MobRecon: Mobile-Friendly Hand Mesh Reconstruction from Monocular Image – Supplementary Materials

Xingyu Chen[1]* Yufeng Liu[3] Yajiao Dong[1] Xiong Zhang[2] Chongyang Ma[1]
Yanmin Xiong[1] Yuan Zhang[1] Xiaoyan Guo[1]
[1]Y-tech, Kuaishou Technology [2] YY Live, Baidu Inc.
[3]SEU-ALLEN Joint Center, Institute for Brain and Intelligence, Southeast University, China.

1. Complement Dataset

Motivation. As shown in Table 1, many datasets are developed for 3D hand pose estimation [16, 18, 13, 14, 19, 9, 6, 17, 2, 10, 5, 15, 12]. To collect real-world hand data, existing datasets are usually captured using a multi-view studio and annotated via semi-automatic model fitting [19, 6]. However, these model-fitted datasets usually suffer from noisy annotation, lack of background diversity, and costly data collection. An alternative way is computer-aided synthetic data [18, 3], which are superior in scalability, distribution, annotation, and collection cost. In addition, a good training dataset should avoid long-tailed distributions. That is, both hand poses and viewpoints should be uniformly distributed. Unfortunately, we are not aware of any existing dataset that fits this requirement. Some datasets try to alleviate the problem of limited viewpoints by multi-view rendering (*e.g.*, MVHM [3] contains 8 views), but they are still too sparse to cover all the possible views. Boukhayma *et al*. [1] uniformly sampled MANO PCA components to produce various hand poses. However, the PCA space does not describe physical factors, so the corresponding sampling results cannot be intuitively controlled. Thereby, we are inspired to generate a more comprehensive hand dataset with sufficient and uniformly distributed hand poses and viewpoints.

Data Designs. We design a high-fidelity hand mesh with 5633 vertices and 11232 faces. Different from existing hand datasets, we uniformly design hand poses. First, as shown in Figure 1, we set two states for each finger, *i.e.*, total bending and extending. Then, we obtain 32 base poses by combining five finger states. The combination of these base poses results in 496 pose pairs. For each pair, we uniformly interpolate three intermediate poses from one pose to another in Maya software[1] (as shown in Figure 2). In total, we obtain 1520 uniformly distributed hand poses.

Dataset	Type	Size	Mesh	UP	MV
STB [16]	real	36K	×	×	×
RHD [18]	synthetic	44K	×	×	×
GANerated Hands [12]	synthetic	331K	×	×	×
SeqHAND [15]	synthetic	410K	✓	×	×
EgoDexter [13]	real	3K	×	×	×
Dexter+Object [14]	real	3K	×	×	×
FreiHAND [19]	real	134K	✓	×	×
YoutubeHand [9]	real	47K	✓	×	×
ObMan [8]	synthetic	153K	✓	×	×
HO3D [7]	real	77K	✓	×	×
DexYCB [2]	real	528K	✓	×	×
H2O [10]	synthetic	571K	✓	×	×
FPHA [5]	synthetic	105K	×	×	×
H3D [17]	real	22K	✓	×	✓(15)
MHP [6]	real	80K	×	×	✓(4)
MVHM [3]	synthetic	320K	✓	×	✓(8)
InterHand2.6M [11]	real	2.6M	✓	×	✓(80)
ours	synthetic	328K	✓	✓	✓(216)

Table 1. Comparison among RGB-based 3D hand datasets. "MV" means multi-view, and the number in brackets shows the total number of views. "UP" denotes uniform pose distribution.

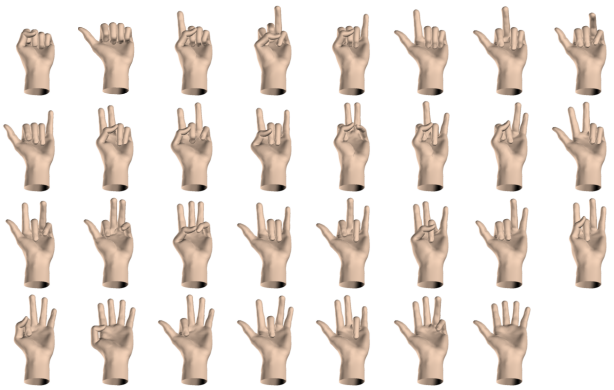

Figure 1. Base poses. Under the consideration of politeness, one pose with middle finger extending is not shown.

