HandNeRF: Neural Radiance Fields for Animatable Interacting Hands (Supplementary Material)

Zhiyang Guo[1] Wengang Zhou[1,2*] Min Wang[2] Li Li[1] Houqiang Li[1,2*]
[1]CAS Key Laboratory of Technology in GIPAS, EEIS Department,
University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
guozhiyang@mail.ustc.edu.cn, {zhwg, lil1, lihq}@ustc.edu.cn, wangmin@iai.ustc.edu.cn

1. Implementation Details

1.1. Model

HandNeRF is built mainly with multi-layer perceptrons (MLP). The density and color fields of the canonical NeRF are composed of linear layers with Softplus and Softmax activations, respectively, and both include a skip connection in the intermediate layer. The error-correction network in the deformation field has a similar residual architecture with ReLU as the activation. The whole model is trained in an end-to-end manner using the Adam optimizer. The learning rate is initialized to be 0.0005 and gradually decays in an exponential form.

1.2. Dataset

HandNeRF is trained on the 30FPS version of InterHand2.6M [1]. This dataset does not provide the split setup for novel view/pose synthesis tasks, nor does any previous work proffer an applicable setting. Therefore, taking the view and pose distribution into consideration, we propose a reasonable setting as interpreted below.

Camera views. Totally, around 80 different views are available in the dataset, with around 46 of them captured by monochrome cameras. Among the 34 RGB views, we remove 6 ones that use landscape orientation instead of portrait, and further filter out 13 views where the target hands are heavily obscured by the photographic equipment. We then select the training and testing camera views within the remaining ones. The model is trained on a single sequence with 4, 7, or 10 views to show the effect of the number of available camera views. For all sequences, 18 common views are selected as the test views. The detailed settings are listed as follows:

10 training views:
1. 400002
2. 400004
3. 400010
4. 400012
5. 400013
6. 400016
7. 400018
8. 400042
9. 400053
10. 400059

7 training views:
1. 400002
2. 400004
3. 400013
4. 400016
5. 400018
6. 400042
7. 400059

4 training views:
1. 400004
2. 400016
3. 400018
4. 400042

18 testing views:
1. 400009
2. 400017
3. 400019
4. 400023
5. 400026
6. 400027
7. 400028
8. 400029
9. 400030
10. 400031
11. 400037
12. 400039
13. 400041
14. 400048
15. 400051
16. 400060
17. 400063
18. 400064

Training pose sequences. Each quantitative result reported in our paper is an average of all results trained and tested on 10 (for single hand) or 5 (for interacting hands) sequences randomly chosen from the large-scale dataset. Each sequence has 55 frames on average (1.8 seconds at 30FPS). The detailed sequence names are listed as follows:

Single Hand:
1. 0004_star_trek
2. 0008_thumbup_rigid
3. 0010_thumbtuckrigid
4. 0011_aokay
5. 0016_fist
6. 0026_four_count
7. 0035_palmdown
8. 0037_fingerspreadrelaxed
9. 0042_pucker
10. 0051_dinosaur

Interacting Hands:
1. 0259_dh_rightclaspleft
2. 0261_dh_fingergun
3. 0263_dh_leftfistcoverright
4. 0264_dh_interlockedfingers
5. 0266_dh_pray

2. Additional Experiments and Results

2.1. Deformation Methods

We conduct additional ablative experiments to validate the design of the proposed deformation field in HandNeRF.

*Corresponding Authors.

Figure 1. **Additional rendering results produced by HandNeRF.** The interaction results are produced by directly taking self-occluded interacting hands as the training samples.

	PSNR ↑	SSIM ↑	LPIPS↓	DE ↓
w/o EC	29.1289	0.9488	0.1165	0.1453
w/ NBW	32.7641	0.9667	0.0615	0.1405
w/ NBW & EC	32.9191	0.9684	0.0572	0.1389
HandNeRF (Ours)	**33.0204**	**0.9737**	**0.0475**	**0.1360**
Independent canonical	29.8202	0.9470	0.0939	0.2165
Learnable composition	30.9025	0.9541	0.0728	0.1878
HandNeRF (Ours)	**30.9256**	**0.9570**	**0.0700**	**0.1840**

Table 1. **Additional ablation studies.** We explore different designs of the deformation field with the single-hand setting, ablating the error-correction network (EC) and the neural blend weight field (NBW). Note that without EC and NBW, the deformation field is degraded to a deterministic transformation based on the human priors. In addition, two variants for modeling both hands are presented to validate our simple but effective design.

The neural blend weight field proposed by [2] is employed aside from the error-correction network. The performance comparison is shown in Tab. 1. Apparently, jointly optimizing the blend weights does not bring significant performance improvement.

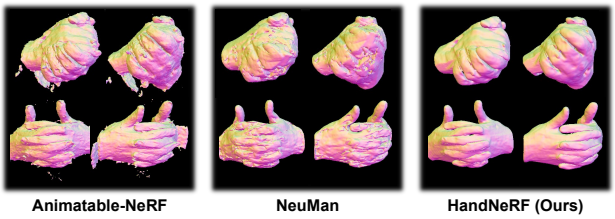

Figure 2. Reconstructed geometries shaded by surface normal.

2.2. Canonical Sharing for Both Hands

We replace the shared canonical model with an individual NeRF model for each hand. Tab. 1 shows that the performance drops without the canonical sharing between two hands. Moreover, it takes an additional 10% of training time to converge, compared with the canonical-sharing model.

It is worth noting that all our training samples capture only one human's interacting hands, since we focus on practical scenarios like sign language conversations. Therefore, we assume in this paper that the two hands are highly sym-

metric and share the same canonical model, enabling the self-complementation of geometry and texture in rarely-observed areas. We believe this is a worthy trade-off against the minor error caused by possible differences of two hands.

2.3. Composition strategy

We use additional learnable weights conditioned on the interacting poses for samples in ray-tracing. The model tends to converge slightly faster, but Tab. 1 shows that no significant performance gain is achieved.

2.4. Additional Visualization

More rendering results of HandNeRF are exhibited in Fig. 1, which are captured with a generated circumferential moving trajectory of camera. Please also refer to the video we provide along with this supplementary document for the continuous transformation of views and poses. The synthesis results prove that our method is able to handle various complex interacting poses and maintain excellent multi-view consistency at the same time.

We also provide some qualitative results of geometry reconstruction for HandNeRF and two baselines. The meshes and surface normals are extracted using marching cubes. As exhibited in Fig. 2, HandNeRF outperforms the baselines, mainly thanks to the depth-guided density optimization.

3. Limitations and Discussions

As the first NeRF model designed for photo-realistic novel view/pose image synthesis of interacting hands, there is still room for improvement in HandNeRF. It can be observed in our rendering results that some black artifacts still exist for specific views, especially in the frequently-obscured areas of interacting hands. That is probably caused by the lack of visible texture information. Appropriate post-processing operations need to be explored to remove those artifacts and further improve the visual quality of the rendered images. Besides, without the stage-two pose adaptation, the direct generalization of our method for unseen poses cannot properly handle relatively large deformations from the limited training samples. One possible solution is introducing a pre-trained pose estimation network to enhance the capability of pose representations, thus improving the model's adaptability for out-of-distribution poses.

References

[1] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 1

[2] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14314–14323, 2021. 2