ACR: Attention Collaboration-based Regressor for Arbitrary Two-Hand Reconstruction **Supplementary Material**

Zhengdi Yu[1,2], Shaoli Huang[1,*], Chen Fang[1], Toby P. Breckon [2], Jue Wang
[1]Tencent AI Lab [2]Durham University
{zhengdiyu,shaolihuang,fcfang}@tencent.com toby.breckon@durham.com arphid@gmail.com

1. Introduction

In this supplementary material, to showcase the superiority of our method and its potential for real-world application, we first demonstrate more qualitative results, including a video of a real-time demo of ACR (Sec.2) and reconstruction results on in-the-wild images or video (Sec.3). Then in Sec.4, we present more detailed ablation study results to better understand the performance of our method design.

2. Real-time Demo of ACR

We implement a real-time two-hand reconstruction demo based on our proposed method and an ordinary webcam. Due to simplicity, our approach can run in real-time on a laptop with an RTX 2080 GPU. We provide the results of the demo in Fig.1 and a video ***acr_live_demo.mp4*** for more details. Our method can produce high-quality reconstruction results and effectively handle various inputs such as interacting hands, truncated hands, and hand-object interaction. Besides, our algorithm bypasses the requirement of a hand detector or constraint inputs, while IntagHand [4] requires two-hand in a pre-defined region. These advantages are significant for advancing hand-reconstruction technology in real-world applications.

3. In-the-wild Qualitative Comparisons

In addition to the extra results on the InterHand2.6M dataset in Fig. 5, this section also provides more qualitative results on in-the-wild datasets or web videos (watch video ***acr_in_the_wild.mp4*** for more detail). First, we compare our method with the previous state-of-the-art, IntagHand[4] on RGB2Hands [6] and Ego2Hand datasets [1]. We also provide a qualitative comparison of two approaches on web videos (obtained from YouTube). Since IntagHand can only deal with well-cropped hand regions, we acquire the result by employing a hand detector to crop out the hand region from an in-the-wild image while keeping the aspect ratio. For visualization, we project the rendering results of IntagHand back to the **original image** instead of the cropped image.

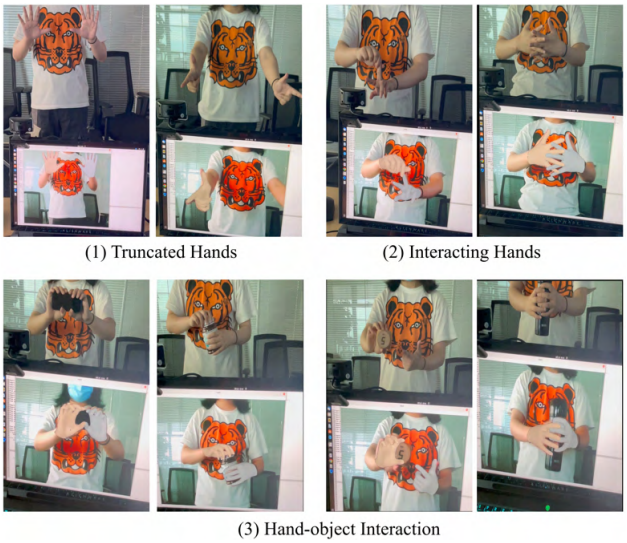

Figure 1. Results of the ACR real-time demo (see video ***acr_live_demo.mp4*** for more detail.) Our method produces high-quality results on a live video stream from a cheap webcam.

Our method performs better than IntagHand under nearly all cases, particularly in challenging cases such as truncated hands, severe occlusion, and hand-object interaction. Fig 3 and 4 show some representative examples. For instance, if one hand is severely impeded by the other hand (like Fig.3(h)), our approach yields more reasonable results than IntagHand. Another case that IntagHand usually fails to handle is truncation (i.e., 4(f), hand pars truncated by image boundary). In contrast, our method built on part-level representation learning is not relatively sensitive to this situation. Moreover, our approach also performs much better than IntagHand on hand-object interaction data. This is be-

*Corresponding author.

cause IntagHand is very sensitive to external occlusion, as it may treat the object occlusion as interacting hand occlusion, resulting in failure estimation.

