Sequential Voting with Relational Box Fields for Active Object Detection Supplementary Material

Qichen Fu Xingyu Liu Kris M. Kitani

Carnegie Mellon University

A. Overview

In this document, we provide additional implementation and experimental details, as well as qualitative results and analysis. We present the details of the self-attention layer in Appendix B.1 and the confidence calculation in Appendix B.2. We validate that the proposed method is more robust to detect active objects under occlusion in Appendix C. We show the effect of reinforcement learning with additional ablation study in Appendix D. We illustrate additional qualitative results and visualizations in Appendix E. The inference running time is presented in Appendix F.

B. Implementation Details

B.1. Self-attention Layer

As described in Sec. 3.3 of the main paper, inside the image feature extractor, we use a self-attention layer between the encoder and the decoder to further exploit the synergy between hands and objects. The architecture of the self-attention layer is illustrated in Fig. 1. The self-attention layer takes the image feature $\mathcal{F}_{\text{deep}}$ from the encoder as input and computes the query, key, and value embeddings (Q, K, and V) from $\mathcal{F}_{\text{deep}}$ using learnable embedding matrices W_q, W_k and W_v. Then the relationships between every spatial location in the feature map are computed using query Q and key K, which is used as the weight to average V. Finally, a two-layer MLP with layer normalization [1] is applied as

$$\begin{aligned} &Q = W_q\mathcal{F}_{\text{deep}}, K = W_k\mathcal{F}_{\text{deep}}, V = W_v\mathcal{F}_{\text{deep}} \\ &\mathcal{F}^{+}_{\text{deep}} = \text{MLP}(\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V) \end{aligned} \tag{1}$$

where d_k is the feature dimension of the key and $\mathcal{F}^{+}_{\text{deep}}$ is the feature map after applying self-attention, which is forwarded to the decoder. Empirically, we find marginal improvement from positional encoding, so we omit it for simplicity.

In order to avoid exhaustive computation, we set $d_k = 256$, and reduce the feature dimension of $\mathcal{F}_{\text{deep}}$ from 2048 to 512 by a convolutional layer with a kernel size of 1×1.

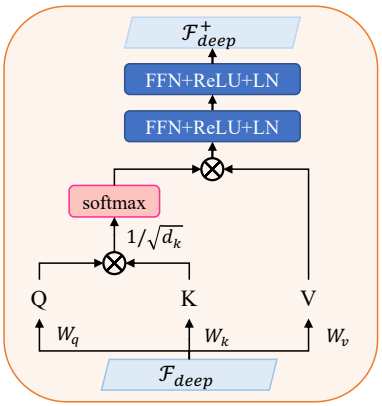

Figure 1. The architecture of the self-attention layer

Following [4] , we use 8 attention heads to address multiple relations between hands and objects.

B.2. Confidence Calculation

By definition, an active object must be manipulated by a human hand. We first predict a contact score $s^{\text{contact}}_{b^h}$ representing the probability that a given hand b^h is manipulating an object. Besides, we predict a object probability score $s^{\text{obj}}_{\hat{b}^o}$ for the final object estimation $\hat{b}^o$ of the given hand. To compute $s^{\text{contact}}_{b^h}$ and $s^{\text{obj}}_{\hat{b}^o}$, we use the average of confidence scores inside predicted hand-to-object $\hat{F}^{ho}$ and object refinement $\hat{F}^{oo}$ box fields as

$$s^{\text{contact}}_{b^h} = \frac{\sum_{u,v\in b^h}\hat{c}^{ho}_{u,v}}{|b^h|}, s^{\text{obj}}_{\hat{b}^o} = \frac{\sum_{u,v\in \hat{b}^o}\hat{c}^{oo}_{u,v}}{|\hat{b}^o|} \tag{2}$$

We suppress the object detection by a object probability threshold t_{obj}. The final confidence $\hat{c}_{b^h}$ of the hand b^h is the fusion of the hand contact score and the object probability score defined as

$$\hat{c}_{b^h} = \begin{cases} 1 - s^{\text{contact}}_{b^h} & \text{if } s^{\text{contact}}_{b^h} < t_{\text{contact}} \\ s^{\text{contact}}_{b^h} \cdot s^{obj}_{\hat{b}^o} & \text{otherwise} \end{cases} \tag{3}$$

We use $t_{\text{obj}} = 0.2$, $t_{\text{contact}} = 0.1$ in all our experiments.

C. Analysis: Robustness to Occlusions

To analyze the robustness of our method to occlusions, we compute the recall on the hand-object pairs with three different occlusion levels on 100DOH dataset. The occlusion level of a hand-object pair is measured by the IoU between their bounding boxes. In 100DOH dataset, there are 2222 hand-object pairs with an IoU $\in [.25, .5)$, 348 pairs with an IoU $\in [.5, .75)$, and 14 pairs with an IoU $\in [.75, .1]$. The quantitative comparison in Tab. 1 shows that our method is more robust in detecting active objects under occlusions over all baselines.

