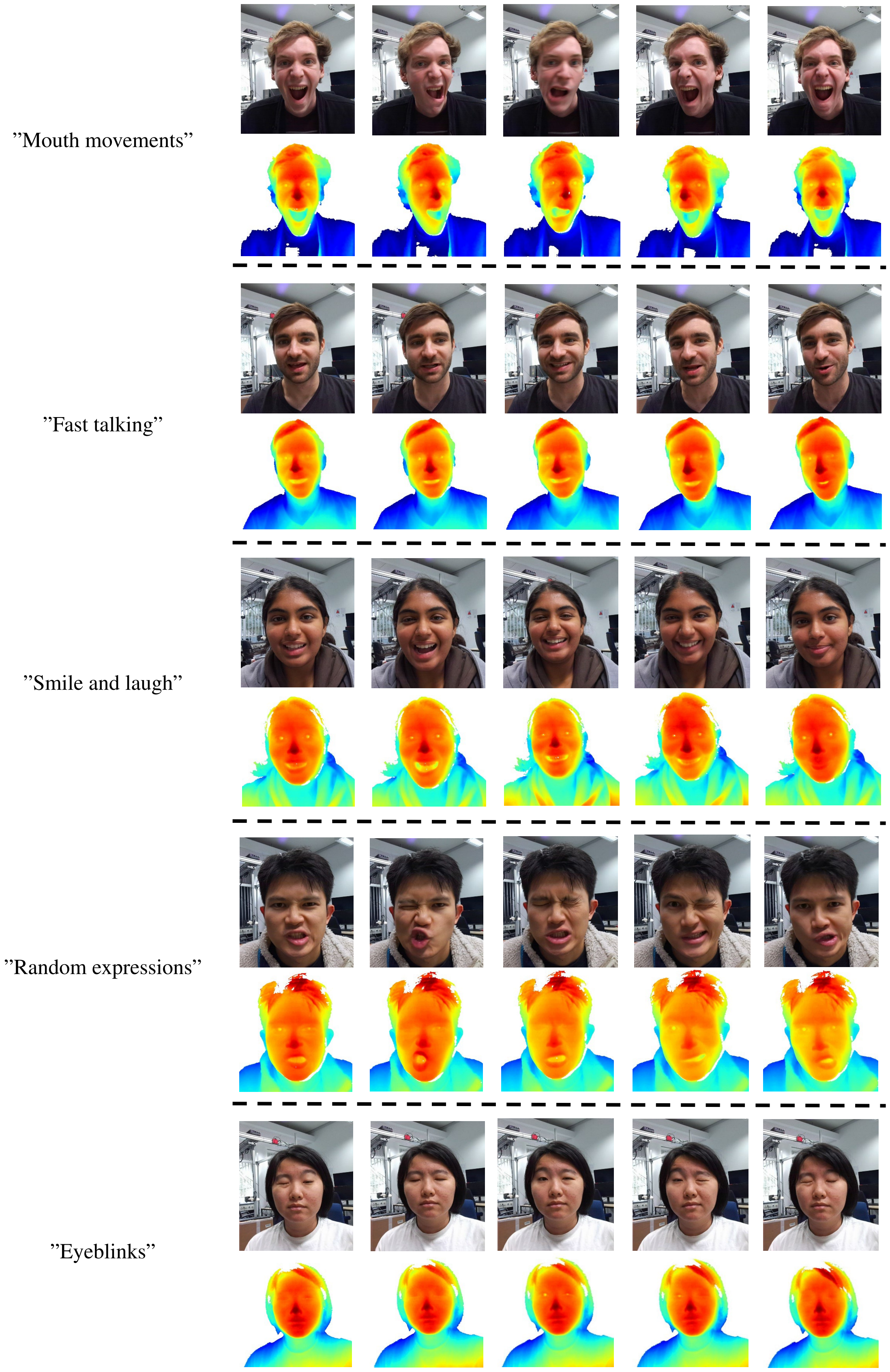

Figure 3. Example sequences of our **DPHM-Kinect dataset**.