More interestingly, IntagHand mostly fails to reconstruct two hands on the ego-view dataset (as shown in Fig.2). One possible reason is its GCN, and transformer-based attention mechanism overly relies on two-hand interacting dependency to reason about two-hand reconstruction. At the same time, the two hands are primarily separate and coupled with slight object occlusion. Nevertheless, our method consistently performs well on this dataset thanks to its independent features for each hand and its powerful collaborative representation.

4. Detailed ablation study

In this section, we provide further information about our network and ablation studies for 1) aggregation method of the **G**lobal representation (G) and **P**art representation (P) and **C**ross-hand representation (C). And 2) supervision method of hand part segmentation branch.

Ablation study of aggregation method To explore a proper way to aggregate the global representation and part representation while maintaining their own advantages, we have conducted different kinds of aggregation methods ($mode$ in the Tab 1), where $offset$ means a simple summation and $concat$ means feature aggregation illustrated in the methodology section. We found that aggregating the representations by concatenation always yields better performance under different cases. Thus, we report the final results on both InterHand2.6M[5] and FreiHand[8] and claim the state-of-the-art in this manner.

Ablation study of supervision In Tab 1, in addition to decoupling each module of our network, we also explored 1) different ways to aggregate the global representation and part-based representation. We first tried to remove the supervision of L_{seg} from our network to see if the part segmentation can work as implicit attention guidance, and vice versa, if the part segmentation can explicitly work as an attention mask. Finally, we use a hybrid training strategy for our network, which only supervise the L_{seg} for the first two epoch. This strategy significantly speeds up the training process and yields the best performance. Please note that it is unnecessary to have a superior segmentation mask to learn a part-based representation. On the contrary, it is reasonable to have a slightly lower mIoU, as this would expand the attention area of each part to let the network focus on the visible parts and aggregate helpful features for the missing part by a reasonable deduction.

5. Details of Our Method

This section provides further details of our methods including the details of data generation, the mechanism of our method, and evaluation metrics.

Global representation: To guide our network to gain a better global representation, we adopt a scale-adaptive Gaussian kernel for our center map generation. As shown in Fig. 6. To generate a scale adaptive Gausssian heatmap, we adopt a Gaussian kernel according to the size of the bounding box. The bounding box is roughly computed by the maximum and minimum values of the visible keypoints of the hand. Specifically, the center-based attention is represented by a Gaussian map where its kernel size K is computed according to the hand box. Let d be the diagonal length of the box, W_b be the width, then the kernel size to generate the supervision map can be computed by

$$k = k_{min} + \delta_k \times (\frac{d}{\sqrt{2}W})^2, \tag{1}$$

where k_{min} stands for the minimum kernel size and we would adjust kernel size depending on difference hand scale. δ is the adjusting factor to control the expanding size of the kernel size. In all of our experiments, we set $k_{min} = 2$ and $\delta_k = 7$ for all of our experiments.

Part representation: Our ground truth segmentation map is rendered by utilizing the ground truth MANO mesh and camera parameter provided by InterHand or FreiHand with a neural renderer [3]. Thus we only supervise the part segmentation brach when the ground truth MANO parameter or the ground truth segmentation is available. Our segmentation map is represented as in the right of Fig. 6. The background class is 0 (black part). The labels for left hand parts are from $1 \sim 16$ and right hand labels are $17 \sim 32$.

6. Evaluation Metrics

In this work, we use four metrics to evaluate the reconstruction quality of our method, which are MPJPE, MPVPE, PA-MPJPE, and PA-MPVPE. Please note that all of the evaluation metrics are performed after **root joint (middle MCP joint) alignment** of each hand. It is worth noting that our concurrent work and prior arts [4, 7] typically need to recover the mesh to ground truth scale by using **extra ground truth information** during evaluation, which is not fair for previous methods. However, for fair comparison with them, we also both the originally 'correct' protocol as the previous methods and their protocol of using extra ground truth scale and box as in Table 1 of our main paper.

