Supplementary Material for "Bringing Inputs to Shared Domains for 3D Interacting Hands Recovery in the Wild"

Gyeongsik Moon
Meta Reality Labs
mks0601@gmail.com

In this supplementary material, we provide more experiments, discussions, and other details that could not be included in the main text due to the lack of pages. The contents are summarized below:

A. Qualitative comparison
B. Ablation study on the architecture of TransNet
C. Verification of reproduced results in Table 5
D. Clarification of 2D-based weak supervision in Table 3
E. Architecture of DetectNet
F. Architecture of SHNet
G. Implementation details
H. Limitations

A. Qualitative comparisons

Fig. A shows that ours produces much more robust results than IntagHand [4] on in-the-wild images. Overall, IntagHand produces reasonable 3D hand mesh for a visible hand but fails to recover the other occluded hand. Also, it suffers from depth ambiguity as the first and fourth rows show, where the 2D error is small but the 3D error is large. The fourth row also shows that IntagHand fails to recover 3D relative translation between two hands due to the depth ambiguity, while ours successfully recovers. Finally, we think the reason why IntagHand produces non-hand shape meshes is that IntagHand directly regresses the 3D coordinates of 3D hand meshes. On the other hand, ours regresses MANO parameters and 3D meshes are obtained by forwarding the parameters to the MANO layer.

B. Ablation study on the architecture of TransNet

Table A shows that our fully convolutional network (FCN)-based architecture performs the best compared to widely used fully-connected (FC) network architecture and Transformer [12]. We think this is because our FCN can explicitly utilize the spatial relationship between voxels using the 2.5D heatmap representation. On the other hand, FC-based and Transformer-based settings take 2.5D coordinates as input. This result is consistent with previous studies [6], which demonstrates the superiority of heatmap representation over coordinate representation. As proposing better network architecture for TransNet is not our main focus, we believe developing its network architecture can be an interesting future direction. We used the architecture of Martinez *et al*. [5] for the FC setting, and Zheng *et al*. [14] for the Transformer setting.

Settings	HIC [11]	IH2.6M [7]
FC	56.76	29.85
Transformer	64.49	33.59
FCN (Ours)	**31.35**	**29.29**

Table A. MRRPE comparisons between TransNet that have various architectures.

Settings	Translation align	Scale align	MPVPE
Official model	Middle root	✓	9.36
Reproduced	Middle root	✓	9.34
Official model	Middle root	✗	9.99
Reproduced	Middle root	✗	9.81
Official model	Wrist	✗	15.24
Reproduced	Wrist	✗	14.12

Table B. Reproduce verification of IntagHand [4]. We report MPVPE on interacting hand sequences of InterHand2.6M.

Settings	Scale align	MPVPE
Original	✓	10.40
Reproduced	✓	10.70
Reproduced	✗	14.20

Table C. Reproduce verification of Zhang *et al*. [13]. We report MPVPE on interacting hand sequences of InterHand2.6M.

Figure A. Qualitative comparison between our InterWild and IntagHand [4] on MSCOCO. Ours detects hand boxes from the human images, while IntagHand takes the hand images using GT boxes.

C. Verification of reproduced results in Table 5

Table B and C verify our reproduce results of IntagHand [4] and Zhang *et al*. [13] on InterHand2.6M, respectively. For the verification of IntagHand [4], we used their officially released pre-trained model and evaluation code. As their model is trained only on interacting hand sequences of InterHand2.6M, we also trained a model following their training set for the verification. Their original evaluation setting is the first row of Table B. On the other hand, our evaluation setting in Table 5 of the main manuscript is the last row as most 3D hand recovery works [1, 2, 6, 7, 16] align translation with the wrist joint, a *root joint* in hand kinematic space. Using the middle root joint reduces the errors as the length of the kinematic chain to the fingertips becomes much shorter. Please note that the reproduced numbers can be different from those of Table 5 in the main manuscript as the results in Table 5 are from additional training on MSCOCO.

For the verification of Zhang *et al*. [13], we take numbers

from their paper as their released pre-trained model does not work. Following their training protocol, we trained a model only on the interacting hand sequence of InterHand2.6M. Please note that the reproduced numbers can be different from those of Table 5 in the main manuscript as the results in Table 5 are from additional training on MSCOCO.

D. 2D-based weak supervision in Table 3

We describe how the 2D-based weak supervision in Table 3 of the main manuscript is introduced. We modified TransNet to output 3D global translation of the right hand and 3D relative translation between two hands from the two hand input. We observed this produces better results than estimating 3D global translations of both hands. The 3D global translation of the left hand is obtained by adding the 3D relative translation to the 3D global translation of the right hand. Then, we translate root joint-relative 3D joint coordinates of each hand (*i.e.*, output of SHNet) using the 3D global translation of each hand. The translated 3D joint coordinates of two hands are projected to the TransNet's input space (*i.e.*, union of two hand boxes). Finally, we minimized the L1 distance between the projected and GT 2D coordinates. Please note that we make the shape parameter of MANO of left and right hands be the same during the weak supervision to minimize the scale ambiguity of each hand.