For each pose sample, we provide its dense viewpoints by rendering. To this end, we uniformly define 216

*Corresponding author, chenxingyu@kuaishou.com
[1]https://www.autodesk.com/

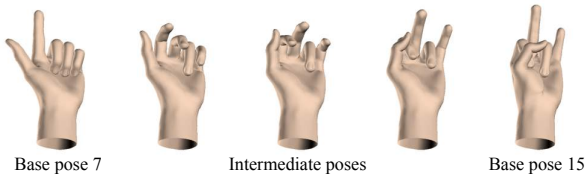

Figure 2. Intermediate poses from base pose 7 to base pose 15.

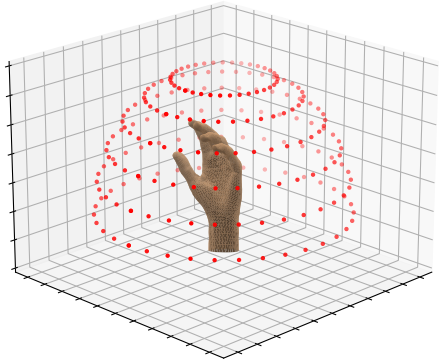

Figure 3. Illustration of viewpoints to render the dataset. Each red point denotes a camera position.The camera points to the palm center.

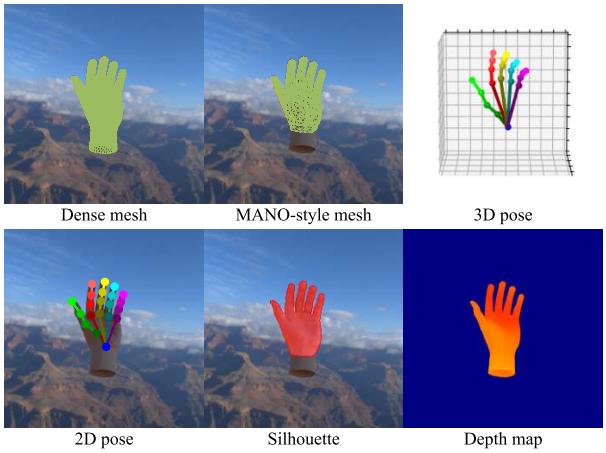

Figure 4. Annotations.

hemispherical-arranged camera positions. As shown in Figure 3, the longitude ranges from 0 to 2π while the latitude ranges from 0 to $\pi/2$. Adjacent positions differ in longitude or latitude by $\pi/18$ or $\pi/12$. All cameras point to the palm center so that the hand locates at the center of rendered images. Because the end of the wrist locates at the sphere center, the hemispherical sampling contains the first-person perspectives. As for the background, we collect high-dynamic-range (HDR) imaging with real scenes and illumination for rendering so that our hand mesh can realistically blend into various scenes. Figure 11 illustrates rendered samples with our viewpoints.

The automatically generated annotations involve no noise. Consistent with mainstream datasets, we design a

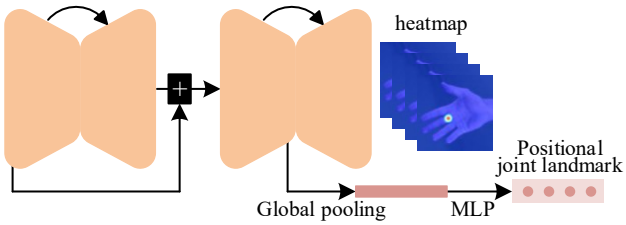

Figure 5. The pre-training architecture using our data.

