SemiHand: Semi-supervised Hand Pose Estimation with Consistency Supplementary Material

Linlin Yang[1,2], Shicheng Chen[1], Angela Yao[1]
[1]National University of Singapore, Singapore
[2]University of Bonn, Germany

In the supplementary material, we present

- the details of 2.5D regression (See Sec. 3.1) in Sec. A,
- the details of HSD dataset (See Sec 4.2) in Sec. B,
- the cross-modal consistency (See Sec. 3.3) in Sec. C,
- the ablation study for label correction and sample selection (See Sec. 3.2) in Sec. D,
- the comparison of our proposed pose registration (See Sec. 3.2) with a state-of-the-art pose prior module in Sec. E,
- the discussion of synthetic dataset in Sec. F,
- more qualitative results in Sec. G.

Note that all the notation and abbreviations here are consistent with the main manuscript.

A. 2.5D Regression for Hand Pose

A standard method for 3D pose estimation of the hand is to use a 2.5D representation $(\mathbf{uv}, \mathbf{d})$ and integrate over the 2D heatmap h_{uv} and latent depth map h_d. Specifically, the output of the network f are the heatmaps h_{uv}, h_d; the 2.5D components $\mathbf{uv}$ and $\mathbf{d}$ are estimated as:

$$\begin{aligned} \mathbf{uv} &= \sum_{g \in \Omega} g \otimes \text{softmax}(\beta h_{uv})(g), \\ \mathbf{d} &= \sum_{g \in \Omega} h_d(g) \otimes \text{softmax}(\beta h_{uv})(g), \end{aligned} \tag{1}$$

where Ω is the set of all pixel locations, $\otimes$ is the element-wise product, β is the learnable parameter and the function $\text{softmax}(\cdot)$ serves as normalization. We show the pipeline of 2.5D regression in Fig. 1. For more details we refer the reader to the paper [4].

In addition, we use the same framework f to estimate the hand mask $\mathbf{w}$ as an auxiliary modality, to encourage the predicted poses to be consistent with the hand masks (See Sec. 3.3).

B. HSD dataset

We apply a semi-supervised annotation framework based on the work [7]. We use 4 RealSense D415 cameras in different views to record 4 RGB-D sequences. Each sequence is performed by one actor and contains 20K frames. We use the first two sequences for training and others for testing. The labeling is done on the multi-view depth images and then applied to the RGB images. Specifically, we use synthetic depth maps as input for warmup training and then train with both manually labelled multi-view depth maps and unlabelled multi-view depth maps iteratively. For multi-view labelling, we initialize the 3D poses with the predictions and project the 3D poses to the image coordinates for each view. We refine the 2D pose of each view manually, and perform a synchronized update of 3D pose and other views' 2D poses until the 2D poses of all the view are reasonable. In each iteration, after the training, we visualize the predictions of unlabelled data and manually revise the worst 100frames based on the model fitting energy. Next, we move three quarters of the revised data into the labelled training data and the remaining revised data into the testing data. We train and evaluate the model iteratively until the model achieves a desired performance on testing data. Specifically, our annotation ends up with a mean EPE of less than 6 mm on testing data, which considered to be comparable to human-annotated labels. We show some examples of the dataset in Fig. 2.

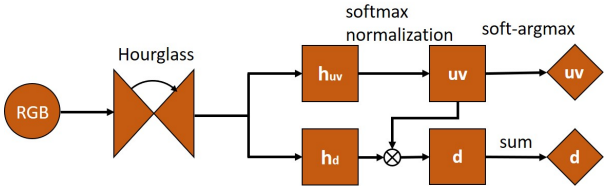

Figure 1: The pipeline of 2.5D regression.

