

Neural Voting Field for Camera-Space 3D Hand Pose Estimation

Supplementary Material

Lin Huang^{1*} Chung-Ching Lin² Kevin Lin² Lin Liang²
Lijuan Wang² Junsong Yuan¹ Zicheng Liu²
¹University at Buffalo ²Microsoft

This is the supplementary materials of the main text. Sec. **A** provides more details regarding the method configurations and training procedure. Sec. **B** presents full quantitative results of the two baselines (*i.e.*, Baseline-Holistic and Baseline-2D-Dense) for camera-space 3D hand pose on FreiHAND [11] with Comp [1] pre-training. Sec. **C** presents qualitative comparisons between our CS-NVF and the state-of-the-art methods (*i.e.*, CMR [2] and I2L-MeshNet [8]) for the task of camera-space 3D hand pose estimation on complex and failure cases (*e.g.*, severe occlusion and extreme poses) on FreiHAND [11]. Sec. **D** provides additional qualitative results from CS-NVF for camera-space 3D hand pose on FreiHAND [11] and RS-NVF for root-relative 3D hand pose on HO3D [4]. To demonstrate the generalization ability of NVF, Sec. **E** provides qualitative results on Real-World dataset [3] using CS-NVF trained on FreiHAND only.

A. Additional Implementation Details

Continuing from Sec. 4.3 of implementation details in the main text, we provide more details regarding the method configurations and training procedure in this section.

Additional Method Configurations. For each single RGB image used for all of our models, different from PIFu [9], no image segmentation is applied. Moreover, for camera-space 3D hand pose estimation, CS-NVF and both baselines take the original image with resolution 224×224 as input without hand detection or cropping applied. As in [6], to remove the ambiguity caused by using images captured by cameras with different focal lengths during training for absolute 3D hand pose estimation, given the provided camera intrinsic parameters from FreiHAND, we remap each input image to a reference pinhole camera with the same focal length which can be arbitrarily chosen. For root-relative 3D hand pose estimation, RS-NVF takes cropped hand-centered image with resolution 128×128 as input without remapping applied. During training, the balancing weight used for CS-NVF and Baseline-2D-Dense is set to 0.1 and the balancing weight used for RS-NVF is set to 10.

*Work done during Lin Huang’s internship with Microsoft.

Training Procedure. We report results achieved by the proposed NVF and the two baselines under various training settings (*i.e.*, with and without hand scale or extra data). For our models without the use of hand scale and extra data as a fair comparison with the state-of-the-art methods, the image encoder (g) and MLP (f_{NVF} , f_{HOL} , f_{DEN}) are initialized via xavier initialization. RMSProp [10] is used for optimization with the batch size of 24, the number of epochs of 650, and the initial learning rate of 0.0001. The learning rate is decayed by the factor of 0.1 at 400-th, 500-th, and 600-th epoch. The models can also be pre-trained first on FreiHAND to estimate relative 3D hand pose. RMSProp is used for optimization with the batch size of 64, the number of epochs of 600, and the initial learning rate of 0.0005. The learning rate is decayed by the factor of 0.1 at 250-th, 350-th, and 450-th epoch. We then fine-tune the models for respective tasks. RMSProp is used for optimization with the batch size of 24, the number of epochs of 60, and the initial learning rate of 0.0001. The learning rate is decayed by the factor of 0.1 at 25-th and 50-th epoch. We found that pre-training on FreiHAND to estimate relative hand pose helps to improve generalization on input images from other domains, especially for unseen poses. For results using extra data, the models are pre-trained first on Comp to estimate relative hand pose and RMSProp is used for optimization with the batch size of 64, the number of epochs of 700, and the initial learning rate of 0.0005. The learning rate is decayed by the factor of 0.1 at 500-th, and 600-th epoch. For our ablation study on hand scale for camera-space 3D hand pose, we directly use the hand scale provided by FreiHAND during evaluation. The hand scale is defined as the metric length of a reference bone which is the phalangeal proximal bone of the middle finger. When hand scale is used, it is concatenated with the input to the MLP for processing.