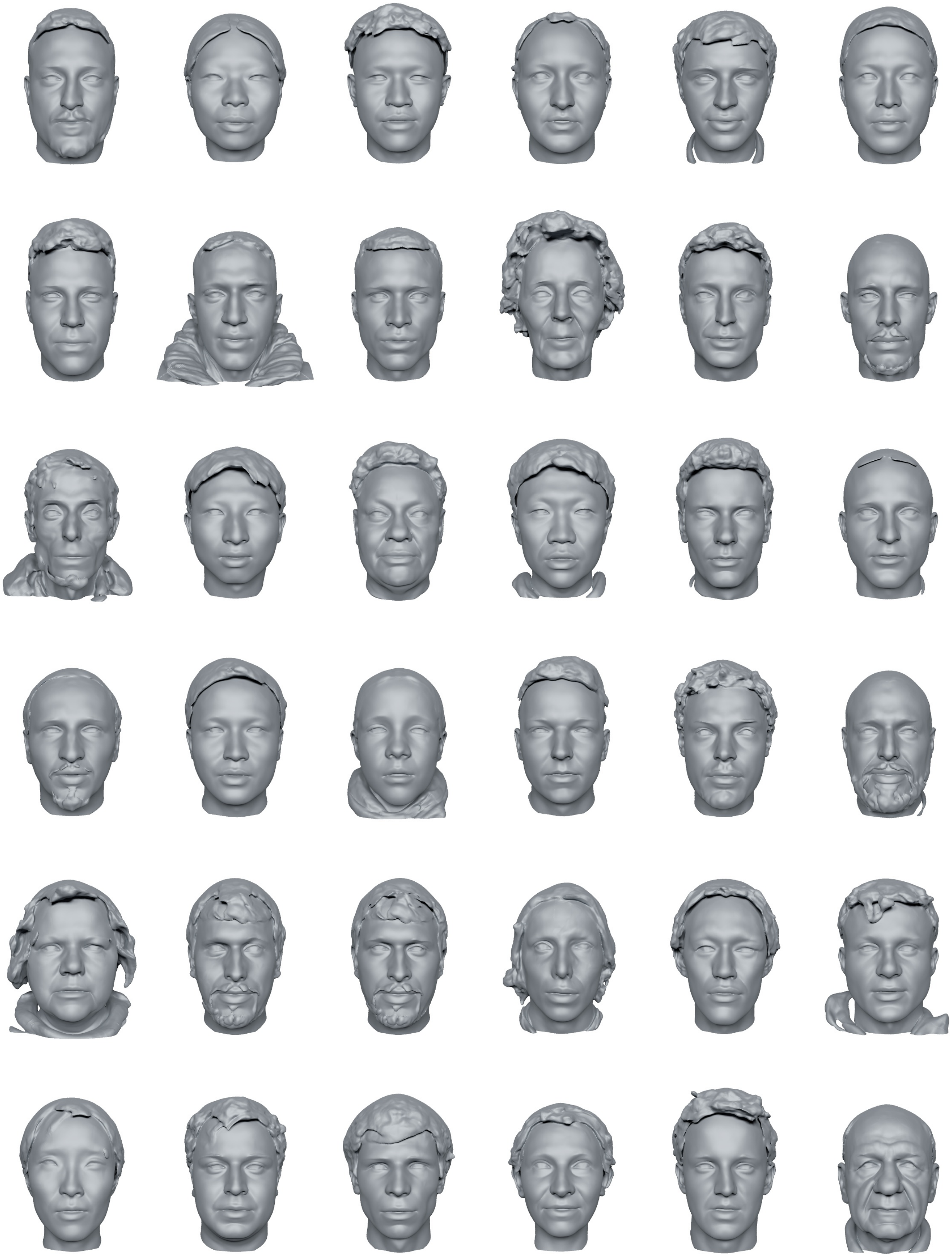

Figure 4. Unconditional sampling results of identity parametric diffusion model. Our approach can generate high-quality head avatars with diverse hairstyles.