C. Cross-modal Consistency

In this section, we outline the details of the circle hand model. Our goal is to estimate the model-fitting energy given 2D poses and hand masks. We start off with the assumption that we already have the 2D poses and hand

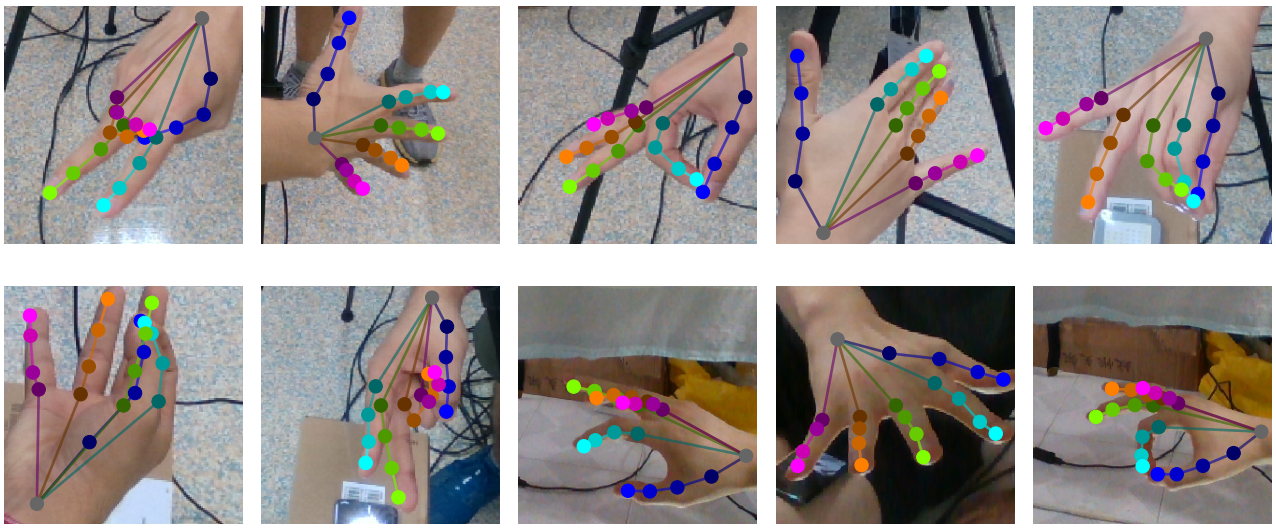

Figure 2: Examples of HSD dataset with ground-truth 2D poses.

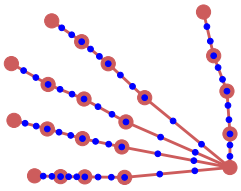

Figure 3: The circle centers based on the hand joints. Indianred dots denote the hand joint locations and blue dots denote the circle centers of our proposed circle hand model.

Figure 4: Each triplet left to right shows groundtruth, circle hand mask (with 2D pose in blue, circles in red), and difference.

masks. For the fingers, we then trisect each bone evenly and get ten keypoints for each finger. Excluding the fingertips, there are 45 keypoints remaining, which serve as centers of the circles. The palm is represented by 5 "bones" radiating from the wrist keypoint to the base of each finger. We similarly trisect these bone evenly and get two keypoints, totally 55 circle centers for our hand circle model. We show the joints (indianred dots) and the circle centers (blue dots) in Fig. 3. Also, Fig. 4 shows two examples of our circle hand masks.

As for the radius, we minimize our proposed energy loss with the model-to-data term and the data-to-model term as described in Sec. 3.3, and learn the radius via a five-layer MLP with BN and leaky ReLU. Our circle hand model is pre-trained on RHD dataset and fixed for our experiments. Given centers and radius, we approximate hand masks by setting the pixel value to 1 if the pixel is inside a circle. The hand mask $\mathbf{w}$ is estimated as below:

$$\mathbf{w}(g) = \begin{cases} 1 & \text{if } \min\limits_{i\in[0,54]} (||g - c^i||_2 - r^i) < 0, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

where $g \in \Omega$. The rendered hand mask can be found in Fig. 4 middle of the main manuscript.