B. Baseline Results with Extra Data

Besides the results of the two baselines shown in Tab. 2 and Tab. 3 of the main paper for camera-space 3D hand pose estimation, we also provide results of the two baselines on the metric of CS-MJE with Comp pre-training in Tab. 7. This then provides the full results from CS-NVF and the

Method	Extra Data	Hand Crop	Hand Scale	CS-MJE↓
ObMan [5]	-	✓	✗	85.2
MANO CNN [11]	-	✓	✗	71.3
I2L-MeshNet [8]	-	✓	✗	60.3
CMR-SG-RN18 [2]	-	✓	✗	49.7
CMR-SG-RN50 [2]	-	✓	✗	48.8
Baseline-Holisitc	-	✗	✗	54.5
Baseline-2D-Dense	-	✗	✗	53.2
CS-NVF (Ours)	-	✗	✗	47.2
Baseline-Holisitc	-	✗	✓	50.4
Baseline-2D-Dense	-	✗	✓	49.0
CS-NVF (Ours)	-	✗	✓	42.4
Baseline-Holisitc	Comp*	✗	✗	51.3
Baseline-2D-Dense	Comp*	✗	✗	50.9
CS-NVF (Ours)	Comp*	✗	✗	44.6
Baseline-Holisitc	Comp*	✗	✓	44.3
Baseline-2D-Dense	Comp*	✗	✓	43.4
CS-NVF (Ours)	Comp*	✗	✓	39.3

Table 7. **Comparison for absolute 3D hand pose on FreiHAND.** *: pre-training on Comp. Note that our CS-NVF and two baselines take the original image without hand detection or cropping.

two baselines under different training settings (*i.e.*, with and without hand scale or extra data) for camera-space 3D hand pose on FreiHAND, as shown in Tab. 7 and Tab. 8.

C. Qualitative Comparisons on Complex and Failure Cases

In Fig. 5, we provide qualitative comparisons between our proposed CS-NVF and the state-of-the-art methods (*i.e.*, CMR [2] and I2L-MeshNet [8]) for the task of camera-space 3D hand pose estimation on complex and failure cases on FreiHAND [11]. Specifically, the complex and failure cases as shown in Fig. 5 can be divided into three categories:

- Severe self-occlusion caused by extreme viewpoint.
- Severe occlusion caused by hand-object interaction.
- Extreme pose.

For each pair of images, the left shows the input RGB image that the method uses during inference and the right shows the predicted camera-space 3D hand pose directly rendered by camera intrinsic parameters. Note that, as shown in the figure, CS-NVF takes the original RGB image as input without hand detection and cropping, while both CMR and I2L-MeshNet take the cropped image. Based on the qualitative comparisons shown in Fig. 5 from three methods, we observe that:

- Self-occlusion: for results shown in row (1-4) of Fig. 5, when only a small portion of the hand is visible in the input image caused by extreme viewpoint, our proposed CS-NVF robustly recovers more plausible and accurate pose structure than CMR and I2L-MeshNet.

Method	Extra Data	Hand Scale	TE↓	DE↓
Baseline-Holisitc	-	✗	50.6	49.1
Baseline-2D-Dense	-	✗	49.2	47.9
CS-NVF (Ours)	-	✗	43.6	42.4
Baseline-Holisitc	-	✓	46.9	45.5
Baseline-2D-Dense	-	✓	45.3	43.9
CS-NVF (Ours)	-	✓	38.9	37.8
Baseline-Holisitc	Comp*	✗	48.7	47.1
Baseline-2D-Dense	Comp*	✗	47.9	46.4
CS-NVF (Ours)	Comp*	✗	41.5	40.4
Baseline-Holisitc	Comp*	✓	41.7	40.1
Baseline-2D-Dense	Comp*	✓	40.5	38.8
CS-NVF (Ours)	Comp*	✓	36.5	35.5

Table 8. **Comparison of 3D Translational and Depth Error for absolute 3D hand pose on FreiHAND.** *: pre-training on Comp.

For example, in row (1), CMR shows three fingers up given the input hand with two fingers up and in row (3), I2L-MeshNet generates implausible structure.

- Object occlusion: for results shown in row (5-6) of Fig. 5, with severe occlusion caused by the interacted bottle, CS-NVF can better recover the occluded four fingers behind the bottle and show an overall reasonable gesture of hand grabbing a bottle. In row (6), CS-NVF is able to provide solid estimation for index finger which is entirely occluded.
- Extreme pose: for results shown in row (7-8) of Fig. 5, while all three methods can generate plausible hand pose structure, the three methods all fail at the part where the two fingers are crossed. Challenging poses shown in the row (7-8) usually are tail-distributed poses in most hand datasets. Thus, to tackle this problem, we argue that improving both fine-grained reasoning towards uncommon gestures for pose estimation pipeline and the pose distribution for hand dataset are required. Both aspects will be investigated in our future work.
- 3D-2D alignment: Besides the plausibility and accuracy of pose articulated structure itself, for the task of camera-space 3D hand pose estimation, we also need to look at the 3D global information (*i.e.*, rotation and translation), which can be indicated by the alignment between the rendered hand pose and input hand area to some extent. With severe occlusion as shown in row (1-6), rendered 3D poses from CS-NVF generally show better 3D-2D alignment.