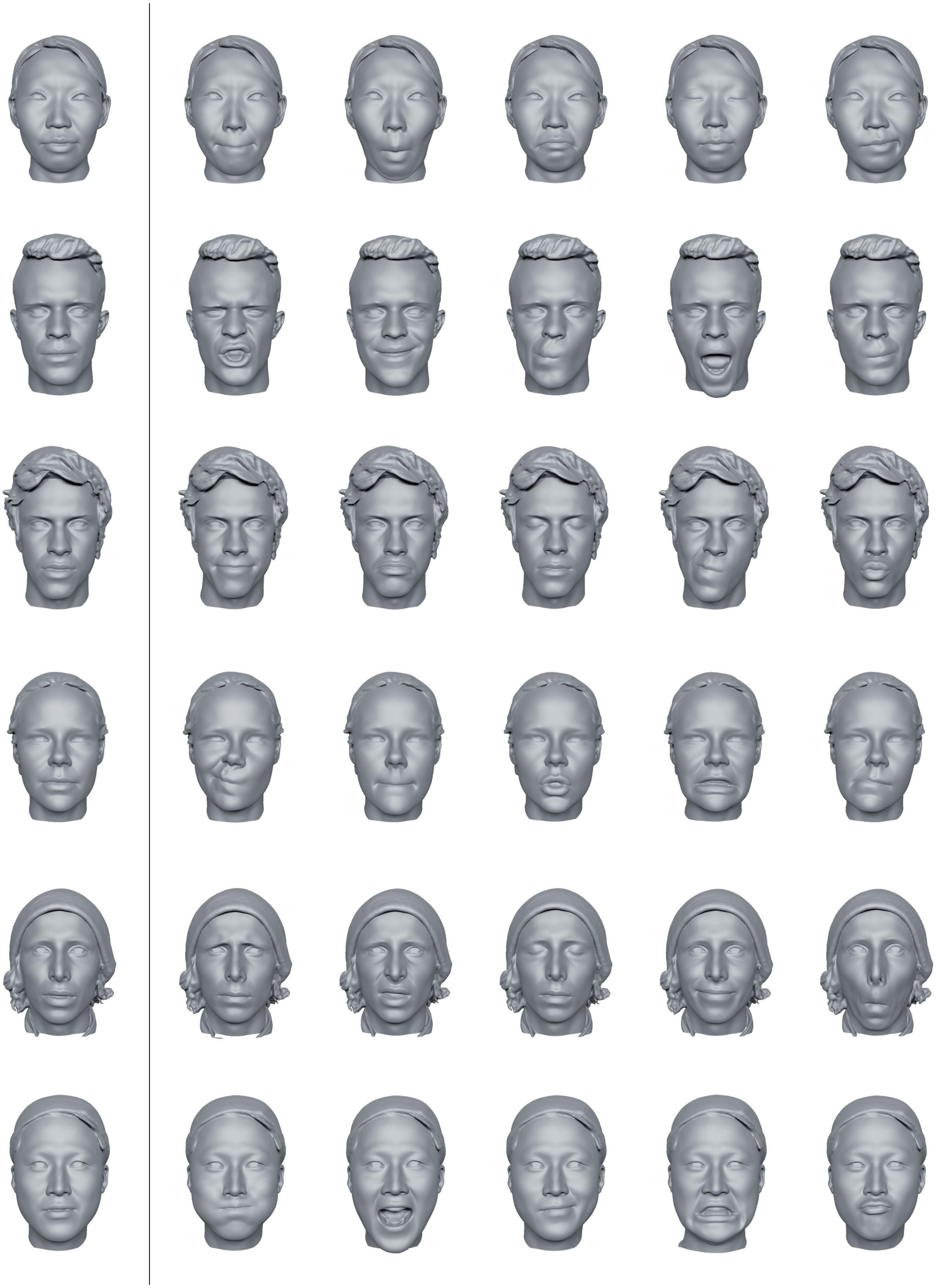

Figure 5. Unconditional sampling results of expression parametric diffusion model. The first column presents the canonical identity geometry with the neutral expression, *i.e.*, zero expression latent vector. Our method can generate a variety of plausible complex expressions.