Method	Recall(IoU$\in [.25, .5)$)	Recall (IoU$\in [.5, .75)$)	Recall(IoU$\in [.75, 1]$)
100DOH Detector	68.68	63.22	78.57
PPDM	53.24	53.45	64.29
HOTR	71.69	68.10	71.43
Ours	**77.22**	**78.45**	**100**

Table 1. Results of hand-object interaction detection for hand-object pairs with different occlusion levels on 100DOH dataset.

D. Ablation: Effect of Reinforcement Learning

Repeatedly applying the voting function trained for one-step prediction (supervised learning) could result in a data distribution shift issue. Specifically, the small error at each step could compound the sequential predictions, which leads to a bad performance towards the final prediction. The application of RL is to mitigate this issue by optimizing over the sequence with an accumulative loss for the sequential predictions. We examine the effect of RL by comparing the performance with and without using RL. The results are shown in Tab. 2, which demonstrate that RL gives significant improvements for AP^{75} and AP^{50} on 100DOH dataset.

E. Visualizations

Qualitative Results The qualitative results on 100DOH dataset [3] and MEECANO dataset [2] are presented in Fig. 2 and Fig. 3 respectively. Each green arrow points from a hand bounding box (blue) to the corresponding active object bounding box (red). The visualization shows that our method is able to robustly detect the active object under scenes with overlapping objects and severe occlusions. Most failure cases are due to wrong hand detection, motion blur, and insufficient feature from tiny hands and objects.

Visualization of Iterative Refinement We further visualize the effect of iterative refinement. In this visualization, we show the initial active object hypothesis (yellow bounding box) and the refined active object estimation (red bounding box) on 100DOH dataset (in Fig. 4) and MEECANO dataset (in Fig. 5). All the examples show that the iterative refinement by applying the voting function multiple times could improve the active object bounding box estimation. For better visibility, every sample only shows one pair of hands and objects.

Dataset	RL	AP^{75}	AP^{50}	AP^{25}
100DOH	✗	23.64	46.84	**57.44**
100DOH	✓	**29.90**	**53.02**	57.15
MECCANO	✗	**13.13**	26.21	34.88
MECCANO	✓	12.99	**26.25**	**34.88**

Table 2. Ablation studies on reinforcement learning (RL) on 100DOH and MECCANO datasets.

Visualization of Pixel-wise Voting To validate the design of pixel-wise voting, we visualize more examples about the heatmap of the IoU between pixel-wise bounding box predictions and the final predicted bounding boxes after voting in Fig. 6. In this visualization, we clearly observe that the final estimated bounding boxes picked by the voting are related more closely to the predictions in the regions of informative patterns such as fingers and objects as opposed to irrelevant information such as the background. For better visibility, every sample only shows one pair of hands and objects.

F. Running Time

We report the runtime on a desktop with a Ryzen 3900X CPU and an RTX 2080Ti GPU. For a 512×512 input image with 2 hands on average, the proposed method runs at 18 frames/second, with 13 ms for network forward inference, and 42 ms for active object localization with voting.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021. 2

[3] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 2

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

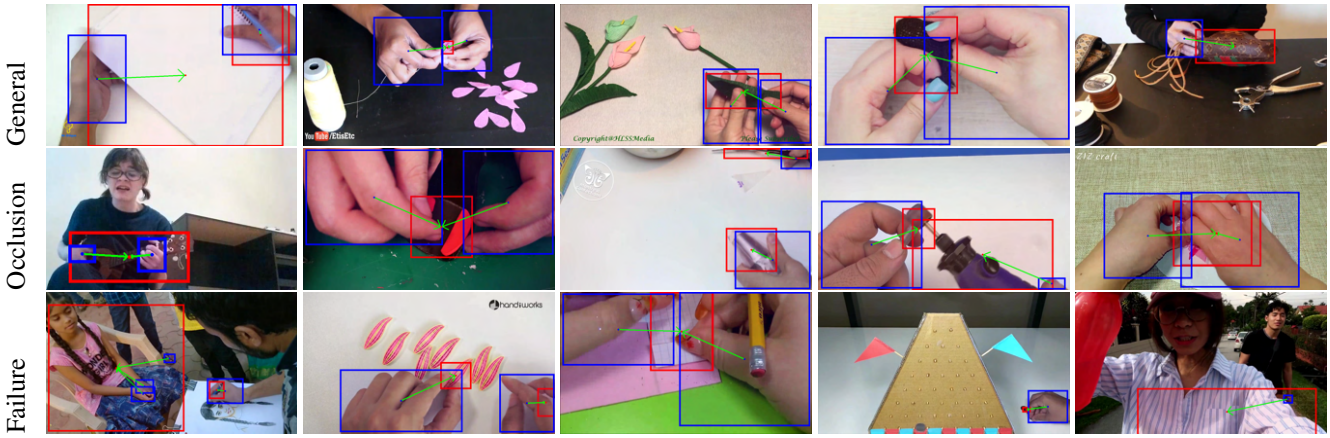

Figure 2. Qualitative Results on the 100DOH dataset. Each green arrow points from a hand bounding box (blue) to the corresponding active object bounding box (red).

