Supplementary Material: Self-supervised Transfer Learning for Hand Mesh Recovery from Binocular Images

Zheng Chen Sihan Wang Yi Sun* Xiaohong Ma
Dalian University of Technology, China
{czheng,sylbia_w}@mail.dlut.edu.cn,{lslwf,maxh}@dlut.edu.cn

1. Detailed architecture of the encoder-decoder for hand prior learning.

We design an encoder-decoder model for pre-learning the hand prior from existing hand dataset. We use the ResNet-50 [1] as our encoder where the fully connected layer is removed and use the mesh regressor in [2] as our decoder which outputs lixel-based 1D heatmap of the hand mesh vertex.

2. Detailed structure of the mesh regressor in our self-supervised model.

As mentioned in the manuscript, the mesh regressor in our self-supervised model (binocular) only outputs the u and v coordinates of the hand mesh, which is slightly different from the mesh regressor in [2]. Thus, we illustrate its detailed structure in Figure 1.

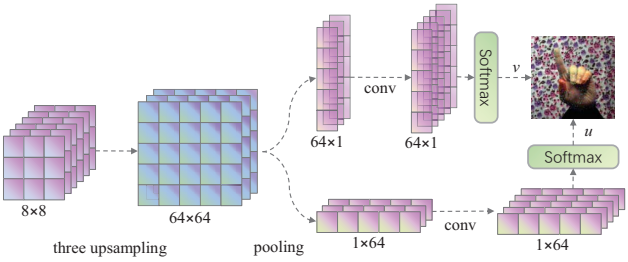

Figure 1. Structure of the mesh regressor in our self-supervised model.

The output feature of the encoder is firstly passed through three upsampling layers. The feature map size is increased from 8×8 to 64×64 and the channel is decreased from 2048 to 256. Then we do average pooling along the row and column, respectively. The obtained feature maps are passed through a convolutional layer with kernel size of 1×1 to get the lixel-based 1D heatmaps of the hand mesh vertex. The channel of the output 1D heatmap is 778, which is the same with the number of the MANO [3] mesh vertex. Finally, the heatmaps are passed through a soft-argmax [4] layer to get 2D coordinates (u, v) of the mesh vertices.

*Corresponding author.

3. Training details.

We crop image patches of the hand region with size of 256×256 from the original images. The cropped image patches are normalized to $[0, 1.0]$ and then used as input without data augmentation. For the STB dataset[5], we crop the image patches with the ground truth root joint of the middle finger as the center. For our collected dataset, we manually label the 2D root joint of the middle finger and use it as the center to crop the image. In the future, we will apply the advanced object detection model to automatically detect hand from the original images.

We pre-train the encoder-decoder end-to-end on the FreiHAND dataset[6] with batch size of 32. We use Adam optimizer and set the initial learning rate to 1e-5. We train the model for 25 epochs and the learning rate is decayed by a factor of 10 after the 17th and 21st epoch.

For training our self-supervised model on the STB dataset and our collected real hand dataset, we set batch size to 4. The model weight is initialized with the pre-trained encoder-decoder. We use Adam optimizer and the initial learning rate is set to 1e-5. We train the model end-to-end for 13 epochs and decay the learning rate by a factor of 10 after the eighth epoch and tenth epoch.

4. More qualitative results.

We provide more qualitative results on the STB dataset and our real hand dataset in the form of video, shown as the MP4 files in the same folder. From the video we can see, our model can output accurate hand pose even in the case of large pose and illumination variation. But there are some distorted hand mesh estimations and the discussions have been given in our manuscript. We will focus on these problems in the future.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] 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. *arXiv preprint arXiv:2008.03713*, 2020.

[3] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017.

[4] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018.

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

[6] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 813–822, 2019.