Supplemental Materials: Fine-grained Egocentric Hand-object Segmentation: Dataset, Model, and Applications

Lingzhi Zhang*[1], Shenghao Zhou*[1], Simon Stent[2], and Jianbo Shi[1]

[1] University of Pennsylvania
[2] Toyota Research Institute

In this supplementary materials, we first describe the following details of this work: 1). details of video frame collection; 2). training details of segmentation network; 3). how we chose the quantity of data augmentation; 4). the performance of using 100-DOH followed by PointRend. Please note that we also provide results of per-frame hand-object segmentations and "see-through hand" application in egocentric videos. Please check our ".mp4" file in the supplementary materials for details.

1 Dataset

Our labeled dataset consists of video frames sparsely sampled from multiple sources, including 7,458 frames from Ego4D [2], 2,212 frames from EPIC-KITCHEN [1], 806 frames from THU-READ [6], and 350 frames of our own collected egocentric videos with people playing Escape Room. In Ego4D videos [2], we use the videos from the hand-object interaction challenges, which consist around 1,000 videos. Among these videos, we first sparsely sample one frame per three seconds, and then use 100-DOH detector [5] to filter out the frames that actually contain hand-object interaction. In order to make our labeled data as meaningful as possible, we ask humans to manually select 7,458 frames with diverse and interesting hand-object interactions among these extracted frames. Similarly, we uniformly sample one frame per three seconds in EPIC-KITCHEN [1] videos across 37 participants, then filter out frames that contain hand-object, and finally manually select the interesting frames to label. The THU-READ [6] dataset consists of short video clips of 40 classes of hand-object interaction, which we uniformly sample a few images in each category. Finally, for our own collected GoPro videos, we also sparsely sample one frame per three seconds, and manually filter out frames to label.

2 Training Details of Segmentation Networks

In this section, we discuss the details of our segmentation network training. We use ResNet-18 backbone [3] and HRNet head [8] as our base model for all of our experiments. For each experiment, we train the network for 80,000 iterations with SGD optimizer with batch size of 8, learning rate of 0.01, momentum of 0.9, and

weight decay of 0.0005. We use random flip and photometric distortion(random brightness, random constrast change, etc.) as our data augmentation techniques on top of our proposed context-aware data augmentation.

3 Choosing the Augmented Data Quantity

Our proposed Context-aware Compositional Data Augmentation (CCDA) technique generates the composite data before the training starts, in an offline fashion. A key question is how much data we should generate for the augmented dataset. To this end, we run an experiment to evaluate how the segmentation performance would vary when gradually increasing the number of composite images, as shown in Fig. 1. We found that the hand and object IoU reaches to the maximum IoU, when augmenting 16K and 8K composite images respectively, on the YouTube testset.

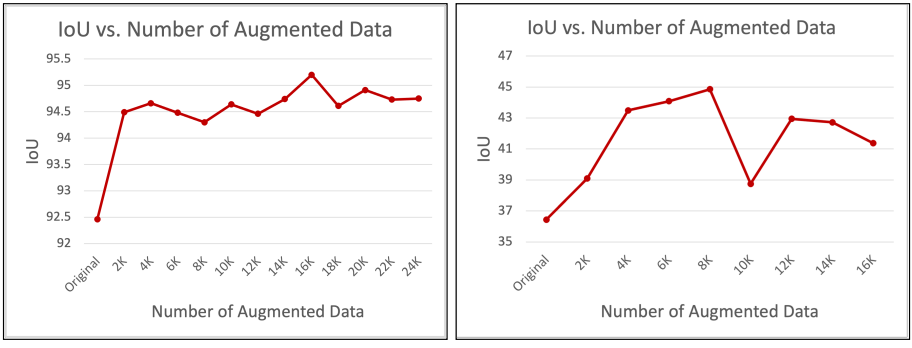

Fig. 1. Averaged hand and object IoU scores vs. the number of augmented data on the left and right, respectively.