4.2. Expression Generation

In Figure 5, we present randomly sampled expressions for given head identities under the neutral expression with a closed mouth. Our expression parametric latent diffusion demonstrates the capability to generate various complicated facial expressions.

5. Additional Comparisons

5.1. Additional comparisons on the DPHM-Kinect dataset

In Figure 6, we visualize additional comparisons on the monocular depth sequences from our DPHM-Kinect benchmark.

5.2. Additional comparisons on the Multi-view Video dataset

In Figure 7, we provide more qualitative comparisons on the single-view depth sequences reconstructed from the multi-view videos of NerSemble [6].

5.3. Evaluations on Unobserved Regions

During evaluations, we use partial depth scans used for test-time optimization as the target to calculate metrics. For DPHM-Kinect sequences, we do not have the ground truth of dynamic head scans. Thus, we can only use the single-view Kinect depth scans for evaluation. To better evaluate the reconstruction of unobserved regions, we conduct additional comparisons on the NerSemble [6] dataset by only using single-view depth videos as input during optimization, while using more complete scans from multi-view depths as targets during evaluation. In Tab. 1, our method consistently outperforms existing methods in all metrics, which further confirms the effectiveness of our approach in reconstructing more accurate unobserved geometries.

Method	FLAME	ImFace	ImFace*	ImAvatar	NPHM	Ours
ℓ_2	2.947	9.471	3.065	3.255	1.024	**0.894**
NC	69.50	78.74	84.36	78.13	88.07	**89.02**
RC	24.21	10.79	31.86	16.23	85.88	**91.02**
RC2	55.15	22.11	63.29	56.63	97.10	**97.99**

Table 1. Head tracking reconstructed from single-view depth scans. The results are evaluated on multi-view depth scans.

5.4. Comparison with Template-based Non-rigid Registration Method.

Additionally, we include a classic template-based non-rigid registration method, NICP [1], into our comparisons. We re-implement NICP using a template mesh from FLAME [7] to obtain mesh deformations that align with depth scans. We evaluate against NICP using the DPHM-Kinect dataset, which contains more challenging expressions compared to VOCA [2]. As illustrated in Fig. 8, NICP cannot recover plausible identities with hair and correct expressions because it does not have effective priors to handle imperfect observations of noisy and partial scans. It also cannot perform expression transfer, without the disentanglement of identity and expression. The quantitative comparisons presented in Tab. 2 demonstrate the superiority of our approach again.

Metric	$\ell_2 \downarrow$	NC ↑	RC ↑	RC2 ↑
NICP	3.926	81.47	32.32	70.50
Ours	**1.465**	**86.80**	**70.79**	**90.98**

Table 2. Quantitative comparisons against NICP on the DPHM-Kinect dataset.

6. Robustness Analysis

In Figure 9 and 10, we provide qualitative comparisons against NPHMs on imperfect observations with different noise and sparsity levels. Detailed quantitative results from our robustness analysis are summarized in Table 3 and Table 4. The results illustrate the superior robustness of DPHMs compared to NPHMs across a range of imperfect observations.

Method	FLAME	ImFace	ImFace*	ImAvatar	NPHM	Ours
0 mm	21.01	40.71	43.71	14.84	84.49	**89.87**
0.5mm	21.06	40.61	43.65	14.01	82.93	**88.41**
1mm	21.05	40.57	43.42	14.35	79.67	**86.12**
2mm	21.08	39.86	42.64	13.85	69.29	**79.41**

Table 3. Quantitative results at different noise levels of the input point cloud at each frame.

Method	FLAME	ImFace	ImFace*	ImAvatar	NPHM	Ours
10,000	21.01	40.74	43.69	14.96	84.25	**89.62**
5,000	21.06	40.71	43.72	14.84	84.49	**89.87**
2,500	21.05	40.68	43.70	14.53	83.55	**89.65**
250	21.08	40.30	42.86	14.37	81.78	**85.31**

Table 4. Quantitative results at different sparsity levels of the input point cloud at each frame.