D. Label Correction and Sample Selection

In this section, we first explore the effect of τ for diversity augmentation, the confidence threshold defined in Sec. 3.2 and Eq. 8 of the main manuscript. In Fig. 5, we show the STB testing performance when fine-tuning pre-trained model on STB training set with different τ. When $\tau = 0$, which corresponds to rejecting all samples, our Semi-Hand degrades to baseline with consistency training and achieves worst performance. When the value of τ is small, probably not enough pseudo-labels are selected for training. As the value of τ increases, the accuracy first increases and then slightly decreases. This indicates the fine-tuning benefits from incorporating of enough high-confidence pseudo-labels for training at first. But, as τ further increases, more noisy pseudo-labels are selected for training and this hurts the training procedure. When $\tau = +\infty$, this corresponds to accepting all samples. Experimentally, we get the best

performance when $\tau = 1.5$. Also, the comparison between $\tau = 0$ and $\tau = +\infty$ indicates that having noisy labels may help the training procedure more than not using these labels at all.

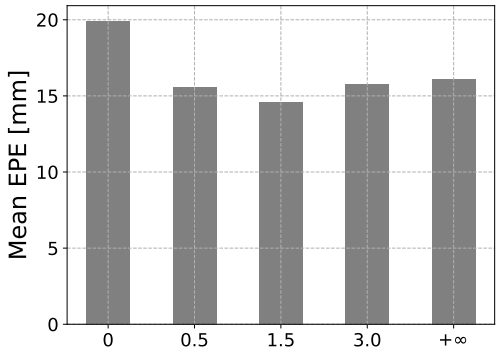

Figure 5: Comparison of different τ for SemiHand.

$\tau = 1.5$	w/o	w/
Mean EPE [mm]	21.70	16.30
$\tau = +\infty$	w/o	w/
Mean EPE [mm]	24.31	21.28

Table 1: Mean EPE of baseline with pseudo-labeling with and without label correction.

We also explore the effectiveness of label correction and sample selection for baseline with pseudo-labeling for diversity augmentation. We set τ to 1.5 and $+\infty$ respectively, and verify the pseudo-labeling with and without label correction. After fine-tuning on STB training set, the STB testing performance is shown in Tab. 1. We can see that pseudo-labeling with label correction outperforms the case without label correction by a large margin.

Meanwhile, if we set τ to $+\infty$, the performance of pseudo-labeling without label correction is even worse than that of baseline (24.31 mm vs. 23.83 mm), which indicates naive pseudo-labeling are even detrimental to learning. However, when only selecting high-confidence pseudo-labels (*i.e.* , $\tau = 1.5$) for diversity augmentation, we prevent model degradation and achieve better performance than that of baseline (21.70 mm vs. 23.83 mm). This also verifies the necessity of pseudo-labels with high-confidence and label correction.

E. Pose Registration

In this section, we compare our pose registration with a hand pose prior module, IKNet [9, 8]. Given a hand template, IKNet tackles joint rotation estimation as an inverse kinematics problem. It proposes to regress the pose parameters of MANO from 3D poses with a neural network. The network is pre-trained on large amount of MoCap data.

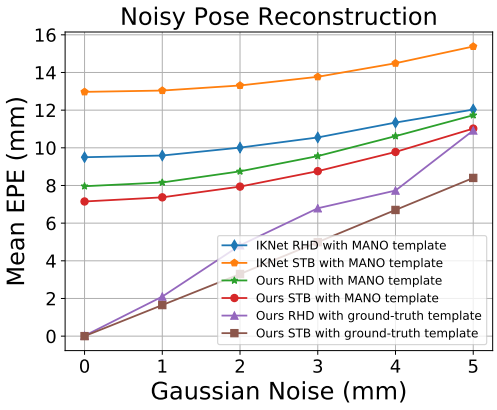

Figure 6: Comparison of our proposed pose registration with IKNet for noisy pose reconstruction.

The fundamental problem with this approach however, is that MANO serves as a hand shape model. The pose parameters of MANO is based the rotation of mesh vertices. Depending on the actual pose, the bone lengths of the hand may actually vary. Also, the MANO template can not match the ground-truth template perfectly. These two lead to a surprisingly significant reconstruction error in the range of 9-15 mm for RHD/STB hand as shown in Fig. 6. Furthermore, as a data-driven method, IKNet may not generalize well to unseen data.