- Overall: while the performances on these challenging scenarios from all three methods are usually worse than performances on common cases, CS-NVF has shown its capability to recover more robust 3D hand pose in camera space in various challenging scenarios, compared with the state-of-the-art methods (i.e., CMR and I2L-MeshNet) with respect to the pose plausibility, pose accuracy, and 3D-2D alignment.

D. Additional Qualitative Results

In Fig. 6 and Fig. 7, we provide more qualitative results for CS-NVF for camera-space 3D hand pose estimation on FreiHAND and RS-NVF for root-relative 3D hand pose estimation on HO3D.

Given an RGB input, for each of the 3D query points densely sampled in camera frustum (CS-NVF) or hand root-relative 3D cube (RS-NVF), NVF regresses: (i) the signed distance between the point and the hand surface; (ii) a set of 4D offset vectors. Each 4D offset vector consists of a 1D voting weight and a 3D unit directional vector from the query point to each hand joint, representing the closeness and direction from the point to each joint. Following a vote-casting scheme, 4D offset vectors from near-surface points (i.e., points for which the predicted signed distance is below the clamping distance) are selected to calculate the 3D joint coordinates by a weighted average. Based on the overall pipeline, for each evaluation sample, we show:

- Column (a): single RGB input image.
- Column (b): 3D hand mesh generated by Marching Cubes [7] from the signed distances predicted at 3D query points in the camera frustum (CS-NVF) or hand root-relative 3D cube (RS-NVF).
- Column (c-h): white circles as the valid 3D voters are the 3D points in the hand surface vicinity (i.e., points for which the predicted signed distance is below the clamping distance). Colored line denotes the predicted 1D voting weight from each near-surface 3D point (white circle) to a joint. Note that brighter line means larger weight, darker or no line means smaller weight.
- Column (i): estimated 3D pose via weighted average over predicted 4D offset vectors from near-surface points, following a vote-casting scheme.

Camera-Space 3D Hand Pose Estimation. Since CS-NVF generates all the predictions in the 3D camera space, all the results shown in column (b-i) of Fig. 6 are directly rendered by camera intrinsic parameters.

Root-Relative 3D Hand Pose Estimation. For RS-NVF, since it generates all the predictions in the root-relative space, all the results in column (b-i) of Fig. 7 are first translated from the root-relative space into the camera space us-

ing the provided 3D ground-truth root location and then rendered by camera intrinsic parameters.

Discussion. Based on qualitative results shown in Fig. 6 from CS-NVF and Fig. 7 from RS-NVF, we observe that:

- Valid 3D voters: as indicated by the white circles in column (c-h) of both figures, among all the 3D points sampled at centers of voxels that fill up the camera frustum (CS-NVF) or hand root-relative 3D cube (RS-NVF) during inference, NVF is able to find points in the hand surface vicinity (i.e., points for which the predicted signed distance is below the clamping distance) even in occluded region and use these points as valid 3D voters for 3D pose estimation. Different from classic pixel-level dense regression methods which mainly model foreground pixels, NVF shows its ability to reason points around the whole 3D hand surface.
- 1D voting weight predictions: as shown in column (c-h) of both figures, given the input images with self-occlusion, occlusion caused by object, and complex poses, NVF is generally able to predict large voting weight for points close to hand joint and small value for points far away, resulting in well-shaped symmetric distribution around corresponding joint, even in occluded regions. Note that for each colored line, brighter line indicates larger voting weight, darker or no line indicates smaller voting weight. This demonstrates NVF’s ability to model relation between each near-surface point and each hand joint.
- Signed distance predictions: as shown in column (b) of both figures for the hand mesh rendering (obtained using Marching Cubes from predicted signed distances), even in highly occluded region, NVF provides solid signed distance distribution showing its ability to reason the global hand structure/geometry.
- Estimated 3D hand pose: based on robust 3D point-wise predictions, NVF can then recover accurate 3D hand pose in challenging cases with severe occlusion and complex poses as shown in column (i) of both figures. Overall, these help to verify that, through direct dense modeling in 3D domain, NVF can model 3D dense local evidence and also the global hand structure/geometry, leading to robust 3D hand pose.