7. Discussions

Hair reconstruction. When hair is sufficiently observed, reconstructed hair can be aligned with input (Fig. 1 and second example of Fig. 4 in the main paper). When given

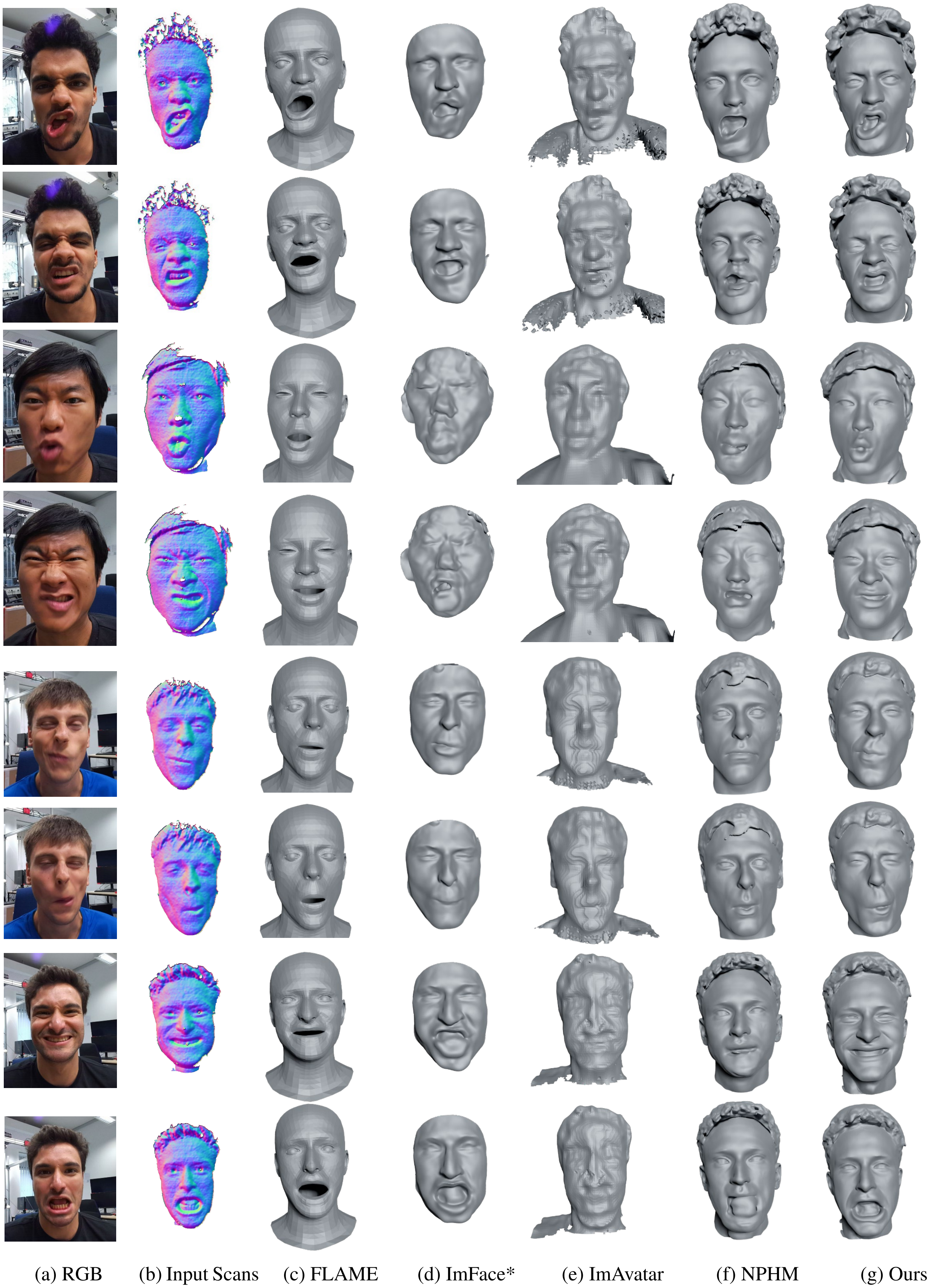

Figure 6. Head Tracking on the DPHM-Kinect dataset. Note that RGB images are only used for reference not used by all the methods except ImAvatar. Compared to state-of-the-art methods, our approach achieves more accurate identity reconstruction with detailed hair geometries while tracking more plausible expressions, even during extreme mouth movements.