In contrast, we tackle joint rotation estimation as a registration problem as done in [7]. We propose a greedy approximation based on the hand's kinematic chain. This does not require any training and is more flexible in that it can be based on arbitrary templates. IKNet, however, is limited to using only the MANO template and requires large amounts of training data. More importantly, our proposed greedy approximation avoids the accumulation of end point errors and achieve more accurate pose reconstruction. This makes the approximation method ideal for pose registration and correction.

To verify the effectiveness of our proposed pose registration, we show the results of noisy pose reconstruction in Fig. 6. We use noisy 3D poses as input, *i.e.* , ground-truth 3D poses corrupted with Gaussian noise on three axes and compare the denoising performance of IKNet versus our proposed pose registration. For IKNet, we use the released model [8] which trained on SIK-1M [8] and fine-tuned on RHD training set. Our pose registration is optimization based and does not require any training or fine-tuning. To see the effect of training data for IKNet, we evaluate the modules on two different evaluation sets, RHD testing set and STB testing set. We use as input the same (MANO) template from IKNet to compare the two modules

Figure 7: Examples of different synthetic dataset. From top row to bottom row: RHD, MANO+NN and MANO+Blender.

in a fair way. As shown in Fig. 6, when using ground-truth 3D poses as input, we can see the performance of IKNet on RHD testing set significantly outperforms that on STB testing set (9.5 mm vs. 12.97 mm) while our pose registration without training achieve more close performance (7.96 mm vs 7.15 mm). Also. on both RHD and STB testing set, ours outperforms IKNet. This indicates that our pose registration achieve better performance and generalization. As the noise increases, the accuracy of both modules decrease. Note that our pose registration outperform IKNet whatever the value of noise is. We also compare different templates for pose registration. We can see that pose registration with ground-truth templates outperform that with MANO templates when adding small Gaussian noise. As the noise increases, the performance of pose registration with the two templates becomes close.

F. Synthetic Data

In our main paper, we used only the RHD dataset for pre-training. In this section we investigate the impact that the quality of the synthetic data may have. Besides the RHD dataset, we also pre-train our SemiHand with two different synthesis data. As [1], we synthesize data based on MANO [6] and a neural renderer [5] to generate data online (MANO+NN dataset). Additionally, we synthesis data based on [3] and use MANO and blender [2] to synthesize 40k images (MANO+blender dataset). Compared to the RHD dataset, MANO+NN and MANO+blender are of a much lower quality. They have limited poses, a limited range of hand skin colors and use a fixed lighting scheme for rendering. Also, the procedure of placing foreground hands into background images is naive. The data in MANO+NN lack the wrist, which we find to be important for the hand segmentation. We show the examples of three datasets in Fig. 7. We can see that RHD is the most photo-realistic hand dataset and the image quality of MANO+blender is marginally better than that of MANO+NN.

We pre-train models on those three synthetic datasets respectively and then fine-tune the models on STB training set and DO. We show the mean EPE of STB testing set and DO in Fig. 9 and 10. We can see that training with more realistic synthetic hand data tend to achieve better performance. Regardless of fine-tuning, using RHD as synthetic data always achieves best performance and using MANO+Blender achieves the second best performance. They both outperform using MANO+NN as synthetic hand data.

