Interacting Two-Hand 3D Pose and Shape Reconstruction from Single Color Image "Supplementary Material"

Baowen Zhang[1,2] Yangang Wang[3] Xiaoming Deng[1,2*] Yinda Zhang[4*]
Ping Tan[5,6] Cuixia Ma[1,2] Hongan Wang[1,2]
[1]Institute of Software, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences
[3]Southeast University [4]Google [5]Simon Fraser University [6]Alibaba

In this supplementary document, we provide additional details on our method and more evaluations to complete the main paper. First, we give more details on our method, especially, relative translation prediction network, interacting hand pose and shape reconstruction and pose-aware feature extractor method. Then, we give more evaluations with comparisons to the state-of-the-art methods and ablation study.

1. Details of Our Method

1.1. Architecture for Feature Encoder

In this section, we introduce the detailed network architecture for feature encoder, which is used in Sec. 3.2 and Sec. 3.3 of our main submission.

We use backbone of ResNet-50 architecture to extract the feature of the input color image, and adopt the state-of-the-art interacting hand pose estimation network in [4] to predict 2.5D heatmap [2]. The output 2.5D heatmap ($21 \times 64 \times 64 \times 64$) is downsampled to $21 \times 16 \times 32 \times 32$ for computation efficiency. After reshaped to $((21 \times 16) \times 32 \times 32)$, the downsampled 2.5D heatmap is concatenated with the feature map extracted by ResNet-50, fed to two ResNet blocks to get the feature map $\mathbf{F} \in \mathbf{R}^{2048\times8\times8}$.

1.2. Architecture for Initial Reconstruction

In this section, we introduce the detailed network architecture for initial estimation in Sec. 3.3 of our main submission.

We use predicted 2.5D heatmap $\mathcal{H}_{2.5D} \in \mathbf{R}^{21\times64\times64\times64}$ to generate initial attention map. After conducting concatenation and max-pooling, we use nearest neighbor sampling to change the size of the attention map to 8×8 in order to match the size of feature map $\mathbf{F}$.

We conduct average pooling to get two 2048-dimensional vectors which are fed into parallel fully connected networks to predict the MANO parameters for each hand respectively. The parallel networks have the same architecture. Each network consists of 5 fully connected layers. The output size of the first four layers is 512, and the last layer outputs 58-dimensional parameter containing 10-dimensional shape parameter and 48-dimensional pose parameter.

Our relative transformation prediction network consists of two fully connected layers. The first layer is of the size 2048×1024 followed by Relu activation function, and the second layer is of the size 1024×4. It takes the average pooling feature of $\mathbf{F}$ as input, and predicts the translation $\Delta \in \mathbf{R}^3$ and the scale $s \in \mathbf{R}$ from the left hand hand coordinate system to the right hand coordinate system.

1.3. Generate 2D Heatmap from MANO Prediction

In this section, we introduce how to generate 2D heatmap from MANO prediction, which is used in the cascaded context-aware refinement in Sec. 3.3 of our main submission. Assuming that we have the predicted MANO parameters, we aim to generating refined attention map by projecting 3D hand joints to the image plane, in which the 3D hand joints are recovered from the predicted MANO parameters.

We use the weak perspective camera model in Eq. (1) for 3D joint projection [3], and get the camera parameters (i.e. the translation $\mathbf{t}$ and the scale s_c) for joint projection

$$\mathbf{x} = s_c \Pi(\mathbf{J}) + \mathbf{t} \tag{1}$$

where Π is an orthographic projection, $\mathbf{J}$ is the 3D joint position of a joint, and $\mathbf{x}$ is the 2D joint position.

We can get the camera parameters for joint projection by alignment or network prediction. In the alignment approach, we align the 3D joints with the predicted 2.5D heatmap (this approach is "fit camera parameters" in Fig. 8 of the main submission). Specifically, we obtain 2D joint position $\mathbf{x}$ and confidence c for each joint through 2.5D heatmap. Then the camera parameters can be calculated as

*indicates corresponding author

follows:

$$\arg\min_{s,\Delta} \sum_{i=1}^{K} ||c_i(s_c \Pi(\mathbf{J}_i) + \mathbf{t} - \hat{\mathbf{x}}_i)||^2 \qquad (2)$$

where K is the number of joints, $\mathbf{J}_i$ is the 3D joint position of the i-th joint, $\hat{\mathbf{x}}_i$ and c_i are the 2D joint position and confidence of the i-th joint obtained through 2.5D heatmap, respectively. The optimization in Eq. (2) can be solved by the least square method.