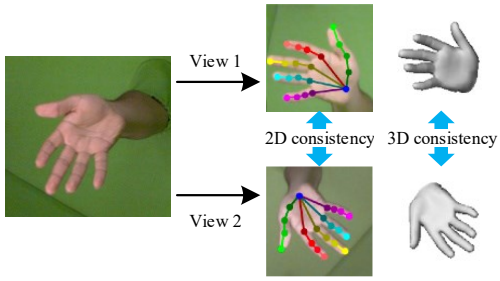

Figure 6. Consistency loss based on data augmentation.

pose-agnostic matrix to map our dense topology to MANO-style mesh with 778 vertices and 1538 faces. As shown in Figure 4, we provide annotations of our designed dense mesh, MANO-style mesh, 3D pose, 2D pose, silhouette, depth map, and intrinsic camera parameters.

Discussion. The limitation of our data is the lack of shape/texture diversity. Additionally, we only consider finger bend, and we plan to model finger splay to extend this dataset to cover the entire pose space uniformly.

Network pre-training. To pre-train the 2D encoding network, we design a 2D pose estimation task without the need of 3D annotation. In the main text, we analyzes 2D representations with heatmap and position regression. Hence, as shown in Figure 5, we equally consider these representations during the pre-training step. That is, both heatmap and positional joint landmark are supervised. The model is pre-trained for 80 epochs with a mini-batch size of 128. The initial learning rate is 10^{-3}, which is divided by 10 at the 20th, 40th, and 60th epochs. The input resolution is 128×128.

2. Analysis and Application

Diagram of our consistency loss. As shown in Figure 6, two views are derived with data augmentation with an input image. Then consistency loss can be designed in both 2D and 3D spaces as Equation 11 in the main text.

Explanation of dataset setting. During the ablation study in the main text, we use RHD, FreiHAND, and HO3Dv2 to evaluate different properties. Because FreiHAND and HO3Dv2 do not release ground truth and the of-

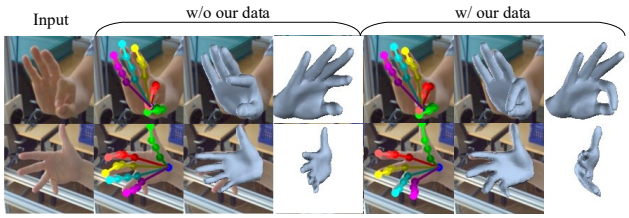

Figure 7. Qualitative visualization of 2D pose, aligned mesh, and side-view mesh.

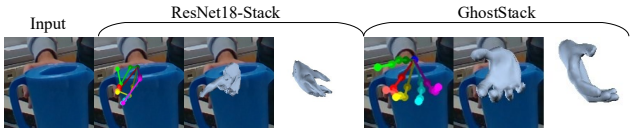

Figure 8. Comparison of GhostStack and ResNet18 on a challenging HO3Dv2 sample.

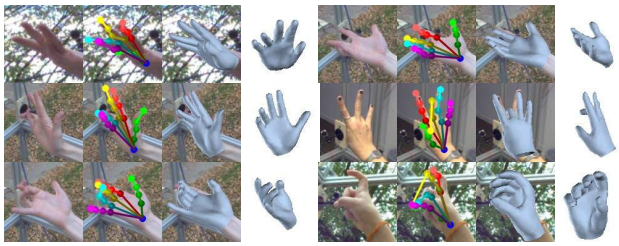

Figure 9. Typical failure cases.

ficial tools do not support 2D evaluation, RHD is employed for testing 2D accuracy. HO3Dv2 is a sequential dataset, so it is adopted to reflect temporal coherence. However, HO3Dv2 highlights hand-object interaction, which is not our topic. In contrast, FreiHAND highlights various hand poses, lighting conditions, *etc*., so we use it for evaluating 3D accuracy.

The effect of our complement data during fine-tuning. As shown in Figure 7, when our data are employed in fine-tuning step, it can improve the model performance on difficult pose prediction.

