Semi-Supervised 3D Hand-Object Poses Estimation with Interactions in Time Supplementary Material

Shaowei Liu[*1] Hanwen Jiang[*1] Jiarui Xu[1] Sifei Liu[2] Xiaolong Wang[1]
[1] UC San Diego [2] NVIDIA
Project Page: https://stevenlsw.github.io/Semi-Hand-Object

Table of Content

Appendix A: Implementation Details

The network architecture is illustrated in Table 1. The input RGB image is sent to the shared feature encoder with Res-50-FPN as the backbone. We utilize ROIAlign [4] to crop the hand and the object features $\mathcal{F}_h$ and $\mathcal{F}_o$ with channel dimensions 256 and spatial resolution 32 from the most fine-grained level P2 features of the FPN.

Contextual Reasoning Module The contextual reasoning (CR) module takes the object features as the query and the hand-object intersecting regions as the key. After passing the input features by two 1-D convolutions to reduce their channel size into 128, we reshape them into $\mathbb{R}^{1024\times128}$ and $\mathbb{R}^{128\times1024}$ for the query and the key respectively. Then we perform matrix multiplication to get a $\mathbb{R}^{1024\times1024}$ pairwise similarity matrix and use the softmax operation on the top to obtain the synergy map. We apply another 1-D convolution to the input key features and do the reshape to get the key embedding $\mathbb{R}^{1024\times128}$. We perform another matrix multiplication between the synergy map and the key embedding to get the intermediate representation of $\mathbb{R}^{1024\times128}$. The target representation is computed by lifting the intermediate representation's channel dimension to the original ones of 256. The output is further added to the original query features by residual connection.

Hand Decoder The 2D heatmaps $\mathcal{H}$ in the joints localization network takes the form of $\mathcal{H} \in \mathbb{R}^{32\times32\times N_h}$ where each channel corresponds to one joint. $N_h = 21$ is the number of joints. Then the 2D joint positions $\mathcal{J}^{2D} \in \mathcal{R}^{N_h\times2}$ can be calculated by the weighted sum of heatmap values and corresponding 2D pixel coordinates as $\mathcal{J}_i^{2D} = \sum p \cdot H_i(p)$ for each joint i, where p represents the pixel coordinates in the heatmap. The mesh regression network predicts the MANO pose parameters $\theta \in \mathbb{R}^{48}$ and shape parameters $\beta \in \mathbb{R}^{10}$ by first using the residual blocks combined with max-pooling layers to downsample the hand features, and three fully-connected layers afterward to regress the target MANO parameters θ and β.

Stage	Configuration	Output
0	Input image	$512 \times 512 \times 3$
	Feature extraction	
1	Res-50-FPN (P_2) [6]	$128 \times 128 \times 256$
1	Hand-RoiAlign [4]	$32 \times 32 \times 256$
1	Object-RoiAlign [4]	$32 \times 32 \times 256$
	Contextual reasoning	
2	enhanced object feature	$32 \times 32 \times 256$
	Hand Decoder	
3	Hourglass Module [7]	$32 \times 32 \times 21$
3	4 Residaul Block [5]	$2 \times 2 \times 512$
3	Flatten	2048
3	3 FC Layers	58
	Object Decoder	
4	4 Shared 2D Convolution	$32 \times 32 \times 256$
4	2D Convolution of localization	$32 \times 32 \times 21 \times 2$
4	2D Convolution of confidence	$32 \times 32 \times 21 \times 1$

Table 1. Network architecture and configurations of the proposed model. FC layers denote the fully-connected layers.

Object Decoder The two-stream object decoder has two outputs separately, the control points 2D coordinates in $\mathbb{R}^{32\times32\times N_o\times2}$ and the corresponding 1D confidence values in $\mathbb{R}^{32\times32\times N_o\times1}$ for all grids, where $N_o = 21$ is the number of control points. To predict the i-th control point pixel coordinates from grid g, the network estimates a offset $v_{g,i}$ between the grid's pixel location p_g and the target control point position t_i. Then the residual error $\delta_{g,i}$ between the prediction in grid g to the target control point i is $\delta_{g,i} = p_g + v_{g,i} - t_i$. The confidence score $c_{g,i}$ is obtained

models	Hand mean distance (↓)	Object mean distance (↓)
Tekin *et. al* [9]	**15.8**	24.9
Hasson *et. al* [3]	18.0	22.3
Ours-sup	16.5	**20.4**

Table 2. Hand pose and object pose performance between state-of-the-art methods [9, 3] and the proposed method on FPHA [2] dataset. **sup** means our model is trained under the supervised learning phase. The error is in mm.

methods	Hand AUC(↑) Joint	Mesh	F-score(↑) F@5	F@15
Boukhayma *et al.* [1]	73.1	74.1	43.2	90.9
Radosavovic *et al.* [8]	73.9	74.7	43.0	91.9
Ours	**74.8**	**75.9**	**45.3**	**92.9**

Table 3. Comparison against different semi-supervised hand pose estimation approaches on the cross-domain FPHA dataset.

by apply the sigmoid function at the top of the second stream output.

Appendix B: Performance on FPHA Dataset with Fully Supervision

We report our method's hand and object pose estimation performance on FPHA [2] dataset. Considering that there is a significant appearance change between the FPHA dataset and other datasets because of the introduction of visible markers used for annotation, we do not use the FPHA dataset to conduct semi-supervised learning. We compare our model's performance trained under supervised learning against other state-of-the-art approaches [9, 3], the results are shown in Table 2. As can be seen from the table, our method has the lowest object estimation error and outperforms other approaches by a large margin, as well as a comparable hand pose estimation performance against [9] and much better than [3]. Our model with only supervised learning could achieve the best overall performance on the FPHA dataset, which demonstrates the superiority of the joint learning framework and the effectiveness of the contextual reasoning module.

Appendix C: Cross-domain Semi-supervised Learning Results on FPHA Dataset

We compare our method with other semi-supervised learning methods [1, 8] on the FPHA dataset as shown in Table. 3. Specifically, we implemented [8] by removing the spatial-temporal constraints in our method for generating pseudo labels. Note that all the methods are not trained with the FPHA dataset. Our method performs significantly better than previous approaches in this cross-domain setting. Our method earns such benefit from using real-world hand-object videos that contain a wide range of domains with different colors, hand scales, lighting, objects, and backgrounds. Training on these videos covers most domains in different test datasets. Thus we can improve both the performance and generalization across different domains and datasets.

Appendix D: Qualitative Generalization Results on FPHA and Freihand Dataset

We visualize our proposed model's generalization results on the FPHA dataset [2] and Freihand dataset [**?**]. We compare our method against the baseline which doesn't include the semi-supervised learning phase and the contextual reasoning model. The qualitative results of the FPHA dataset and the Freihand dataset can be seen in Figure 1 and Figure 2 respectively. As can be seen from these figures, by introducing more training data that covers diverse hand poses and subjects in the wild, our model can obtain much better generalization performance across different backgrounds, viewpoints, and subjects.

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.

[2] 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. *CVPR*, pages 409–419, 2018.

[3] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. *CVPR*, 2020.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *CVPR*, pages 2961–2969, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[8] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, pages 4119–4128, 2018.

[9] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019.

Figure 1. Qualitative comparisons of the proposed method against the baseline on FPHA dataset [2]. The top row shows the predictions of the baseline, while the bottom row shows the predictions of our proposed model.

Figure 2. Qualitative comparisons of the proposed method against the baseline on Freihand dataset [**?**]. The top row shows the predictions of the baseline, while the bottom row shows the predictions of our proposed model.