E. Architecture of DetectNet

DetectNet detects left and right hands from an input image $\mathbf{I}_{\text{det}} \in \mathbb{R}^{3\times H_{\text{det}}\times W_{\text{det}}}$, downsampled from a high-resolution image $\mathbf{I} \in \mathbb{R}^{3\times 2H_{\text{det}}\times 2W_{\text{det}}}$, by predicting two bounding boxes of the left and right hands. $H_{\text{det}} = 256$ and $W_{\text{det}} = 192$ denote height and width of $\mathbf{I}_{\text{det}}$, respectively. The downsampling is necessary to save computational costs. To this end, we extract the image feature from $\mathbf{I}_{\text{det}}$ using ResNet-50 and pass the feature to three consecutive deconvolutional layers, which upsample the feature map by 8 times. We denote the upsampled feature map by $\mathbf{F}_{\text{det}} \in \mathbb{R}^{C_{\text{det}}\times H_{\text{det}}/4\times W_{\text{det}}/4}$. $C_{\text{det}} = 256$ denotes the number of channel of $\mathbf{F}_{\text{det}}$. We use the original ResNet-50 after dropping global average pooling (GAP) and following fully-connected layers. Then, a 1-by-1 convolutional layer takes $\mathbf{F}_{\text{det}}$ and predicts a 2D heatmap of two hand bounding box centers. Soft-argmax [10] extracts 2D hand bounding box center coordinates from the 2D heatmap in a differentiable way. Then, we extract bounding box center features of left and right hands by performing a bilinear interpolation at the box center positions of $\mathbf{F}_{\text{det}}$. The extracted bounding box center features of each hand are passed to two fully-connected layers, which produce a scale of the bounding box. By decoding the bounding box centers and scales, we obtain two bounding boxes of left and right hands.

F. Architecture of SHNet

SHNet processes right and left hand images in the same way except for the input and output of the left hand image are flipped to the right hand and flipped back to the left hand, respectively. Hence, we omit right and left hand notations in the following description.

SHNet predicts 2.5D joint coordinates $\mathbf{P} \in \mathbb{R}^{J\times 3}$, MANO parameters, and 3D global translation $\mathbf{g} \in \mathbb{R}^3$ from a single hand image $\mathbf{I}_{\text{hand}} \in \mathbb{R}^{3\times H_{\text{hand}}\times W_{\text{hand}}}$. $H_{\text{hand}} = 256$ and $W_{\text{hand}} = 256$ denote height and width of $\mathbf{I}_{\text{hand}}$, respectively. $J = 21$ denotes the number of single hand joints. MANO parameters include 3D joint rotations $\theta \in \mathbb{R}^{16\times 3}$ and hand shape parameter $\beta \in \mathbb{R}^{10}$. The 3D global translation $\mathbf{g}$ is used in two cases: 1) loss calculations and 2) mesh rendering to visualize results.

2.5D joint coordinate estimation. ResNet-50 extracts an image feature $\mathbf{F}_{\text{hand}} \in \mathbb{R}^{C_{\text{hand}}\times H_{\text{hand}}/32\times W_{\text{hand}}/32}$ from a single hand image $\mathbf{I}_{\text{hand}}$. $C_{\text{hand}} = 2048$ denotes the number of channel of $\mathbf{F}_{\text{hand}}$. Then, the extracted image feature $\mathbf{F}_{\text{hand}}$ is passed to a 1-by-1 convolutional layer, which outputs JD-dimensional feature map, where $D = 8$ denotes discretized depth size. The feature map is reshaped to the dimension of $\mathbb{R}^{J\times D\times H_{\text{hand}}/32\times W_{\text{hand}}/32}$, which is a 3D heatmap of hand joints. Soft-argmax [10] extracts 2.5D joint coordinates $\mathbf{P}$ from the 3D heatmap.

MANO parameter regression. SHNet firsts reduces the channel dimension of $\mathbf{F}_{\text{hand}}$ from 2048 to 512 to reduce computational costs. Then, SHNet extracts joint features by performing a bilinear interpolation at the (x, y) position of the 512-dimensional feature map. The joint features contain essential articulation information about hand joints. Finally, a single linear layer outputs MANO pose parameter θ from a concatenation of the joint features with 2.5D joint coordinates. The pose parameter θ is initially estimated in the 6D rotational representation [15] and transformed to the axis-angle representation. The MANO shape parameter β and 3D global translation $\mathbf{g}$ are estimated from GAPed $\mathbf{F}_{\text{hand}}$.

G. Implementation details

PyTorch [8] is used for implementation. For the training, we use Adam optimizer [3] with a mini-batch size of 64. Data augmentations, including scaling, rotation, random horizontal flip, and color jittering, are performed during the training. The initial learning rate is set to 10^{-4} and reduced by a factor of 10 at the 4^{th} epoch. All other details will be available in our codes.

H. Limitations

Figure B shows failure case of ours, highlighted by yellow circles. It recovers the 3D mesh of the front view correctly; however, there are collisions in the 3D meshes of

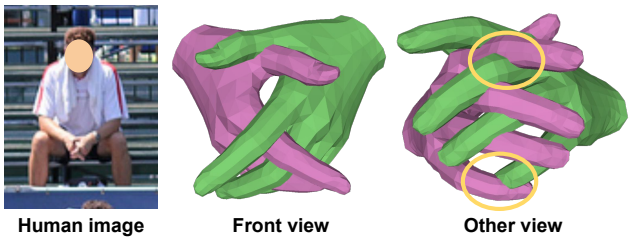

Figure B. Failure case of our InterWild.

other views. This can be addressed by introducing a collision avoidance loss function, similar to IHMR [9]. We leave this as our one of future works.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[2] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 2

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 3

[4] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022. 1, 2

[5] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1

[6] 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

[7] 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

[8] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3

[9] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3D reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, 2021. 4

[10] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 3

[11] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 1

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1

[13] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3D pose and shape reconstruction from single color image. In *ICCV*, 2021. 1, 2

[14] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. In *ICCV*, 2021. 1

[15] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[16] 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. 2