Visualization on HO3Dv2. Referring to Table 6 in the main text, our MobRecon outperforms some ResNet-based models. We observe that this phenomenon is related to generalization performance. As shown in Figure 8, HO3Dv2 contains massive seriously occluded samples. Under this extreme condition, our model can produce a physically correct prediction while the ResNet-based model collapses.

Failure case analysis. As shown in Figure 9, MobRecon could suffer from failure cases as for challenging poses. Typically, self-occlusion by finger splay is hard to accurately predict because they are tail-distributed poses in most datasets. We will solve this problem by improving our complement data, as stated in the above section.

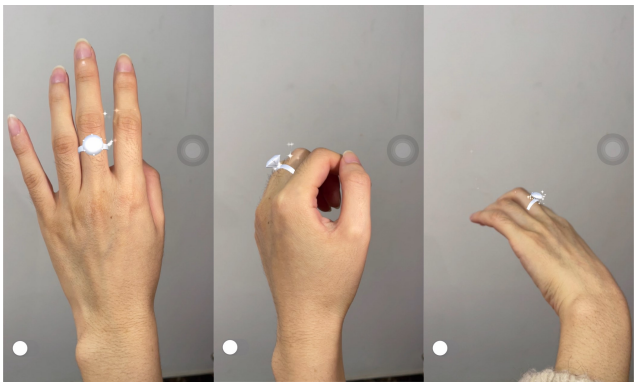

Figure 10. We develop an AR effect with MobRecon and deploy it on mobile devices. This figure is captured with iPhone12.

More qualitative results. Figure 12 illustrates comprehensive qualitative results of our predicted 2D pose, aligned and side-view mesh. The challenges include challenging poses, object occlusion, truncation, and bad illumination. Overcoming these difficulties, our method can generate accurate 2D pose and 3D mesh.

Qualitative comparison on temporal coherence. We record a video snippet to demonstrate temporal coherence, where we keep the camera and hand static to produce low acceleration. Despite the static condition, the network input could be temporally unstable because of detection jitter *etc*. The ground-truth pose is straightforward (see Figure 13), and all compared models can easily obtain high accuracy. Hence, temporal performance can be exclusively revealed in this experiment. As shown in Figure 13, our MobRecon performs better than CMR [4] in terms of 2D/3D pose consistency. In addition, we also compute the root coordinates with the method in [4] and achieve better root recovery stability. Besides, we also complement 2D PCK curves on RHD, which demonstrate that our method has better 2D pose accuracy. Beyond accuracy and temporal coherence, our MobRecon with MapReg can produce better articulated structures because of global receptive field and adaptive inter-landmark constraints (see Figure 6 in the main text).

Mobile application. Based on our MobRecon, a virtual ring can be worn with AR technique (Figure 10).

