Supplementary Material for Mining Cross-Person Cues for Body-Part Interactiveness Learning in HOI Detection

Xiaoqian Wu[1]*, Yong-Lu Li[1,2]*, Xinpeng Liu[1], Junyi Zhang[1], Yuzhe Wu[3], and Cewu Lu[1]**

[1] Shanghai Jiao Tong University
[2] Hong Kong university of Science and Technology
[3] DongHua University
{enlighten,yonglu_li,junyizhang,lucewu}@sjtu.edu.cn,
{xinpengliu0907,wuyuzhe486}@gmail.com

1 HOI Hard Cases

In Fig. 1a (main text, the same below), we show the histograms of hard cases in HICO-DET [1] train set and find that hard cases are common in HOI datasets, which impedes interactiveness learning. In Sec. 5.3, we split HICO-DET [1] test set and compare the interactiveness detection performance of TIN++ [5] and our method. We detail the settings as follows:

- **Tiny interactive persons.** In the 1-st histogram in Fig. 1a, for each interactive H-O pair, the ratio r of the area of the human bounding box to the area of the image is calculated. In the test set, the image is considered as "tiny-persons scenes", if it has at least one interactive pair with the ratio $r<0.1$.
- **Crowded Scenes.** We calculate the detected person counts in each image and show them in the 2-nd histogram in Fig. 1a. In the test set, the image is considered as "crowded scenes", if it has at least three interactive pairs.
- **Occlusion.** In the 3-rd histogram in Fig. 1a, for each interactive H-O pair, the average of human joint detection confidence is calculated based on the pose estimation results [2]. In the test set, for each image, we calculate the average of joint detection confidence of all the human bounding boxes in it as j. An image is regarded as "more-occlusion scenes" if $j<0.2$ and regarded as "less-occlusion scenes" if $j>0.6$.

Notably, we split the test set in *image level* for the convenience of model inference, *i.e.*, avoid an image to be seen in inference phrase. For example, in an image tiny persons and normal size persons may co-exist, and more/less-occluded persons may co-exist. After the split, for all the interactive persons in "tiny-persons scenes" images, 84.3% of them has a ratio $r<0.1$. For all the interactive

* The first two authors contribute equally.
** Cewu Lu is corresponding author, member of Qing Yuan Research Institute and Shanghai Qi Zhi institute.

persons in "more-occlusion scenes" images, 83.1% of them has an average of joint detection confidence j<0.2. Thus, with the image-level split, the model performance on hard cases can still be evaluated effectively and accurately with a much higher ratio of hard instances within images compared to the previous split.

Suppl Tab. 1 shows interactiveness AP under different settings on HICO-DET [1]. With our global perspective to learn the holistic relationship of body-part interactiveness, the difficulty of interactiveness learning is alleviated. Comparing our method with open-source state-of-the-art methods [5,7,8,9], the gains of hard cases are larger than non-hard cases, validating the effectiveness of our method, especially for HOI hard cases.

Table 1: Interactiveness AP under different settings.

Method	Full	Sparse/Crowded	Normal/Tiny	Less/More Occ
TIN++	14.35	16.96/9.64	16.11/8.94	16.49/8.06
PPDM	27.34	34.67/26.69	31.79/26.33	29.83/17.25
QPIC	32.96	36.80/27.04	34.02/26.14	32.08/19.75
CDN	33.55	39.92/28.84	36.10/25.11	34.55/21.69
Ours	**38.74**	**43.62/33.10**	**39.85/32.47**	**38.60/22.75**

2 Sparse vs. Crowded Scene

In this section, we detail the discussion about sparse/crowded scenes in Sec. 4 of the main text.

In the widely used HOI dataset HICO-DET [1] and V-COCO [4], we analyze the detected person counts on each image, and find that images with more than two persons account for 47.3%/62.5% in train/test set in HICO-DET [1], and the number is 40.36%/58.6% for V-COCO [4]. Thus, crowded images occupy a large proportion in the HOI dataset, validating the effects brought by our method.