MPJPE measures the mean per joint position error in millimeters, which is the mean Euclidean distance between the predicted 3D joint locations to ground truth 3D joint locations after root joint alignment.

MPVPE measures the mean per vertex position error in millimeters. The average Euclidean distance between the hand mesh predictions and the ground truth MANO hand mesh after aligning them by root joint.

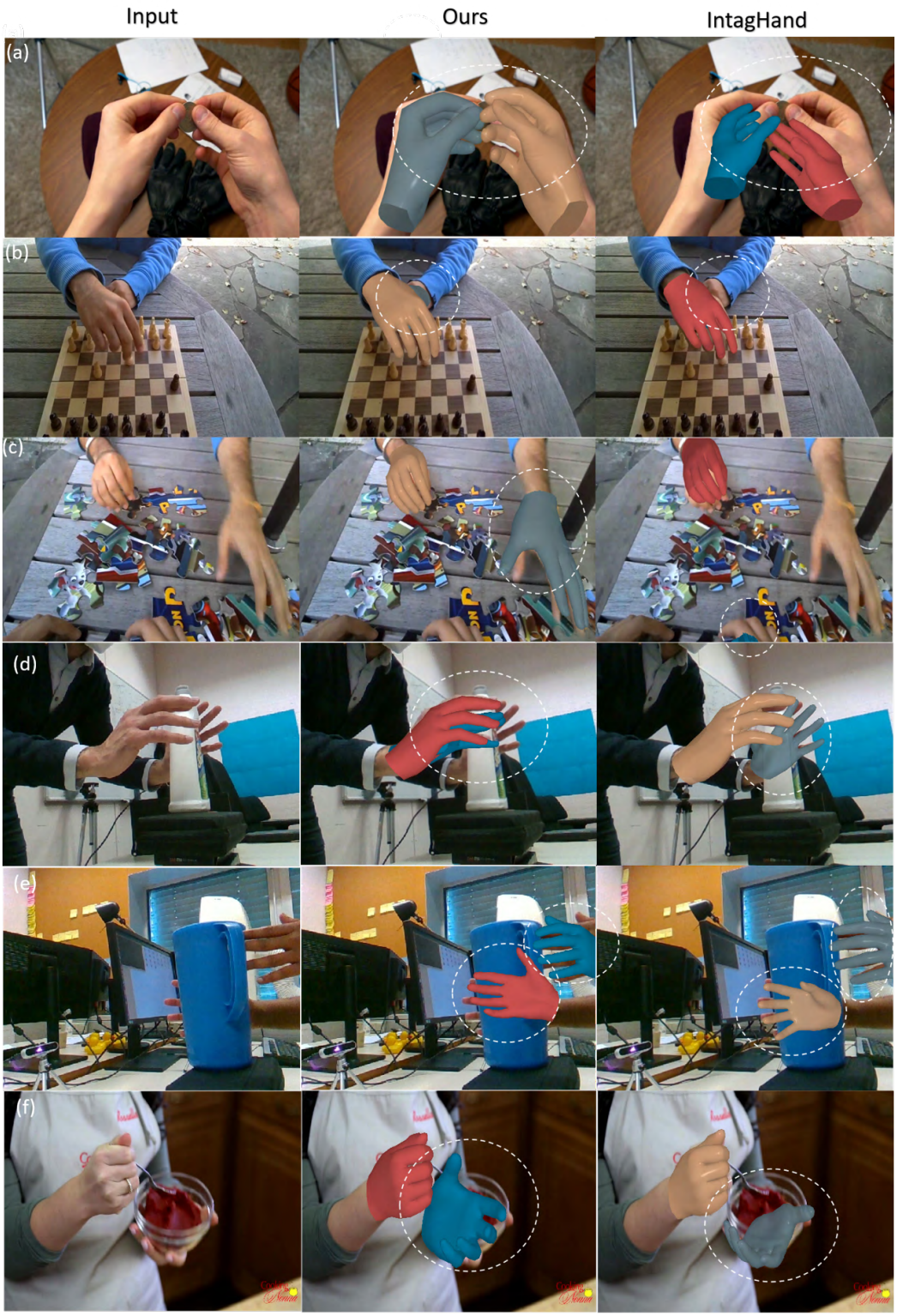

Figure 2. Qualitative comparison results on ego-view data. images in (b)(c) are selected from RGB2Hands benchmark[6]. (d)(e)(f) are selected from H_2O-3D dataset [2]

