Interactiveness Field in Human-Object Interactions*

Xinpeng Liu[1†] Yong-Lu Li[2†] Xiaoqian Wu[1] Yu-Wing Tai[3] Cewu Lu[1‡] Chi-Keung Tang[2]
[1]Shanghai Jiao Tong University [2]HKUST [3]Kuaishou Technology
{xinpengliu0907, yuwing}@gmail.com, {yonglu_li, enlighten, lucewu}@sjtu.edu.cn, cktang@cs.ust.hk

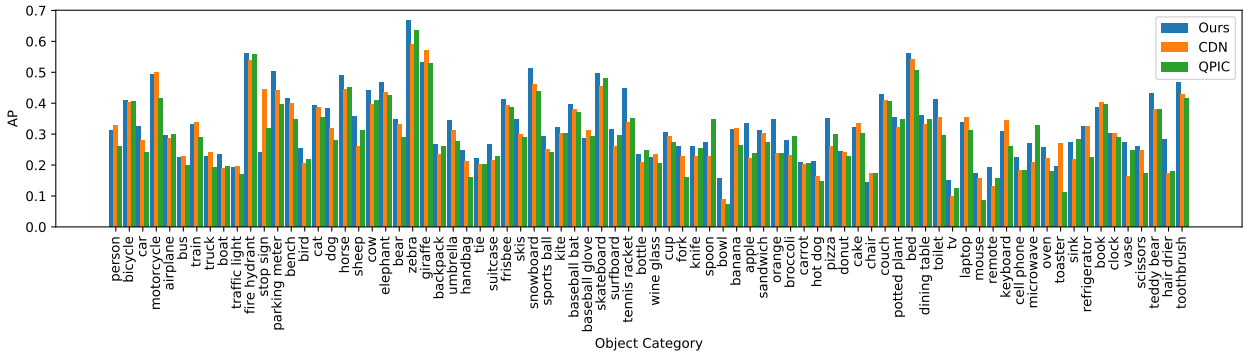

Figure 1. Interactiveness AP for different object categories of our model, QPIC [11], and CDN [12].

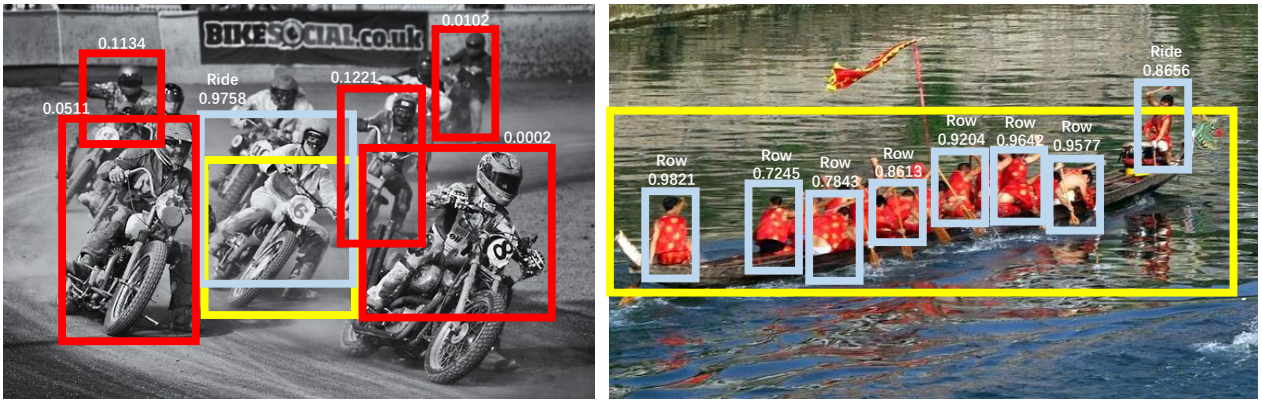

Figure 2. Some typical detection results on HICO-DET [2]. Given an object (yellow), our model efficiently pairs the object with the interactive human (cerulean) and filters out the non-interactive human instances (red). Under both circumstances (non-interactive pairs in majority on the left, and interactive pairs in majority on the right), we achieve satisfactory results.

*The research is supported in part by the Hong Kong Research Grant Council under grant number 16201420.
†The first two authors contribute equally.
‡Corresponding author.