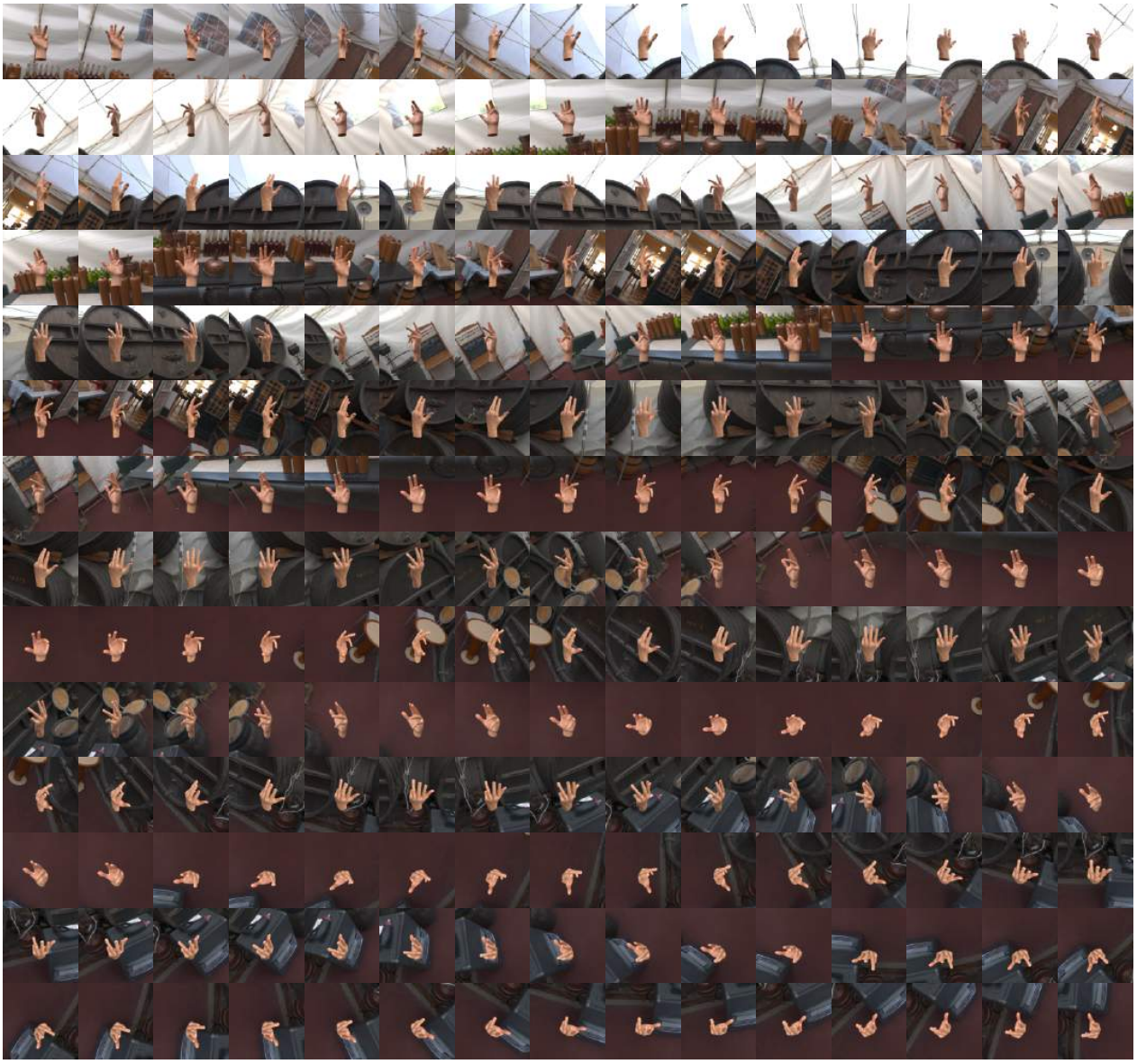

Figure 11. Rendered samples with dense and uniform viewpoints.