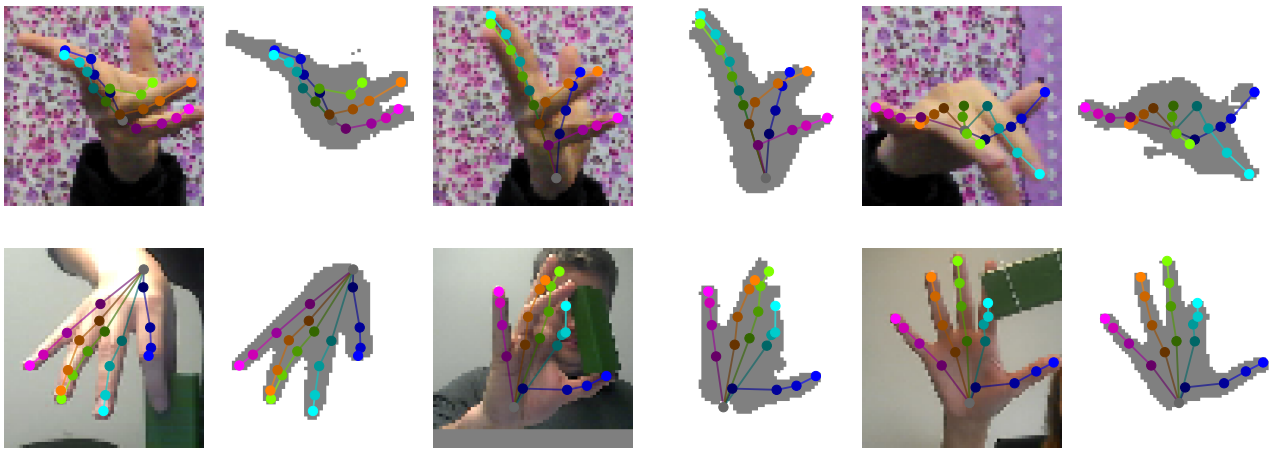

Figure 8: Failure cases with predicted 2D poses and masks.

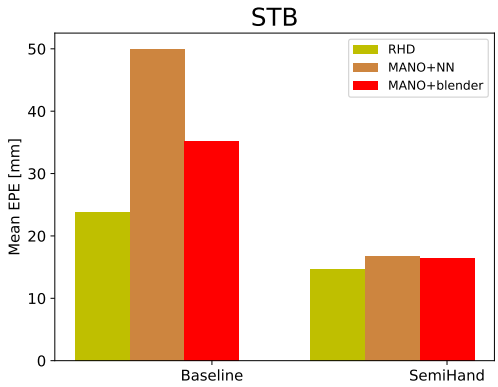

Figure 9: Comparison of different synthetic datasets for pre-training on STB.

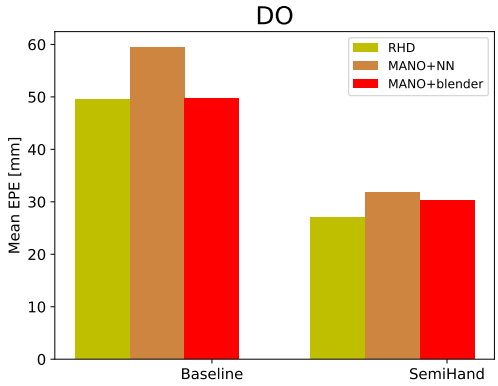

Figure 10: Comparison of different synthetic datasets for pre-training on DO.

However, interestingly, using MANO+NN as synthetic hand data achieves the most significant improvement after fine-tuning. Even its performance is still the worst, the gap with using other two synthetic data is incremental after fine-tuning. Take STB for example, the gap between RHD and MANO+NN is decreased from 26.19 mm to 2.1 mm after fine-tuning as shown in Fig. 9. This indicates our SemiHand still works on STB even pre-training with low quality synthetic data.

G. Qualitative Results

In this section, we show the qualitative results and some failure cases. In Fig. 11, we visualize the predictions of the baseline and our SemiHand as well as the ground-truth. We can see that the predicted poses of baseline may be located in the background. With consistency training and pseudo-labeling, the refined predictions tend to be centered on the fingers and match the shape of hand.

Also, we show some failure cases in Fig. 12 and Fig. 8. In most cases, we find the failures are highly related to the poor predictions of hand mask. As the hand is occluded or in challenging lighting conditions, the model fails to outline the complete hand silhouette and the hand keypoints tend to locate only inside the shrunken hand silhouette as shown in Fig. 8. This encourages us to further explore the multi-task framework with cross-modal consistency.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 4

[2] Blender Online Community. Blender - a 3d modelling and rendering package. http://www.blender.org. 4