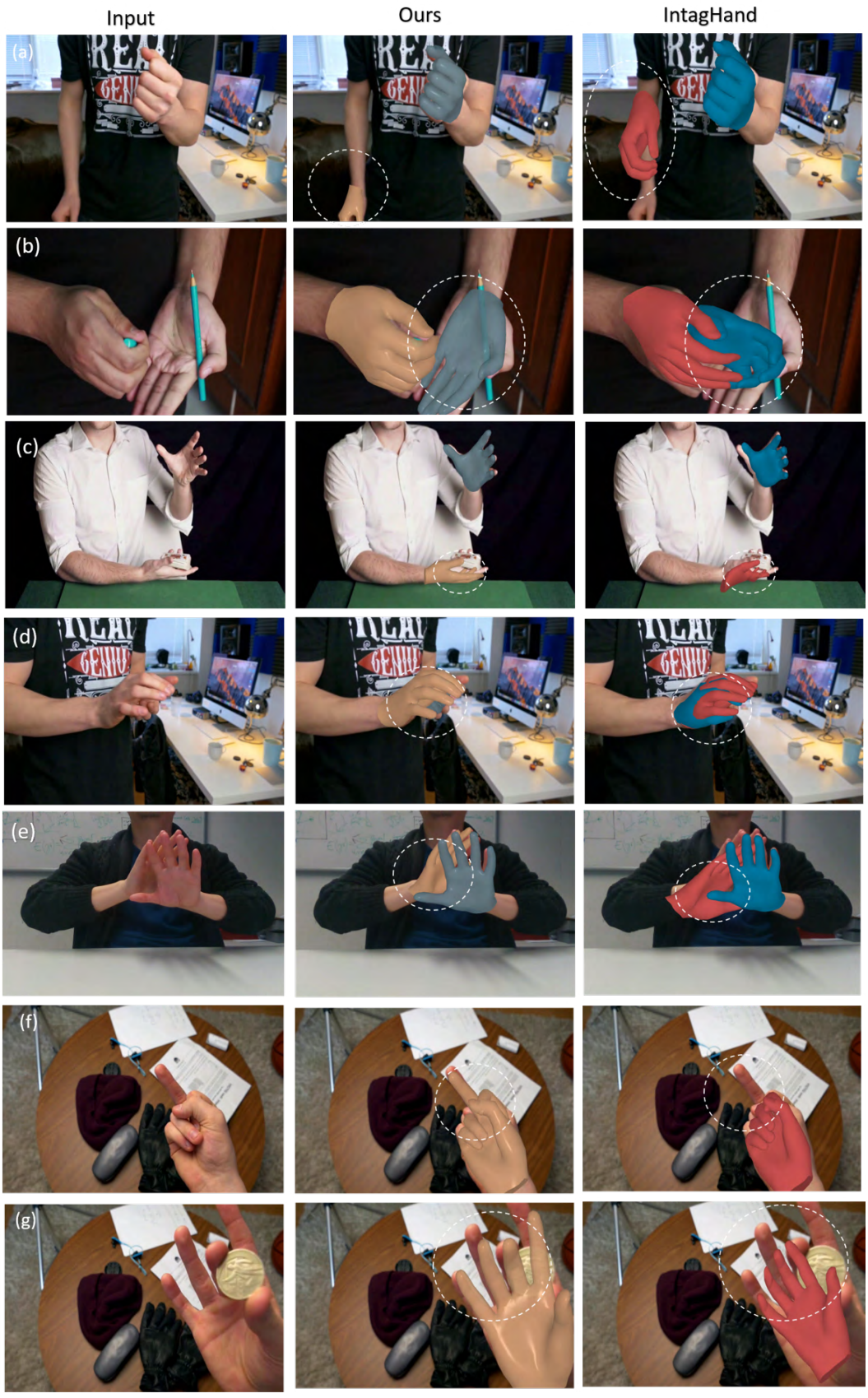

Figure 3. Interacting hand and single hand reconstruction. Here, the images in (e) are selected from RGB2Hands benchmark[6]. The others are from web videos.