4 PointRend vs. BoxInst for 100-DOH

Since 100-DOH [5] only predicts the bounding of hands and objects, we compare two ways to further convert the bounding boxes to segmentation. In the first way, we use 100-DOH [5] detector to generates pseudo labels of hand-object, and train BoxInst [7] model to segment hand and objects. In the second way, we use 100-DOH [5] detector to localize the hand-object bounding boxes, and use PointRend [4] to segment the masks. We find that the 100-DOH [5] + PointRend [4] has better performance on the left/right hand segmentation, and 100-DOH [5] + BoxInst [7] has better performance on the binary hand segmentation, as shown in Table 1 and 2, respectively. Nevertheless, the model trained on our dataset surpass the performance both approaches by an obvious margin.

Datasets	mIoU	mPrec	mRec	mF1
100-DOH[5] + BoxInst[7]	36.30/37.51	50.06/61.63	56.91/48.94	53.27/54.55
100-DOH[5] + PointRend [4]	61.83/62.72	76.17/78.41	76.66/75.8	76.41/77.09
Ours	79.73/82.17	84.26/90.38	93.68/90.04	88.72/90.21

Table 1. Left/Right Hand Segmentation.

Datasets	mIoU	mPrec	mRec	mF1
100-DOH[5] + BoxInst[7]	69.50	84.80	79.67	82.00
100-DOH[5] + PointRend [4]	63.62	78.39	77.14	77.76
Ours	85.45	90.11	94.30	92.15

Table 2. Binary Hand Segmentation.

We also conduct an experiment to use 100-DOH [5] + PointRend [4] to compute the interacting object segmentation masks. However, we observe that such inference pipeline often completely miss or misclassify the interacting objects segmentation, or in other words has low recall, as shown in Fig. 2. Quantatively, we find that the averaged object IoU of 100-DOH [5] + PointRend [4] is only 12.24, which is significantly lower than ours on the YouTube testset.

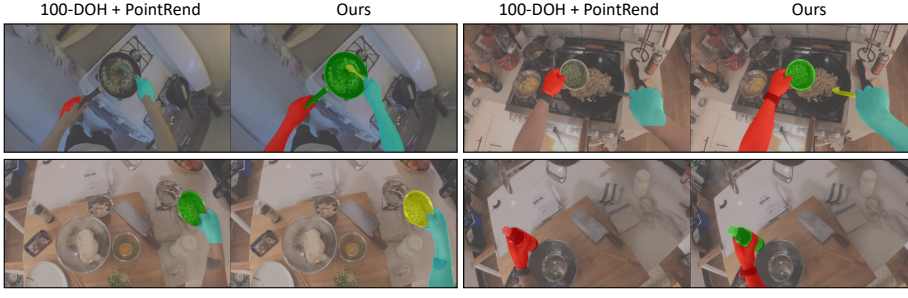

Fig. 2. Visual comparison of hand-object segmentaiton between 100-DOH [5] + PointRend [4] and ours. The color are coded as follows: red → left hand, cyan → right hand, green → left-hand object, yellow → right-hand object, orange → two-hand object.

5 More Segmentation Label Visualization

We show more visualizations of our hand-object segmentation labels, across different data sources, in Fig .3.

6 More Video Results

In the "video_2152.mp4" file, we show more results on hand-object segmentation and the application of "seeing through the hand" in the egocentric videos.