E. Qualitative Results on Real-World dataset

In Fig. 8, we provide qualitative results on Real-World dataset [3] using our CS-NVF which is trained on FreiHAND dataset only. For each pair of images, the left shows the input RGB image and the right shows the 3D hand pose directly rendered by camera intrinsic parameters. As shown in the figure, our method is able to generate solid results when testing on images in the wild/from another domain with various poses. This should demonstrate, to some extent, the generalization capability of our proposed NVF.

Acknowledgments. This work is supported in part under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

- [1] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022. [1](#)
- [2] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *CVPR*, 2021. [1](#), [2](#), [5](#)
- [3] Lihao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. [1](#), [3](#), [8](#)
- [4] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. [1](#)
- [5] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. [2](#)
- [6] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *ECCV*, 2022. [1](#)
- [7] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIG-GRAPH*, 1987. [3](#)
- [8] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. [1](#), [2](#), [5](#)
- [9] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. [1](#)
- [10] Tijmen Tieleman and Geoffrey Hinton. Rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012. [1](#)
- [11] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019. [1](#), [2](#)

Qualitative Comparisons on **Complex and Failure Cases** for Camera-Space 3D Hand Pose between **CS-NVF (Ours)** and the **state-of-the-art methods**



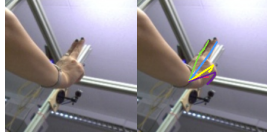

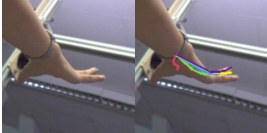





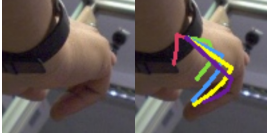
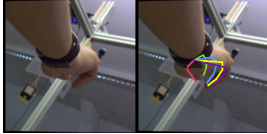
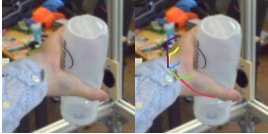
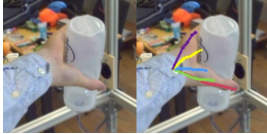
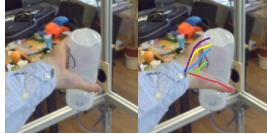






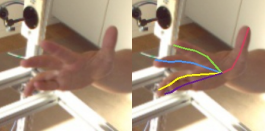
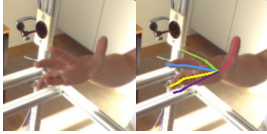
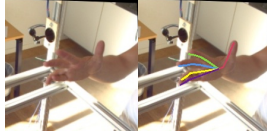
	CMR	I2L-MeshNet	CS-NVF (Ours)
Self-Occlusion caused by Extreme Viewpoint	(1) 		
	(2) 		
	(3) 		
	(4) 		
Severe occlusion caused by Interacted Object	(5) 		
	(6) 		
Extreme Pose	(7) 		
	(8) 		
	CMR	I2L-MeshNet	CS-NVF (Ours)

Figure 5. **Qualitative comparisons on complex and failure cases for absolute 3D hand pose on FreiHAND.** For each pair of images, the left shows the input RGB image that the method uses during inference and the right shows the predicted camera-space 3D hand pose directly rendered by camera intrinsic parameters. Compared with the state-of-the-art methods (*i.e.*, CMR [2] and I2L-MeshNet [8]) on complex and failure cases with respect to the pose plausibility, pose accuracy, and 3D-2D alignment, CS-NVF has shown its capability to recover robust 3D hand pose in camera space when facing severe self-occlusion caused by extreme viewpoint, occlusion caused by interacted object, and extreme pose.



Figure 6. **Additional qualitative results from CS-NVF for camera-space 3D hand pose on FreiHAND.** CS-NVF can handle challenging cases of self-occlusion, occlusion by interacted object, and complex poses, leading to robust 3D hand pose. Please refer to Sec. D for specific meaning of the rendering of camera-space 3D prediction results shown in columns from (b) to (i).



Figure 7. Additional qualitative results from RS-NVF for root-relative 3D hand pose on HO3D. RS-NVF can handle challenging cases of self-occlusion and occlusion by interacted object, leading to robust 3D hand pose. Please refer to Sec. D for specific meaning of the rendering of root-relative 3D prediction results shown in columns from (b) to (i).