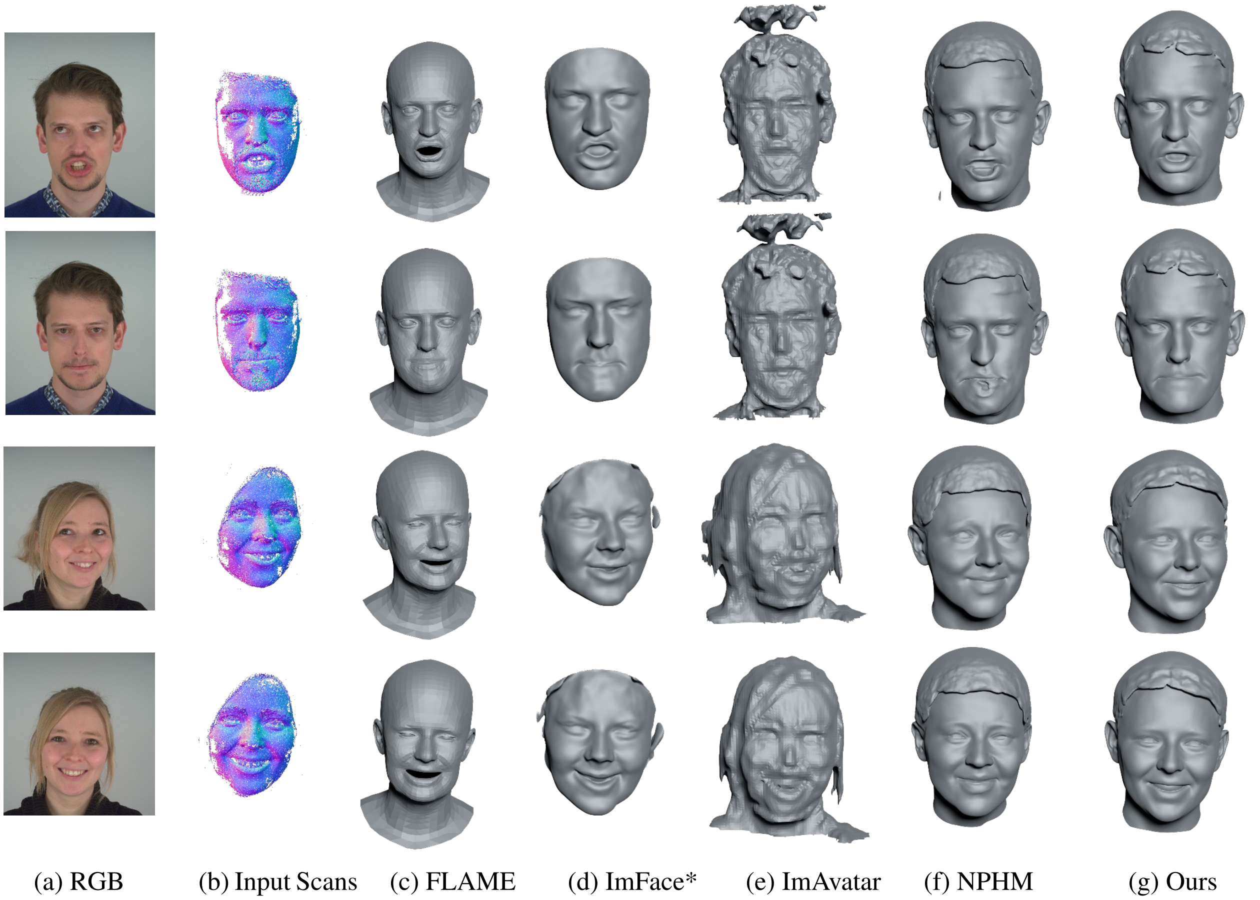

Figure 7. Head Reconstruction and Tracking on the single-view depth sequences of NerSemble [6]. Note that RGB images are only used for reference and not used by all methods except ImAvatar. Compared to state-of-the-art methods, our approach demonstrates the ability to reconstruct realistic head avatars with hairs and accurately capture intricate facial expressions.