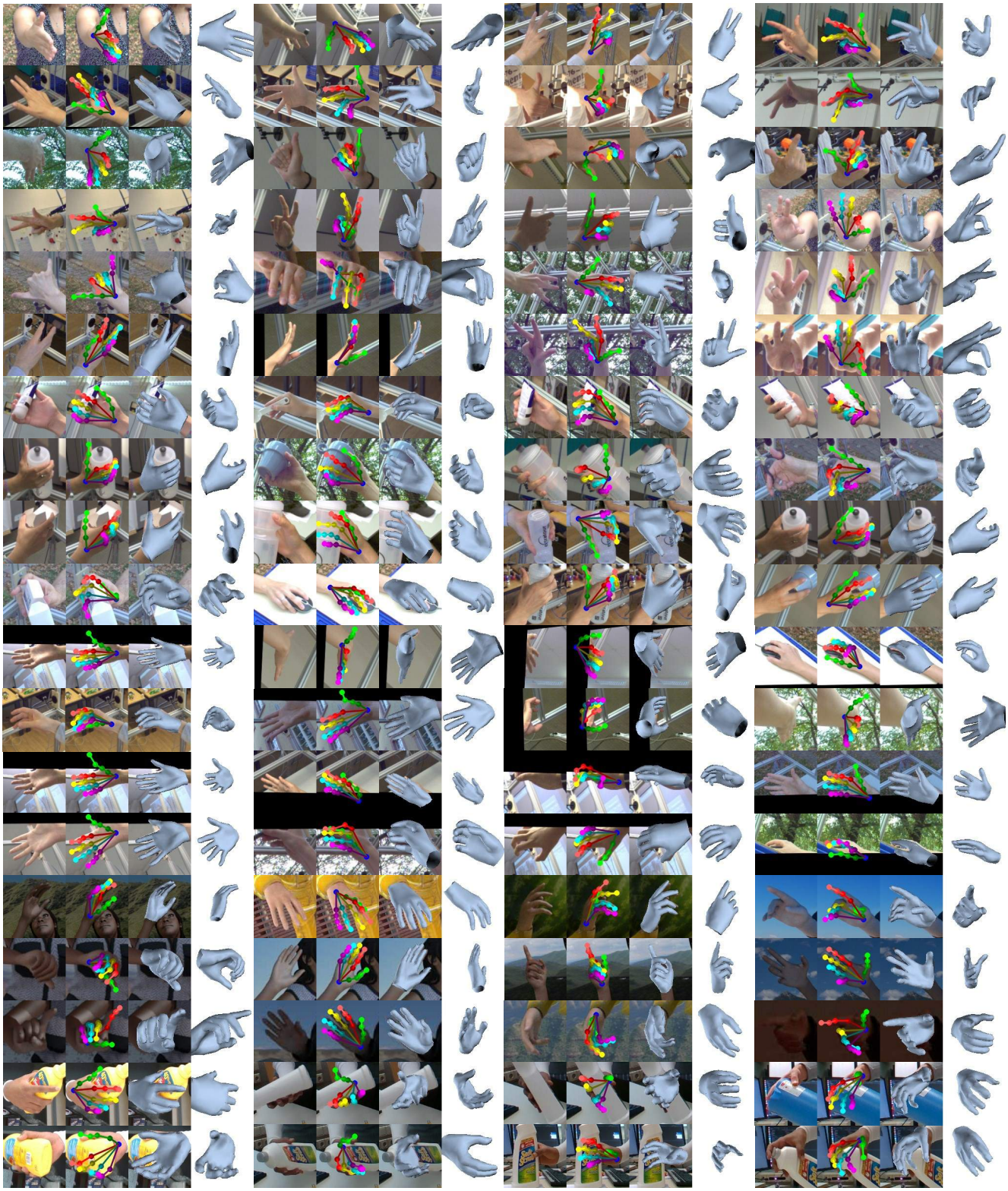

Figure 12. Qualitative visualization of 2D pose, aligned mesh, and side-view mesh on FreiHAND, RHD, and HO3Dv2. Our method is robust enough to handle cases of occlusion, truncation, challenging poses, and bad illumination.