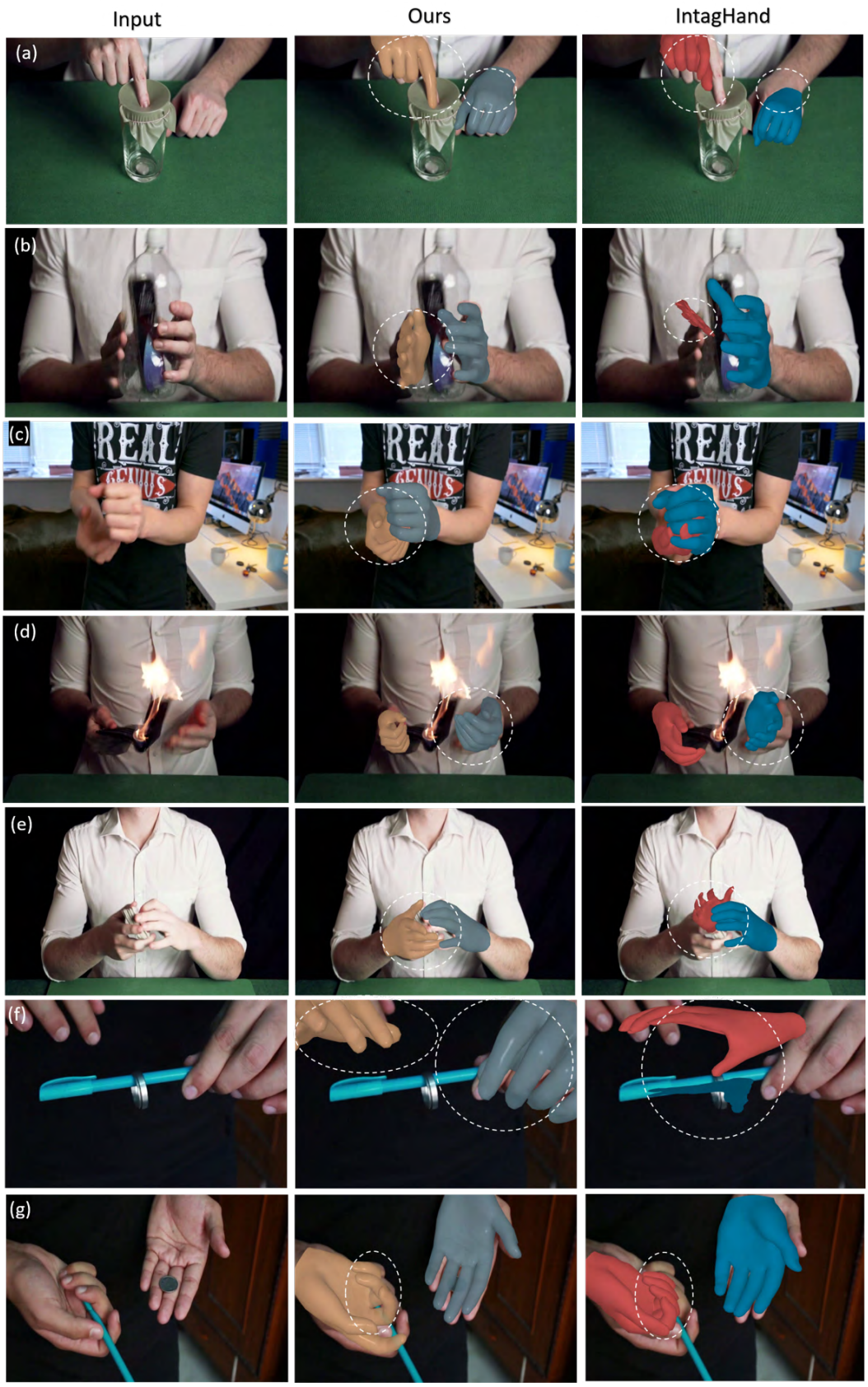

Figure 4. Hand-object interaction on web videos (watch video ***acr_in_the_wild.mp4*** for more detail).

