SUPPLEMENTARY MATERIAL

LiveHand: Real-time and Photorealistic Neural Hand Rendering

Akshay Mundra [1,2], Mallikarjun B R [1], Jiayi Wang [1],
Marc Habermann [1], Christian Theobalt [1,2], Mohamed Elgharib [1]
[1] Max Planck Institute for Informatics [2] Saarland University

In this supplementary document, we first provide the implementation details, followed by a comparison with the state-of-the-art methods on a synthetic dataset. Then, we demonstrate the robustness of our method to MANO fitting inaccuracies. Finally, we show an additional application where the hand geometry can be edited at inference time without any additional retraining.

1. Implementation Details

We use the same positional encoder as [3], with a maximum frequency of $L = 10$ for the canonicalized sampling point (u, v, h) and $L = 4$ for the viewing direction d. Our network is parameterized with a 6 layer-deep MLP as shown in Fig. 1. We use a similar CNN network architecture as in EG3D [1] for our super-resolution network, with an upsampling factor of 2. We empirically choose 16 as the number of samples to draw per ray as it best trades off image quality and rendering speed. Our radiance field module H_α, super-resolution module S_ϕ, and color calibration parameters g_j, b_j are learnt with learning rates of 0.0025, 0.0025, and 0.0001 respectively using Adam optimizer [2] with a decay rate of 0.1. All models are trained for 200K iterations.

To run SMPLpix, we assign gradually varying color values to the posed MANO mesh and render it from the camera view. This colored image is used as an input to the image translation network, which predicts the actual hand image. These input-output pairs are visualized in Figure 2.

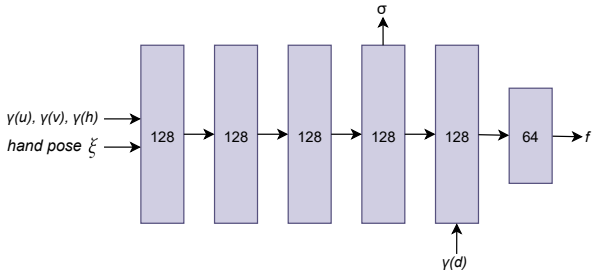

Figure 1. **MLP architecture of our model.** The uvh mapping simplifies learning the radiance field, allowing us to reduce the network size compared to Mildenhall *et al*. [3].

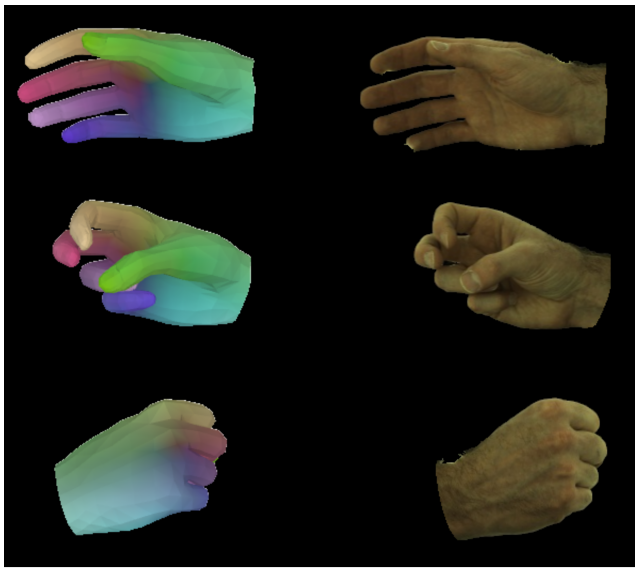

Figure 2. **Data for running SMPLpix.** We use renderings of colored MANO meshed as input to the SMPLpix model to predict the ground truth hand image.

For the naive pose-conditioning ablation ('xyz' in Table 3 of the main paper), we use the original MLP architecture described in [3], while also concatenating hand pose to the positionally encoded input. For the sampling ablation ('w.o. mesh-guided samp.' in Table 4 of the main paper), we first draw 64 stratified samples in the 3D bounding box surrounding the hand, and then use 16 importance samples.