1. Generalization of Interactiveness Bi-modal Prior

A potential issue is the generalization of our proposed bi-modal prior. We reemphasize that the bimodal prior is

universal with good generalization in the context of HOI in two aspects. **First**, the bimodal prior is fundamentally rooted in HOI. The very compositional nature of human and object in HOI makes it susceptible to a severely imbalanced distribution as revealed by *Zipf's Law*. As shown in Table 1, widely used natural image HOI datasets [2–5] all hold the prior **Second**, we claim that the **object-centric**

Dataset	$\frac{\#inter}{\#no-inter} \ll 1$	$\frac{\#inter}{\#no-inter} \approx 1$	$\frac{\#inter}{\#no-inter} \gg 1$
HICO-DET [2]	79.1%	7.3%	13.6%
V-COCO [4]	76.0%	7.9%	16.1%
Ambiguous-HOI [5]	80.1%	6.2%	13.7%
AVA [3]	73.2%	8.4%	18.4%

Table 1. Interactive ratio of different datasets.

bimodal prior exploited in our paper is one subclass of the prior, since the widely-used benchmarks HICO-DET and V-COCO both have this property. Besides the object-centric prior that is more suitable in *multi-person* scene, a similar prior exists in a **human-centric** view for images with *few people*. Even for really sparse scenes containing one person and one object, in a **body-part view** inspired by [10], the interactive body parts are statistically rare. For such sparse scenes, statistics show in images with only *one person* and *one object* from HAKE [6, 7] that only **9.8**% of the existing parts are interactive with objects. That said, the prior still holds as a *learning paradigm*. We believe the object-centric prior is a first step towards deeper exploration on such useful prior.

2. Detailed Analysis on Interactiveness Detection

As stated in the main paper in Section 4.3, we evaluate our model using the interactive AP metric proposed by TIN [8]. In this section, we include more details for interactiveness detection. Figure 1 shows the interactiveness AP for different object categories of our model and previous state-of-the-art QPIC [11] and CDN [12]. Our model achieves superior performance on most of the object categories. In detail, our method takes the lead in **56** of the 80 object categories, while falling behind on only 4 categories. Furthermore, on over 20 object categories, our advantage is more than 5 mAP, indicating the efficacy of our interactiveness detection for various objects.

3. Prediction Visualization

To vividly show the effectiveness of our method, we give some typical results on HICO-DET [2] in Figure 2. Our method can precisely filter out the non-interactive pairs while detecting interactive pairs in complex scenes.

4. Discussion on Limitations

Though the interactiveness field has greatly enhanced H-O pairing and boosted the HOI detection, the room for H-O pairing is still large needing more exploration.

While the proposed bimodal prior is of great efficacy in interactiveness modeling, it is still an issue to precisely discern the correspondence between rare/non-rare pairs and interactive/non-interactive pairs. Since even with a compromised strategy that treats rare pairs as interactive, the performance improvement is considerable, we believe effective inference on the correspondence may lead to very promising enhancement.

The proposed interactiveness field is investigated generally based on the bimodal prior only, while we believe the more fine-grained study is worthwhile, e.g., the interactiveness field for different object categories inspired by Liu *et al*. [9], the interactiveness field for different verb categories, the field under different background contexts, the interplay of the interactiveness fields of different objects, and so on.

5. Societal Impact

As all the data used here come from public dataset thus there is no privacy issue. Our work aims at prompting the HOI understanding, thus may be helpful to the development of health-care robot, etc. However, there could be potentially negative societal implications, such as its potential use in surveillance, military purposes which requires serious moral consideration. We encourage well-intended application of our method.

6. Licenses of Adopted Datasets

V-COCO [4] is released under the MIT License. Our code is mostly derived from DETR [1], QPIC [11] and TIN [8]. DETR [1] and QPIC [11] are released under the Apache License 2.0. While TIN [8] is released under the MIT License.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 2

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2

[3] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

[4] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[5] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2

[6] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: A knowledge engine foundation for human activity understanding, 2022. 2

[7] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019. 2

[8] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 2

[9] Xinpeng Liu, Yong-Lu Li, and Cewu Lu. Highlighting object category immunity for the generalization of human-object interaction detection. *arXiv preprint arXiv:2202.09492*, 2022. 2

[10] Cewu Lu, Hao Su, Yonglu Li, Yongyi Lu, Li Yi, Chi-Keung Tang, and Leonidas J Guibas. Beyond holistic object recognition: Enriching image understanding with part states. In *CVPR*, 2018. 2

[11] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 1, 2

[12] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *arXiv preprint arXiv:2108.05077*, 2021. 1, 2