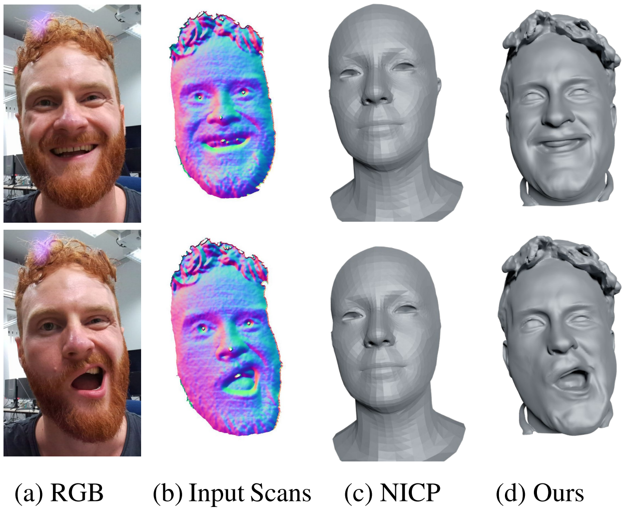

Figure 8. Qualitative comparisons against NICP on the DPHM-Kinect dataset.

a few hair measurements in the depth scans (first example of Fig. 4 in the main paper), we can still output plausible results compared to baselines. We believe that incorporating RGB images with depth scans as inputs could further improve the results.

Arbitrary Length of Sequences. Our method formulates head tracking as an optimization problem. It directly optimizes the identity and expression parametric latent. This eliminates the need for an encoder to process input depth scans. As mentioned in Sec. 3, our non-rigid tracking consists of three stages: identity fitting using the first frame, frame-by-frame expression fitting, and joint fine-tuning of rigid and non-rigid registration parameters. In the final stage, if the sequence length is short, we can jointly fine-tune the parameters of all frames. However, for long sequences, we can improve temporal smoothness by randomly sampling fragments, with each fragment comprising three consecutive frames. This strategy effectively mitigates memory consumption issues. Therefore, our method is capable of handling depth sequences of arbitrary length.

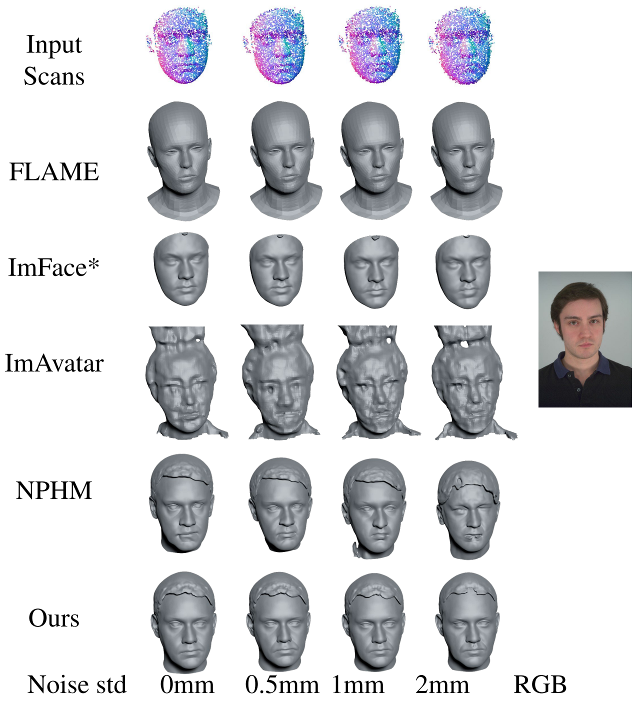

Figure 9. Qualitative comparisons of NPHM and our method with respect to noise in the input scan. We perturb the input scans by additive Gaussian noise with different standard deviations.