[3] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learn-

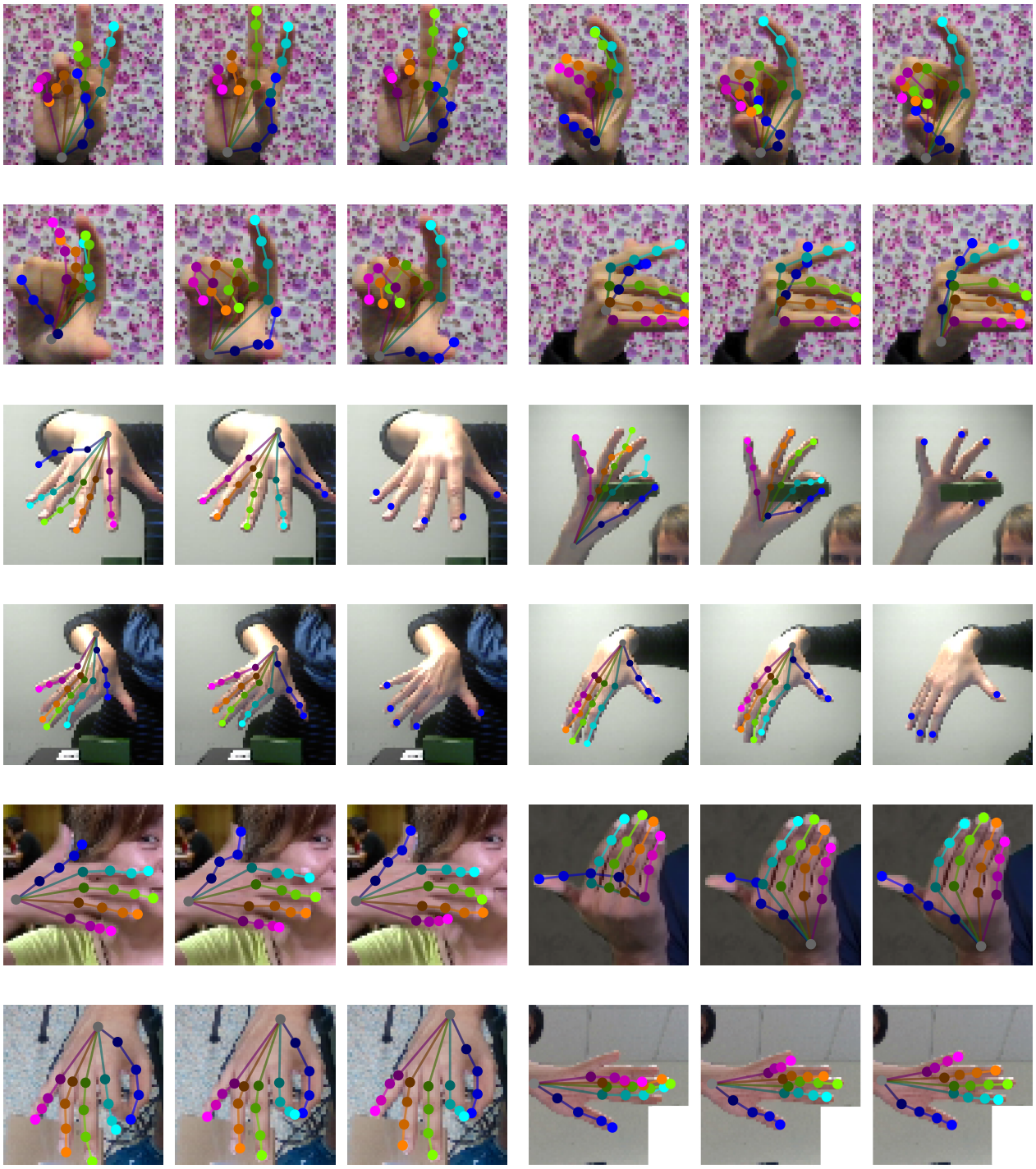

Figure 11: Qualitative results. For each trio, the left most column corresponds to the prediction of baseline, the second column corresponds to the prediction of our SemiHand and the right most column corresponds to the ground-truth.