2. Comparison on Simulated Dataset

For additional benchmarking, we generate a synthetic dataset by applying one of the ground truth texture maps from HTML [4] on the InterHand2.6M MANO meshes and

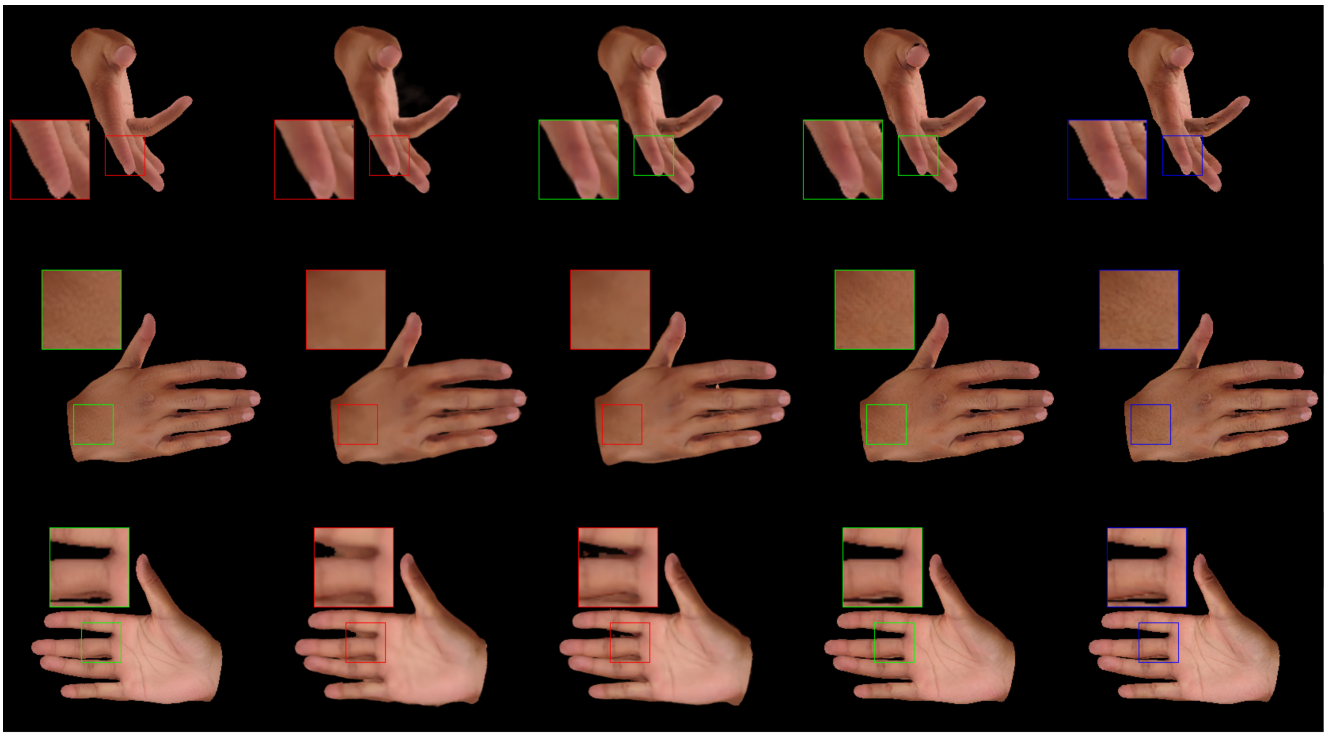

Figure 3. **Comparison on our synthetic dataset.** Our method significantly outperforms the other two volumetric approaches (i.e. A-NeRF and LISA) which struggle in getting the geometry and the appearance right. While SMPLpix performs well on this synthetic dataset because of the availability of perfect annotations, we can still observe some ringing artifacts, as shown in the first row.

render it with the provided hand poses and camera parameters. This dataset is then used to train our method, as well as the other state-of-the-art methods. The results are presented in Tab. 1 and Fig. 3.