In the network prediction approach, we use a network to predict the camera parameters for joint projection (This approach is "predict camera parameters" in Fig. 8 of the main submission). This network uses the feature for MANO parameter regression as input, and then regress weak perspective camera parameters.

Since our cascaded network uses multiple feature maps of different resolutions to extract features, before calculating the camera parameters, we scale the position of 2D joints to match the resolution of the used feature map.

1.4. Architecture for Context-Aware Refinement

In this section, we elaborate on the network architecture for context-aware refinement in Sec 3.3 of our main submission.

In the context-aware refinement, we refine the initial hand reconstruction results with high-resolution features provided by the encoder. The refinement is achieved in a cascaded manner. In the cascaded block, the hand parameters of two hands are predicted by two parallel fully connected networks. Each network takes the predicted MANO parameters of two hands in the previous cascaded stage, the predicted relative transformation, and feature of a certain hand as input, and predicts the MANO parameters for the hand. Each network has 5 fully connected layers. The dimensions of each layer of the network are the same as those of the initial MANO parameter prediction network, except the first layer to match the size of input parameters. Our network contains 3 cascaded blocks. The input MANO parameters of each stage are the output of the previous stage, and the input left and right hand features are extracted from the feature maps with sizes of 8, 16, and 32 in the encoder.

2. More Experiments

2.1. More Ablation Study Results

To the effect of context-aware refinement, we show more comparisons between the model that feeds MANO parameters of single hand to cascaded blocks, and our full model. As shown in Fig. 1, predicting MANO parameters conditioned on previously predicted MANO parameters of two hands helps to improve the accuracy of hand geometry details.

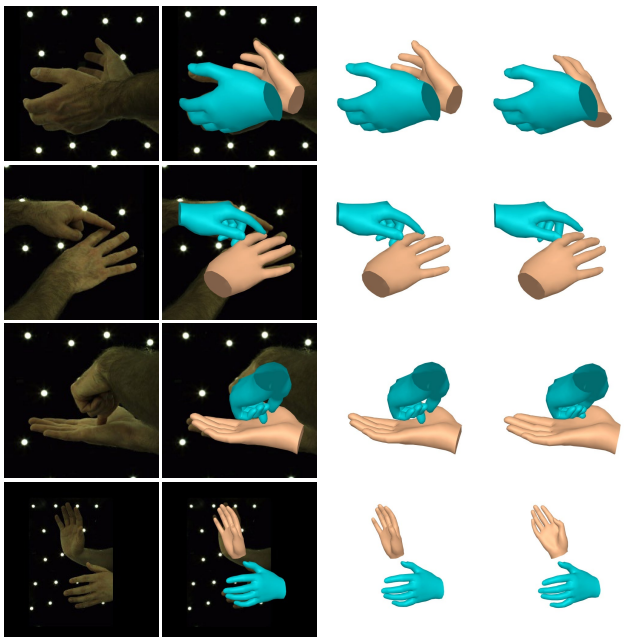

Figure 1. Qualitative study of context-aware refinement. We compare our full model and the model consisting cascaded blocks using MANO parameters of single hand as input.

2.2. Comparison to State-of-the-art Methods

In this section, we show more comparison results to state-of-the methods on 3D hand shape recovery from single color image.

Fig. 2 shows several comparison results of our method and the methods proposed by Boukhayma *et al*. [1] and Zhou *et al*. [5]. As shown in this figure, the compared methods often fail when hands are interacting, while our method can recover interacting hand shapes even under severe occlusions. With two separated hands, Zhou *et al*. [5] can recover hand shapes. Conceptually, the method proposed in [5] fails for interacting hand shape reconstruction, because interacting hands have higher degree of freedom, more deformations, and more occlusions than separated hands, and all these issues make interacting hand reconstruction very hard. Fig. 3 shows more qualitative results of our interacting hand reconstruction method.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and P. Torr. 3d hand shape and pose from images in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10835–10844, 2019. 2, 4

[2] U. Iqbal, P. Molchanov, Thomas Breuel, Juergen Gall, and J. Kautz. Hand pose estimation via latent 2.5d heatmap regression. *Proceedings of the European Conference on Computer Vision*, pages 118–134, 2018. 1

[3] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1

[4] Gyeongsik Moon, Shoou-I Yu, H. Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *Proceedings of the European Conference on Computer Vision*, 2020. 1