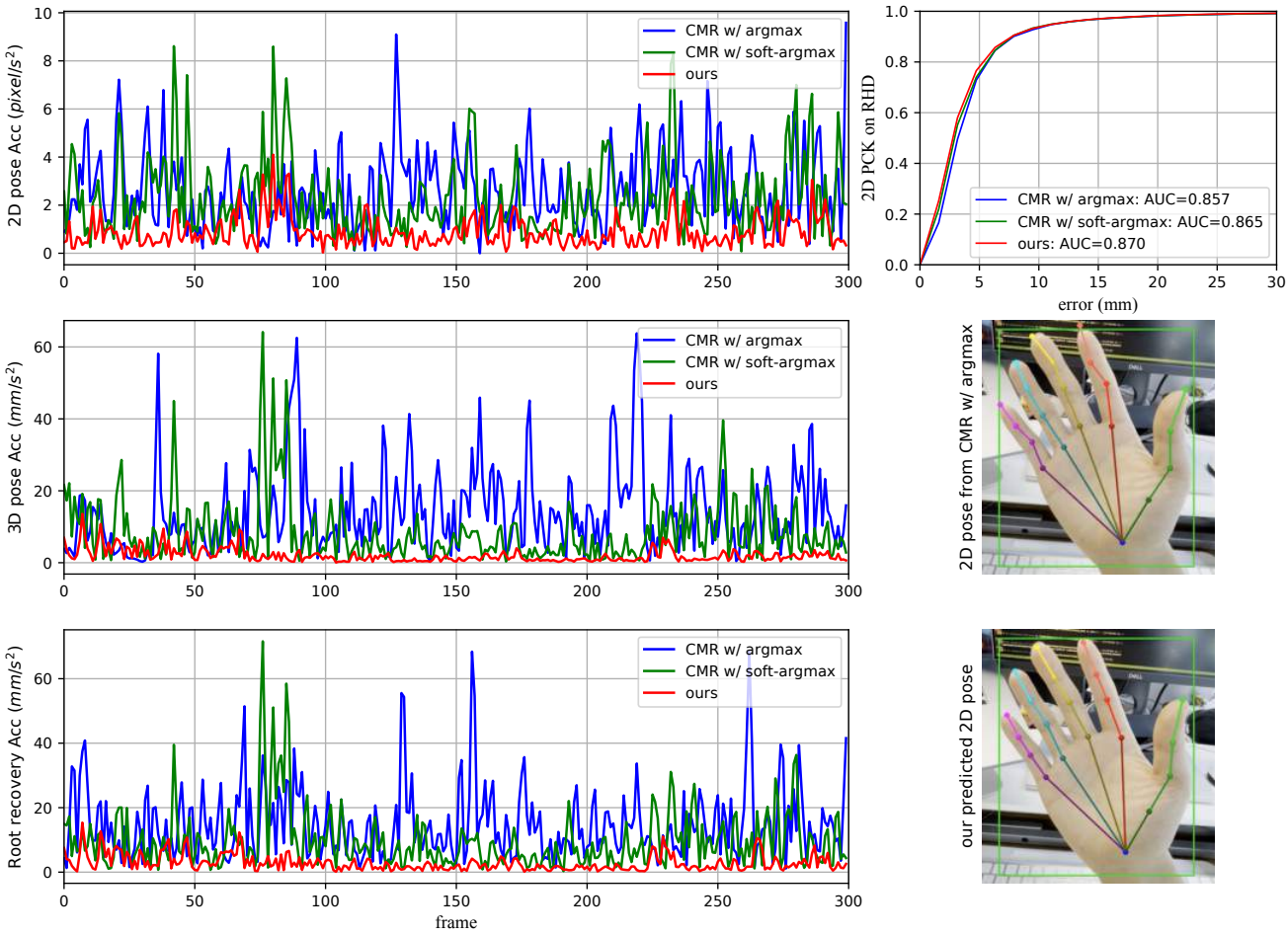

Figure 13. We record a video snippet with a straightforward and static hand pose (see the bottom right corner) to compare the temporal performance and articulated structure.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 1

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 1

[3] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. MVHM: A large-scale multi-view hand mesh benchmark for accurate 3D hand pose estimation. In *WACV*, 2021. 1

[4] 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. 3

[5] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 1

[6] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3D hand pose dataset. *Image and Vision Computing*, 2019. 1

[7] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 1

[8] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1

[9] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 1

[10] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 1

[11] 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

[12] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 1

[13] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *ICCV*, 2017. 1

[14] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *ECCV*, 2016. 1

[15] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. SeqHAND: RGB-sequence-based 3D hand pose and shape estimation. In *European Conference on Computer Vision*, pages 122–139. Springer, 2020. 1

[16] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3D hand pose tracking and estimation using stereo matching. *arXiv:1610.07214*, 2016. 1

[17] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3D-Studio: A new multi-view system for 3D hand reconstruction. In *ICASSP*, 2020. 1

[18] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017. 1

[19] 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