It is evident that our method outperforms the other two volumetric approaches (i.e. A-NeRF and LISA). On the other hand, SMPLpix performs better than our method on this simulated dataset. This is expected because this synthetic setting of perfect MANO annotations reduces SMPLpix's task to a simple one-to-one 2D mapping. But in real datasets, MANO fitting is challenging and it also does not allow capturing identity-specific details in the mesh because of limited PCA space. As a result, input to the neural renderer and ground truth will have pixel-level misalignments. This results in poor multi-view consistency and generalization, which can also be observed in Tab. 2 and Fig. 5 in the main paper. Moreover, this simulated data does not contain pose and view-dependent effects, which can be modeled by our method but not SMPLpix. All of these factors help SMPLpix achieve better results for the synthetic dataset.

Also, note that we do not show a comparison against 'Mesh wrapping' here as a similar technique is used in the first place to generate the synthetic dataset.

	PSNR ↑	LPIPS(x1000) ↓	FID ↓	FPS ↑
SMPLpix	36.66	6.84	23.95	58.82
A-NeRF*	29.47	38.61	68.15	0.83
LISA*	30.95	32.17	63.07	3.70
Ours	32.59	9.36	26.03	45.45

Table 1. **Comparison on our synthetic dataset.**

3. Robustness to MANO inaccuracies

Our method does not require perfect MANO fitting and is robust to some misalignment. The ground truth meshes provided in InterHand2.6M dataset are in fact noisy as the authors report a 5mm average error between the MANO joints and the annotated 3D keypoints. For a more thorough analysis, we introduce pose noise in the accurate MANO fittings of our simulated dataset and report the findings in Table 2. It is evident that our approach can handle tracking errors in the MANO fittings during training. The same can also be observed in Figure 4. Multiple factors in our design lead to this increased robustness: 1) our volume rendering does not rely on exact surface tracking and we only leverage the mesh for sampling and space canonicalization; 2) the perceptual loss does not penalize per-pixel mismatch in

Avg. mesh error	PSNR ↑	LPIPS (x1000) ↓	FID ↓
0 mm	32.59	9.36	26.03
2.99 mm	30.53	14.41	35.04
6.64 mm	29.97	18.24	41.98
10.63 mm	29.99	21.85	56.68

Table 2. **Impact of MANO fittings inaccuracies on reconstruction.** We observe that an average mesh error as high as 10.63 mm does not deteriorate the results significantly, as also observed in the qualitative results in Fig. 4.

the image space, thus avoiding blurred results common in case of poor tracking.

4. Application: Shape Editing

Our uvd encoding and the mesh-guided sampling formulation are not only advantageous in terms of rendering speed and quality, but they also enable easy editing of the hand-avatar geometry. Given the original hand parameter $\psi_{\text{init}} : \{\theta, \beta_{init}, t, R\}$, we can modify the shape parameter to obtain $\psi_{\text{new}} : \{\theta, \beta_{new}, t, R\}$. By using the corresponding mesh $\mathcal{M}(\psi_{\text{new}})$ in the canonicalization procedure, the rendered hand appearance will change accordingly. This allows the geometry of the hand avatar to be modified without retraining. We show the results of this application in Fig. 5, where we modified the first principal component of the MANO shape parameter β.

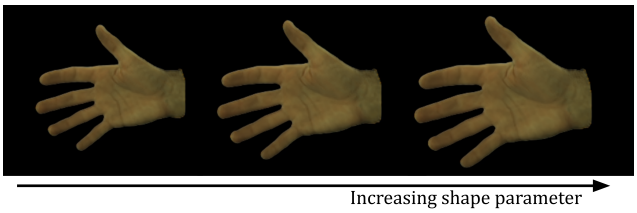

Figure 5. **Application: Shape Editing.** The hand geometry can be edited without any additional retraining of the model.