We split HICO-DET [1] test set into *sparse* and *crowded* scenes respectively and evaluate the interactiveness AP. Here, images with at least three interactive pairs are considered as "crowded". The performances are 16.96/9.64 AP (TIN++ [5]) and 43.62/33.10 AP (ours). From the large performance gap (7.32 for TIN++ [5] and 10.52 for ours), we can see that interactiveness learning is mainly bottlenecked by crowded scenes. Therefore, it matters to focus on crowded scenes for interactiveness learning.

3 Detailed Experiment Settings

In our training process, the interactiveness classifier is first trained and then is the verb classifier. Since the box detector is fine-tuned in the 2-nd stage, for each image we finally get two predicted sets: 1) detection with verb classification

results p_v: $R = \{r^i | r^i = (b(h)^i, b(o)^i, c^i, p^i_{verb})\}_{i=1}^{N_q}$, and 2) detection with interactiveness classification results p : $R^{'} = \{r^{j^{'}} | r^{j^{'}} = (b(h)^{j^{'}}, b(o)^{j^{'}}, c^{j^{'}}, p^j_{int})\}_{j=1}^{N_q}$. We use $R^{'}$ to calculate interactiveness AP in Tab. 1. For HOI detection boosting in Tab. 2 and Tab. 3, for each $r^i \in R$, a matched proposal $r^{f(i)^{'}}$ is found from $R^{'}$ and the matched interactiveness results $p^{f(i)}_{int}$ is used for non-interaction suppression [6], where H-O pairs with lower interactiveness scores $p^{f(i)}_{int}$ are filtered out. Here, the matching function $f(i)$ is obtained via

$$f(i) = \underset{j:1\leq j\leq N_q, c^i=c^{j^{'}}}{argmax} (IoU(b(h)^i, b(h)^{j^{'}}) + IoU(b(o)^i, b(o)^{j^{'}})). \quad (1)$$

i.e., the matched proposal should have the correct object class and maximum human&object bounding boxes IoUs. When no matched proposal is found for r^i, the interactiveness score is 0.

In Tab. 4, we feed the representative two-stage HOI method iCAN [3] (human-object pairs are exhaustively paired) with our detected pairs. The inference score S is obtained via $S = S_v * S_o$, where S_v is the HOI prediction score from iCAN [3] and S_o is the object prediction score from our method. Also, a pairwise NMS with a threshold of 0.6 is conducted. In our method, interactiveness inference results are used for filtering out non-interaction pairs. Additionally, CDN [9] reports similar results by replacing detected boxes (Tab. 1 in their paper) while the results are different. We assume it is because of different experiment settings.

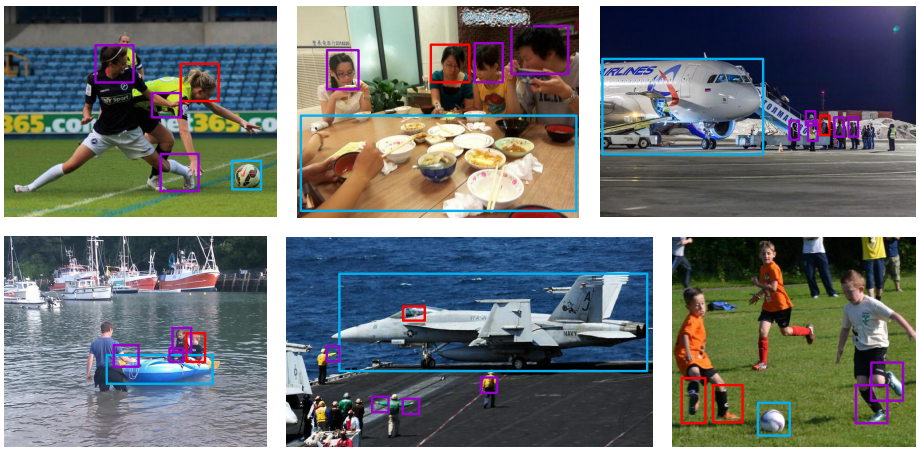

Fig. 1: Some examples where our proposed global perspective facilitates interactiveness learning. In the images, red boxes represent targeted body-parts, blue boxes represent targeted objects, and purple boxes represent body-parts of other persons which provide informative cues for interactiveness classification of the targeted body-parts and objects.