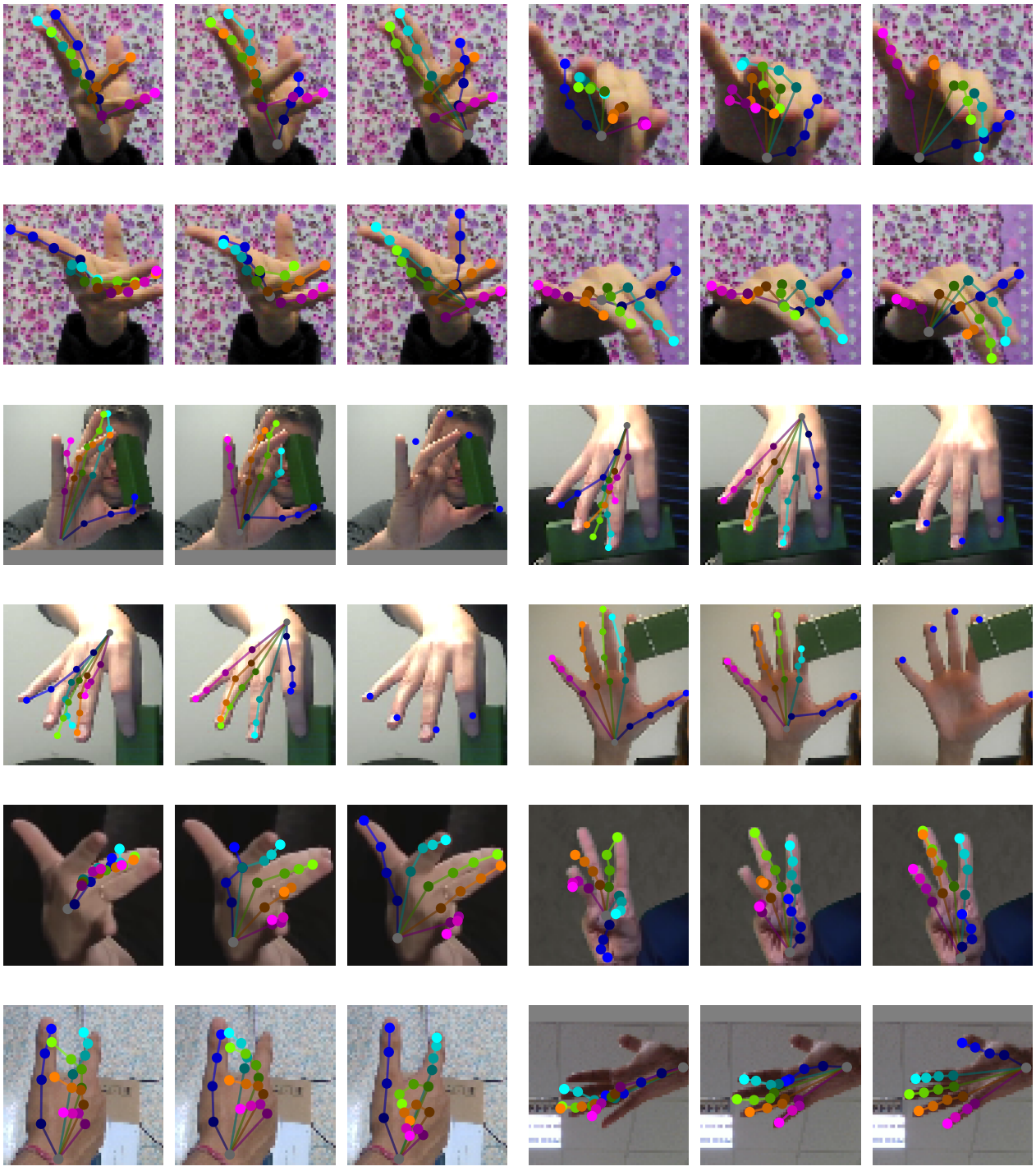

Figure 12: Failure cases. For each trio, the left most column corresponds to the prediction of baseline, the second column corresponds to the prediction of our SemiHand and the right most column corresponds to the ground-truth.

ing joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 4

[4] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 1

[5] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, pages 3907–3916, 2018. 4

[6] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6):1–17, 2017. 4

[7] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dual grid net: hand mesh vertex regression from single depth maps. In *ECCV*, pages 442–459. Springer, 2020. 1, 3

[8] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. *arXiv preprint arXiv:2008.05079*, 2020. 3

[9] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, pages 5346–5355, 2020. 3