References

[1] 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. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] 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

[4] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision, (ECCV)*, pages 54–71. Springer, 2020. 1

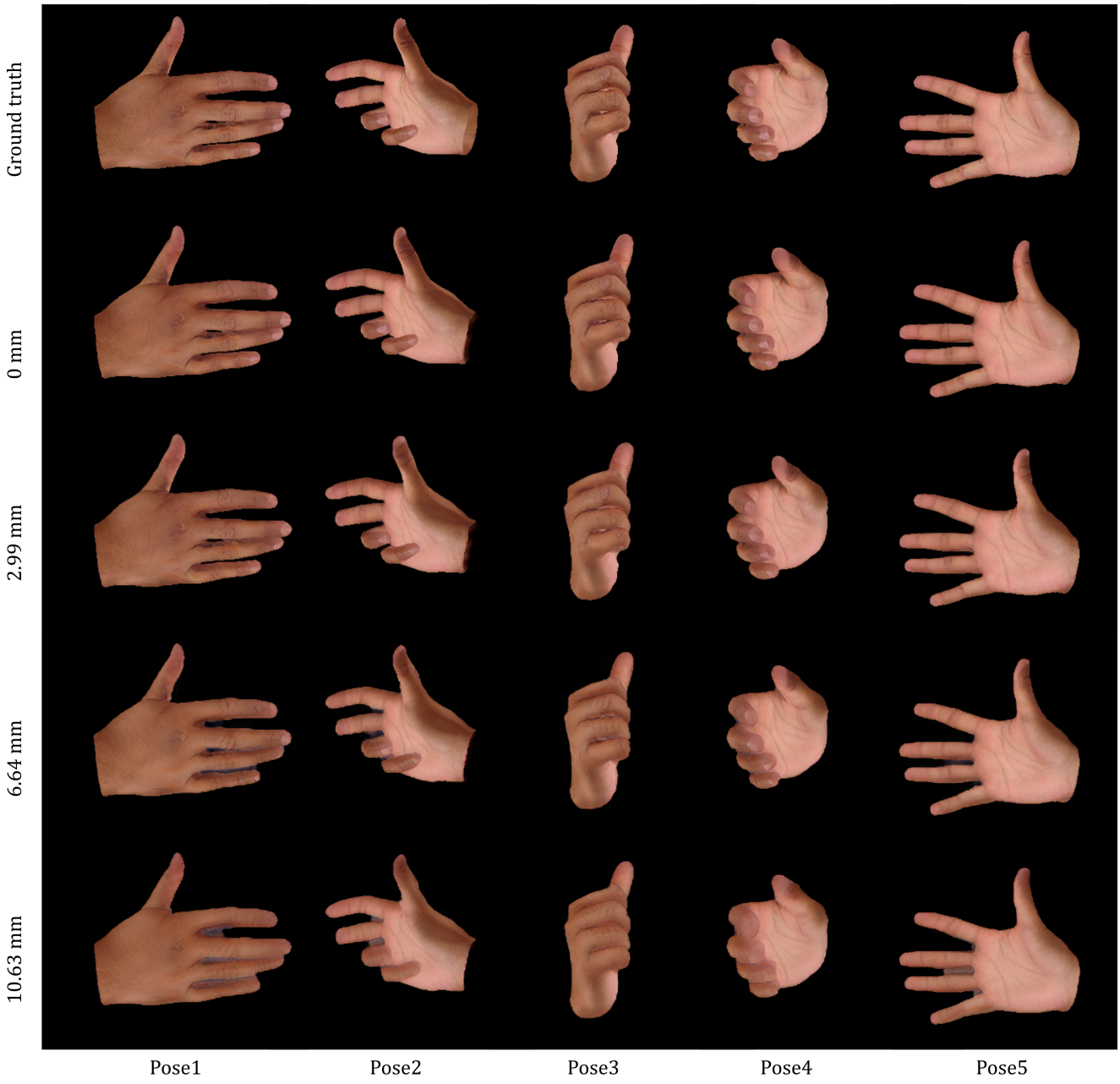

Figure 4. **Robustness to MANO fitting errors.** We analyze the impact of inaccurate MANO fittings on our method. Our use of perceptual loss, as well as mesh-based sampling and canonicalization strategies make the method robust to inaccurate MANO meshes and preserves the details in the learned model.