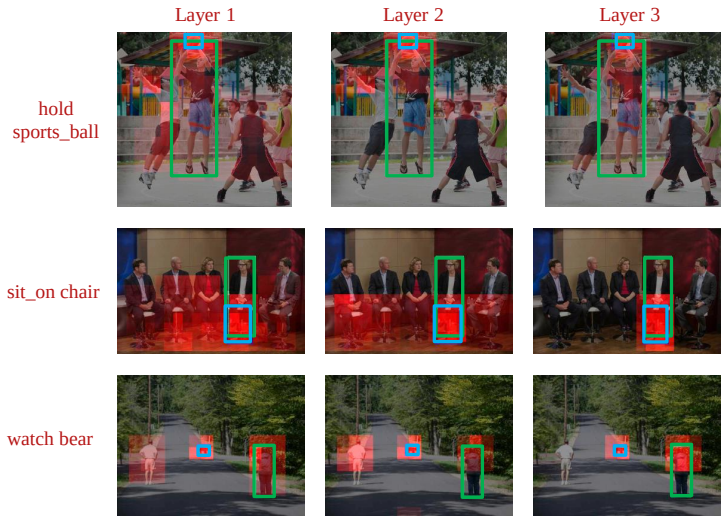

Fig. 2: More visualization results of how the learned attention changes in each layer, which shows the effectiveness of progressively masking and attention concentration.

4 Visualization

In Suppl Fig. 1, we extend Fig. 1b and show some examples where our proposed global perspective facilitates interactiveness learning. For example, in the left-most image of the upper row, the athlete is inspecting a sports_ball, and her competitor who is inspecting and reaching for it can provide some useful information.

Fig. 4 shows some visualization results of the learned attention. Here we provide a detailed analysis. In Fig. 4d, f, our model learns the attention on the same body-parts (head & hands) of the target person and other persons. In Fig. 4a/c, when classifying hands/hands interactiveness of the target person, our model learns to highlight arms & legs/legs & feet of other persons to explore visual cues. For hard cases, train passengers (Fig. 4e) and other athletes in the field (Fig. 4g) provide useful visual cues.

Moreover, Suppl Fig. 2 shows more visualization results of how the learned attention changes in each layer. The 1-st layer allows attention computation from different body-parts of different persons. The 2-nd layer emphasizes the same body-part from different persons in the image, while the 3-rd layer focuses on the body-part of the targeted person. We can see that with the progressive masks throughout transformer layers, different visual patterns are flexibly encoded to facilitate body-part interactiveness learning.

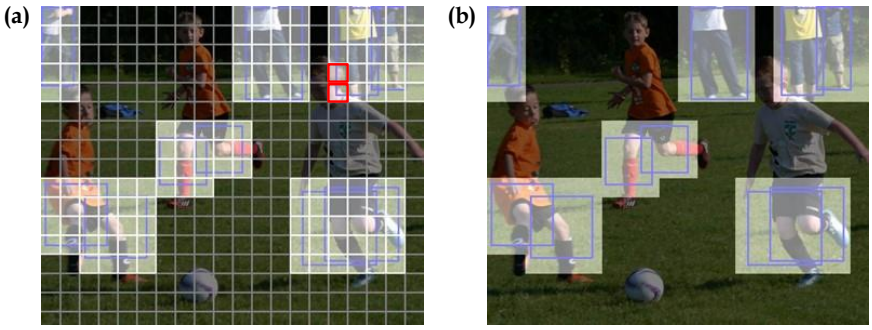

Fig. 3: The resolution of the feature map, the detected body-part boxes and the calculated attention masks. a) The resolution of the feature map (the grid). b) The detected body-part boxes (left/right legs in blue boxes) and the calculated attention masks (highlighted white regions).

5 Detailed Design of Attention Masks

In Sec. 3.2 of the main text, we propose to construct body-part saliency maps via transformer attention masks. In this section, we introduce the detailed design.