[5] Yuxiao Zhou, Marc Habermann, Weipeng Xu, I. Habibie, Christian. Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5345–5354, 2020. 2, 4

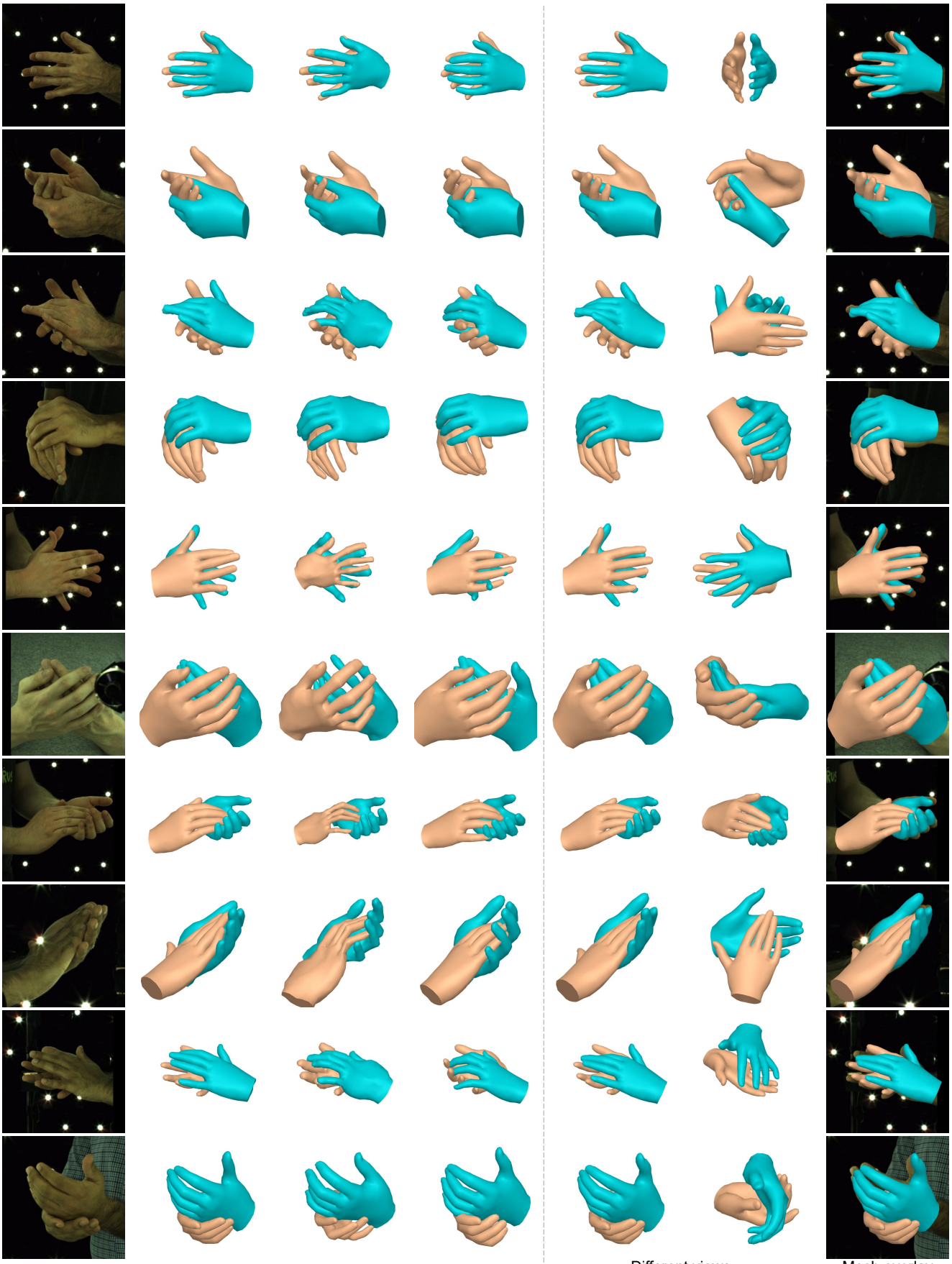

Figure 2. Qualitative comparison of the interacting hand reconstruction with our method and the state-of-the-art single hand reconstruction methods Boukhayma *et al*. [1] and Zhou *et al*. [5] on InterHand2.6M.

Figure 3. Qualitative results of the interacting hand reconstruction with our method on InterHand2.6M (row1-row5) and Haggling dataset (row6-row8).