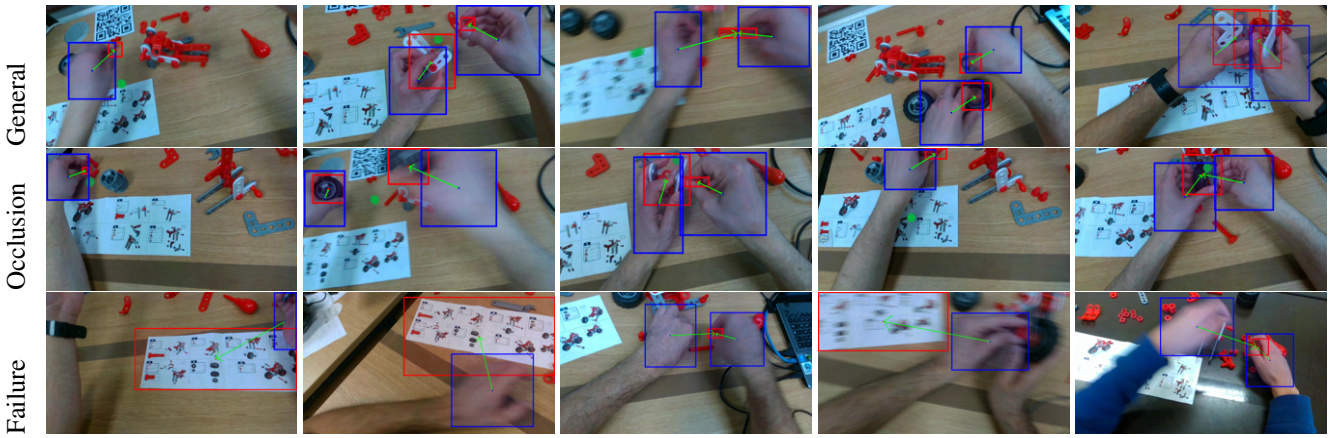

Figure 3. Qualitative Results on the MECCANO dataset. Each green arrow points from a hand bounding box (blue) to the corresponding active object bounding box (red).

Figure 4. Visualization of iterative refinement on the 100DOH dataset. We show the initial active object hypothesis (yellow bounding box) and the refined active object estimation (red bounding box) corresponding to the hand (blue bounding box).

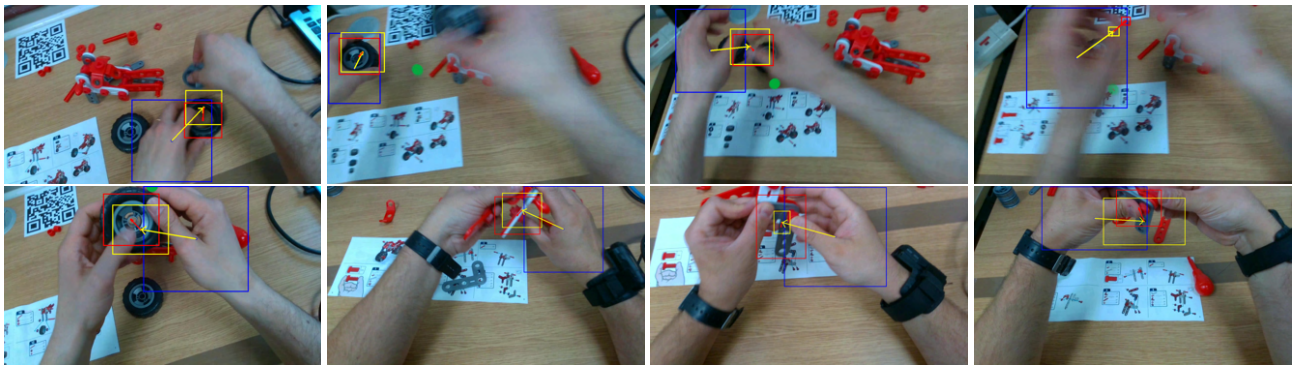

Figure 5. Visualization of iterative refinement on the MECCANO dataset. We show the initial active object hypothesis (yellow bounding box) and the refined active object estimation (red bounding box) corresponding to the hand (blue bounding box).

Figure 6. We show more examples by visualizing the IoU (red indicates higher IoU) between the final active object box estimation (red) and the pixel-wise predictions inside the hand bounding box (blue). The final estimated bounding boxes picked by the voting are more closely related to the predictions in the regions of informative patterns such as fingers and objects as opposed to irrelevant information such as the background.