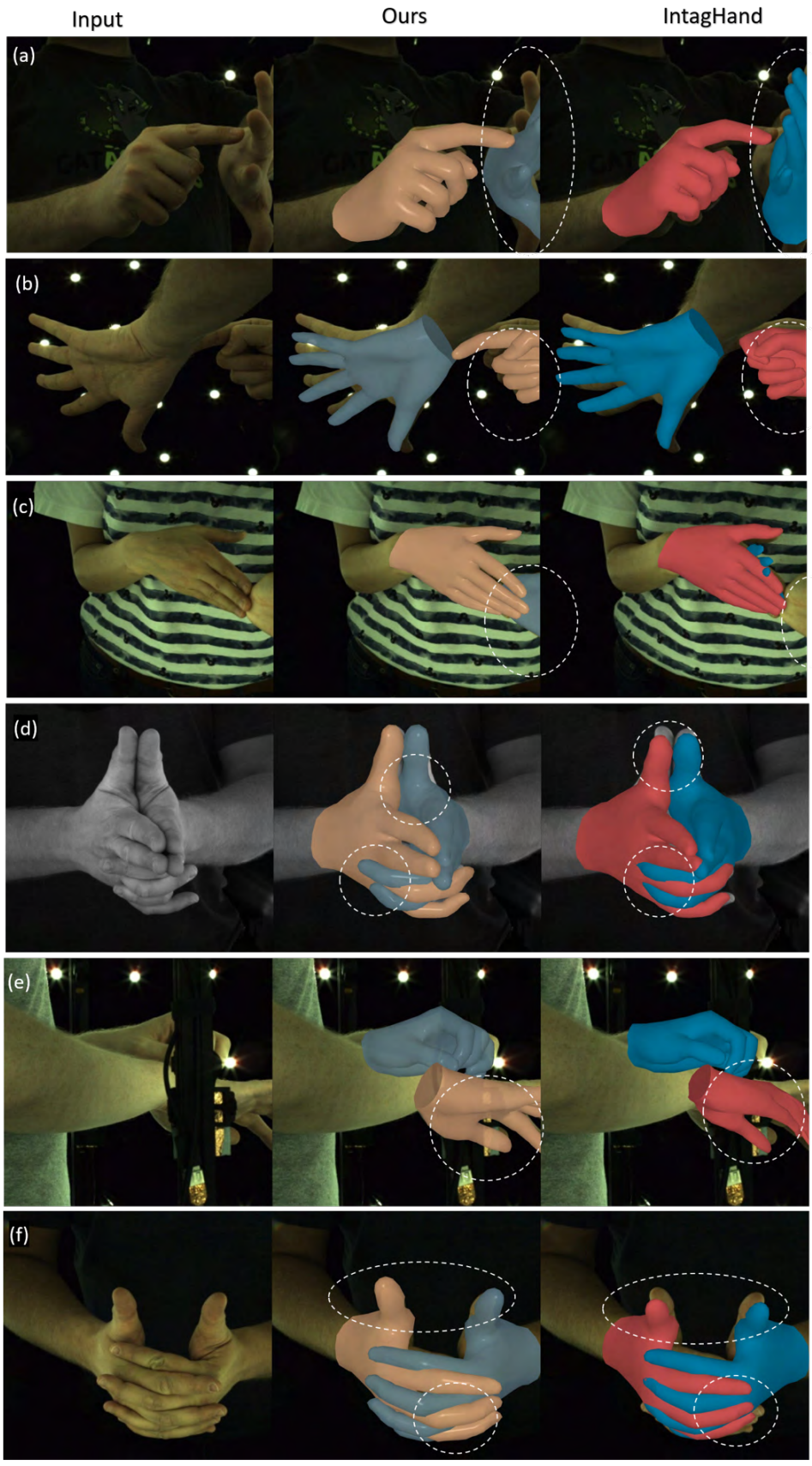

Figure 5. More qualitative results on InterHand2.6M dataset.

	mode	supervision	IH MPJPE	SH MPJPE	PAMPJPE
G	-	-	10.35	8.65	6.41
G+P	Concat	Full	9.71	7.05	5.54
P	-	Full	10.03	7.48	5.68
G+P	Offset	Full	9.82	7.13	5.60
G+P	Concat	Hybrid	9.69	6.87	5.49
P	-	Hybrid	9.76	7.26	5.59
G+P	Offset	Hybrid	9.49	6.91	5.50
G+P	Concat	Unsup.	9.73	7.05	5.54
P	-	Unsup.	10.05	7.52	5.67
G+P	Offset	Unsup.	9.87	7.17	5.61
G+**C**+P	Offset	Hybrid	9.28	7.01	5.38
G+**C**+P	Concat	Hybrid	**9.08**	**6.85**	**5.21**

Table 1. Ablation study of different aggregation methods of part-global representation learning, cross-hand-attention prior module, and part-segmentation branch supervision method. *mode* means the aggregation method of part-global representation. *supervision* suggests different supervision strategies for the art segmentation branch. G, P, and C stand separately for global representation, part-based representation, and cross-hand attention prior module.

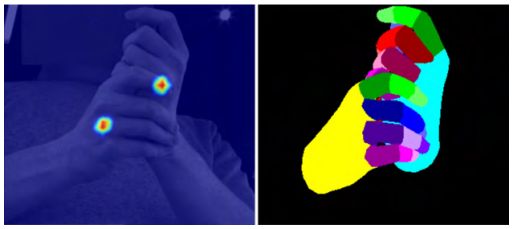

Figure 6. Illustration of hand center and hand part segmentation.

PA-MPJPE is the MPJPE after Procrustes alignment. By Procrustes aligning the predictions and the ground truth mesh, it eliminates the effects of translation, rotation and translation and focuses on the reconstruction accuracy.

PA-MPVPE is the MPVPE after Procrustes alignment, which eliminates the effects of translation, rotation, and translation

7. Discussion

In terms of monocular full-body capturing, hand capture is always the most difficult part due to frequent occlusion, truncation, and fast movements. Typical methods follow a pipeline to crop the single hand by an external hand detector. However, it has been explored that interacting hands can not be well recovered separately by treating them as single hand reconstruction methods. Thus, all the existing interacting hand reconstruction methods naturally adopt a naive strategy to crop the interacting two hands in one box and extend the output to two hands with some tailored mutual fusion part such as transformer-based [4] module. We believe this kind of cropping strategy is an ill-posed pipeline because it is only limited to very closely interacting hands without generalization ability. On the contrary, our method is the only one that can be plugged into any kind of hand pose estimation task. It is also worth noting that our network can be very easy to be extended to multiple hand mode by leveraging the center representation.

References

[1] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957, 2015. 1

[2] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 3

[3] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 2

[4] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 2, 7

[5] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[6] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt.

Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 1, 3, 4

[7] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, and H. Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11354–11363, 2021. 2

[8] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2