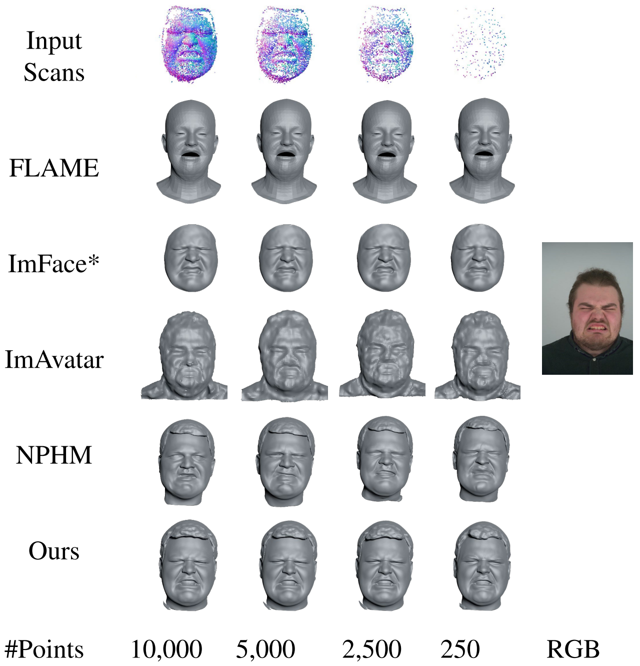

Figure 10. Qualitative comparisons of NPHM and our method with respect to the number of points in the input scan.

Non-marginal quantitative improvements. In Fig. 11, We provide comparisons between NPHM and our method in terms of reconstructed meshes and error maps derived from Scan2Mesh distances, along with the calculated metrics. Notably, 0.263 mm lower ℓ_2 error means significant improvements in reconstructing facial wrinkles and mouth regions, also with 9.63% higher Precision@1.5mm score.

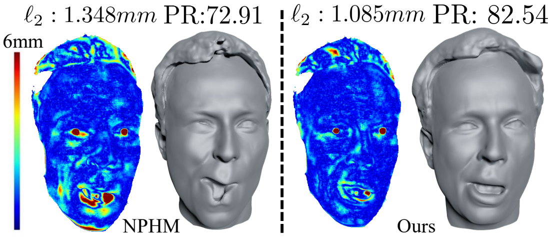

Figure 11. Comparisons between NPHM and our method in reconstructed meshes and error maps from Scan2Mesh distances, along with the calculated metrics.

Expression Transfer. As our approach disentangles identity and expression via two separate latents, it can be applied to expression transfer applications. In Fig. 12, we demonstrate the transfer of our reconstructed expressions to a different person. Given a monocular sequence of depth scans as inputs, we initially obtain the identity reconstruction and track expression transitions by our method. Subsequently, we animate the source identity using the reconstructed expression latent. The transfer result faithfully represents the intricate facial expressions without introducing personalized geometry details such as hairstyle.

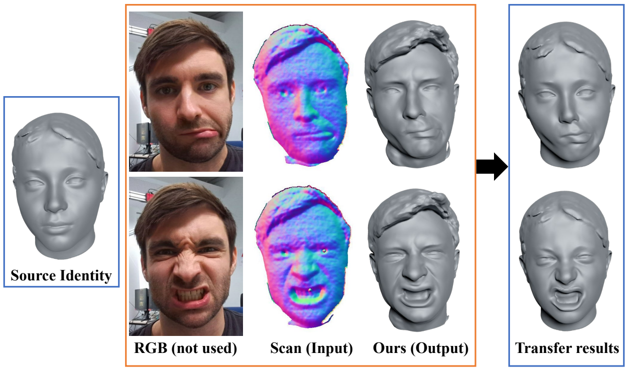

Figure 12. **Expression Transfer**. Given a monocular depth sequence of a head avatar, we first use our tracking method to obtain the identity and expression latent. Then we transfer the reconstructed expression to the source identity using backward deformation fields, which are conditioned on both the source identity latent and the target expression latent.

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 6

[2] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 6

[3] 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. 1

[4] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[6] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *arXiv preprint arXiv:2305.03027*, 2023. 6, 8

[7] 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. 2, 6

[8] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 1

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

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2