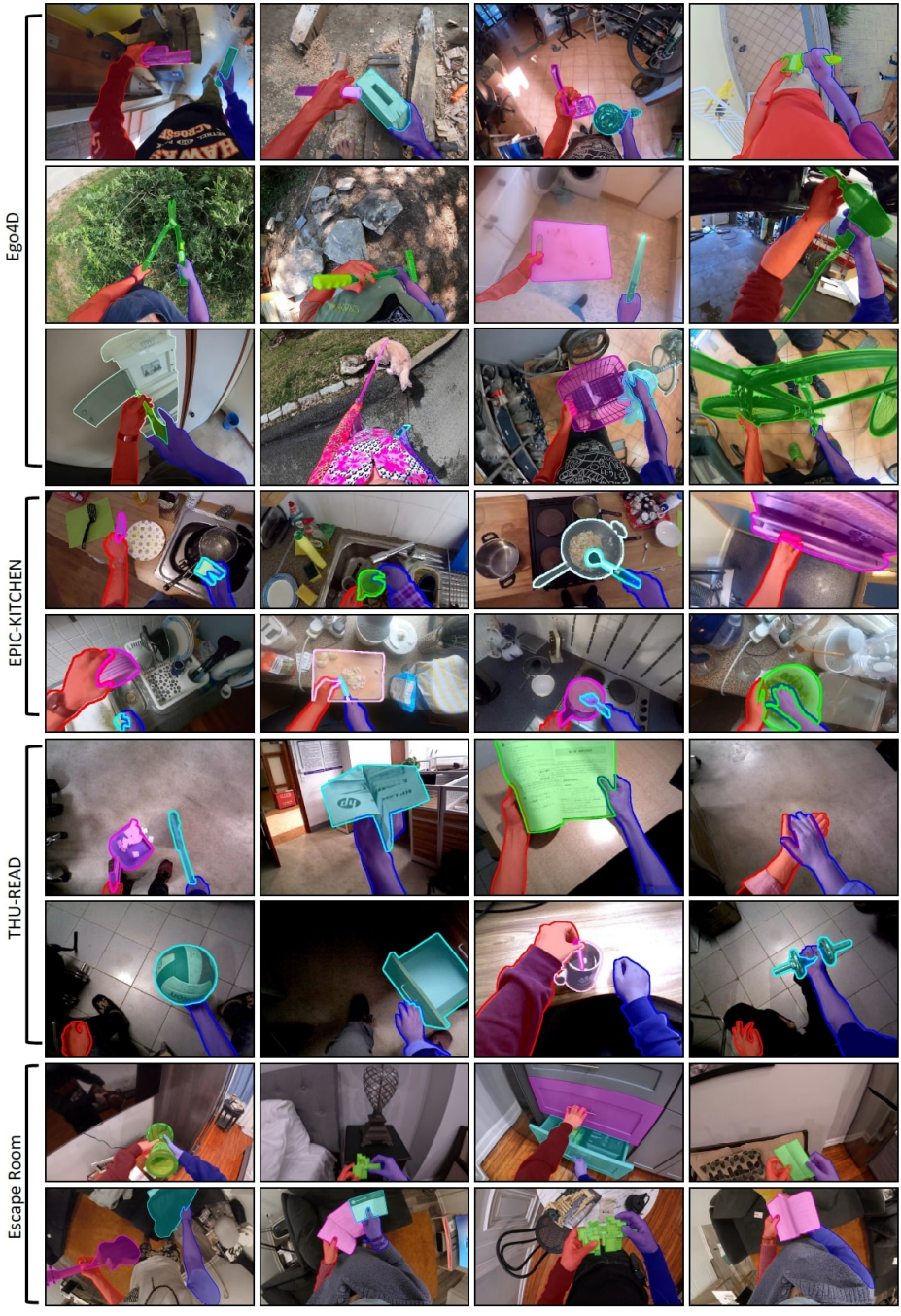

Fig. 3. More visual demonstration of our hand-object segmentation labels.

References

1. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 720–736 (2018) 1
2. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. arXiv preprint arXiv:2110.07058 (2021) 1
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
4. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020) 2, 3
5. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9869–9878 (2020) 1, 2, 3
6. Tang, Y., Tian, Y., Lu, J., Feng, J., Zhou, J.: Action recognition in rgb-d egocentric videos. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3410–3414. IEEE (2017) 1
7. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5443–5452 (2021) 2, 3
8. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020) 1