Qualitative Results on **Real-World dataset** using **CS-NVF** trained on **FreiHAND** only

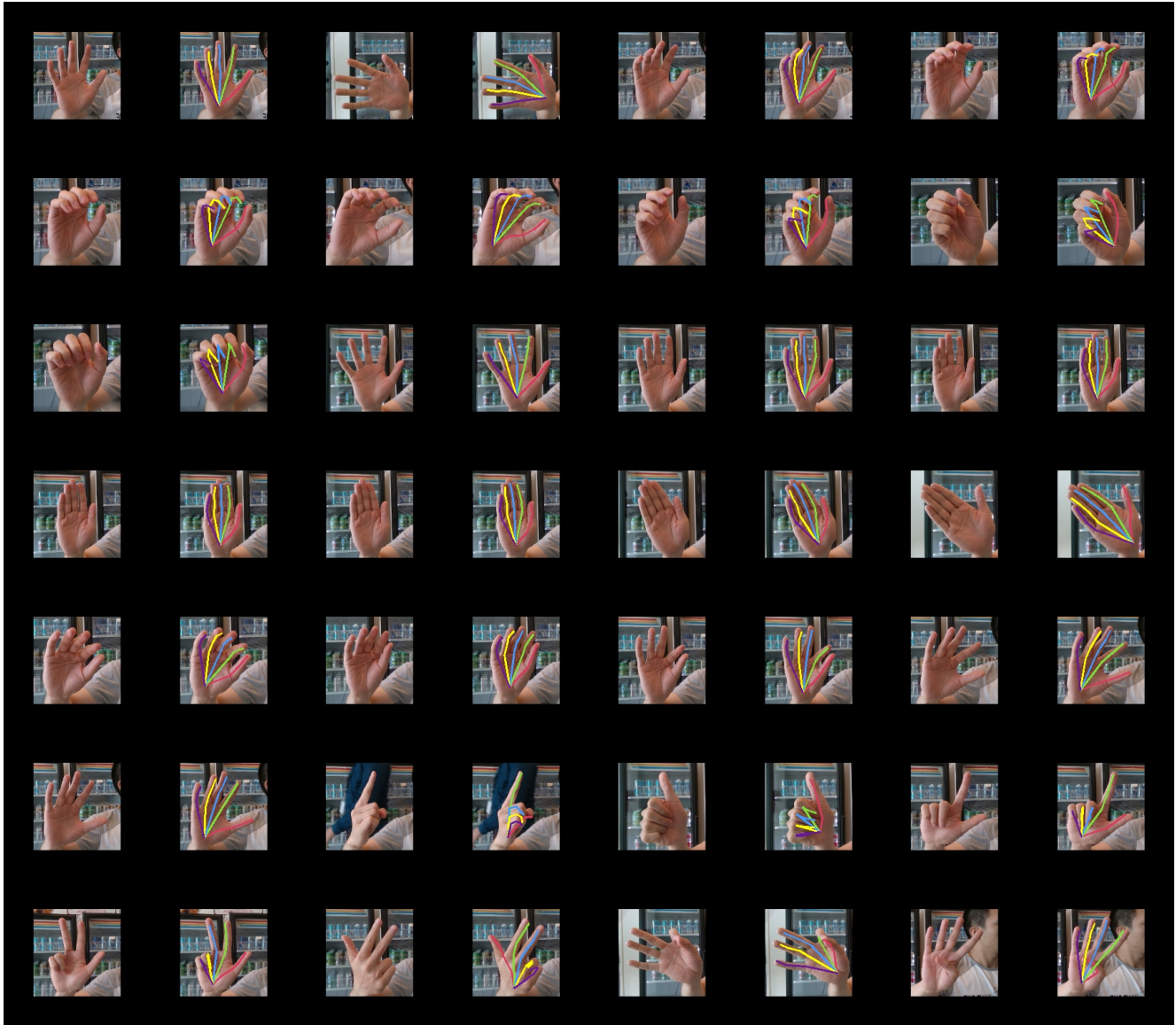


Figure 8. **Qualitative results on Real-World dataset [3] using CS-NVF trained on FreiHAND only.** For each pair of images, the left shows the input RGB image and the right shows the 3D hand pose directly rendered by camera intrinsic parameters. Our method shows its ability to generate solid results when testing on images in the wild/from another domain with various poses. This demonstrates, to some extent, the generalization capability of our proposed NVF.