Given a transformer layer in f_{dec2}, the input queries are $D = \{d^i | d^i \in \mathcal{R}^{D_c}\}_{i=1}^{N_q}$, and the keys and values $K, V \in \mathcal{R}^{S \times D_c}$ ($S = H \cdot W$) are obtained from feature map z. For the i-th proposal, originally we have the self-attention calculation as $Att(d^i, K, V) = softmax(d_i K^T / \sqrt{D_c})V$. With a mask matrix $m^i \in \{0, 1\}^S$(0 for masked) to emphasize body-parts, we have $Att^*(d^i, K, V) = softmax(m^{i*} \circ (d^i K^T)/\sqrt{D_c})V$, where $\circ$ is Hadamard product, $m^{i*} = \{m_s^{i*} | m_s^{i*} = m_s^i (m_s^i = 1)\ or\ m_s^{i*} = -inf(m_s^i = 0)\}$ and inf is numerically big enough (*e.g.*, $2^{32} - 1$). Thus, unrelated tokens are dropped from self-attention calculation.

For the attention mask in Eq. 1, the resolution of the feature map z is scaled down from that of the original image and the scaling factor is 32 with ResNet-50 backbone. Masks are applied on the feature map instead of the original image for the convenience of model design. Despite the down-sampling, masks on the feature map can accurately express body-parts. An example is shown in Suppl Fig. 3. For each of the three boys, their legs occupy more than 15 tokens. Even for the background person in the left upper corner, he occupies 10 tokens.

Another consideration is that, for the tokens near the body-part boxes border, they may have a small overlap with the useful body-part while a large overlap with the background, thus bringing noisy information, or a large overlap with other body-parts, thus bringing confusion. Thus, we propose to randomly drop these tokens based on how much ratio a token is inside the part box. For example, the two red tokens in Suppl Fig. 3a may be randomly dropped. Nevertheless, we find the detection performance is slightly changed (interactiveness AP is 38.72 /38.74 w/o random dropping). Therefore, the mask is accurate enough to express

the semantics of body-parts despite some geometrical ambiguity because most of the tokens contain the correct body-parts.

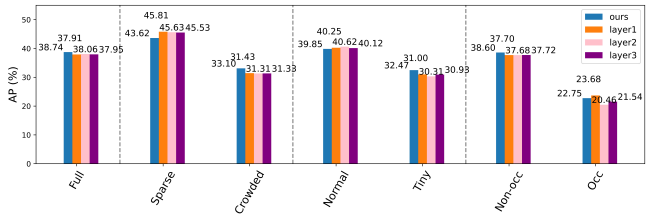

Fig. 4: Interactiveness detection AP on hard cases with progressively masking removed.

6 Analysis of Progressively Masking

In this section, we analyze the effect of progressively masking in detail. As an extension of the ablation studies in Sec. 5.5 of the main text, we remove progressively masking and apply the same attention mask on all decoder layers in f_{dec2} when classifying interactiveness. The interactiveness AP is 37.91/38.06/37.95 when applying masks $m_1/m_2/m_3$ on all layers, while AP is 38.74 with progressively masking. Additionally, we report the detailed performance on hard cases in Fig. 4. Progressively masking has a slight advantage for sparse scenes with only one/two interactive pairs, while an obvious advantage for other cases, especially for hard cases. The results validate its effectiveness to integrate diverse body-part oriented visual patterns flexibly.

7 Discussion of Limitations

Despite the effectiveness of the proposed global perspective, there are still some limitations. One limitation is the accuracy of body-part boxes from pose estimation results. It may be better to integrate body-part box regression into the training. Moreover, body-part interactiveness would be learned better if visual patterns across different images are considered. We plan to improve them in the future work.

References

1. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018)
2. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV (2017)

3. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. In: BMVC (2018)
4. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
5. Li, Y.L., Liu, X., Wu, X., Huang, X., Xu, L., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. TPAMI (2022)
6. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: CVPR (2019)
7. Liao, Y., Liu, S., Wang, F., Chen, Y., Feng, J.: Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020)
8. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021)
9. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. arXiv preprint arXiv:2108.05077 (2021)