Human-Object Interaction Detection via Disentangled Transformer
Supplementary Material

Desen Zhou[1]* Zhichao Liu[1,2]*† Jian Wang[1] Leshan Wang[1,2]† Tao Hu[1] Errui Ding[1] Jingdong Wang[1]
[1]Department of Computer Vision Technology (VIS), Baidu Inc.
[2]ShanghaiTech University
{zhoudesen,wangjian33,hutao06,dingerrui}@baidu.com
{liuzhch,wanglsh}@shanghaitech.edu.cn, wangjingdong@outlook.com

1. More Qualitative Results

To further verify the efficacy of our proposed disentangled strategy for HOI detection, in this supplementary material, we provide more qualitative results of our experiments. We first show the visualization of our disentangled cross attention maps, and then show several predicted samples under challenging scenarios on V-COCO [4] and HICO-DET [2] benchmark.

1.1. Visualization of disentangled cross attentions

Our motivation to disentangle HOI representations is that a network might focus on different spatial regions for instances and interactions of HOI triplets. For example, in instance detection, the features of object extremities might be gathered for regressing object bounding boxes, as suggested by previous transformer detectors [1,5]; while in interaction classification, the features of interactive regions or human postures should be important, prior two-stage methods [3,6] have shown that interactive regions or human parts are informative for interaction classification.

We therefore visualize the cross attention maps of the same HOI triplet in the last layer of two disentangled task decoders to verify the effectiveness of our disentangled strategy, shown in Fig.1. The top rows of V-COCO dataset and HICO-DET dataset show the cross attention maps of interaction decoder, we can observe that the attention maps highlight the interactive regions between human-object instance pairs or informative human parts. While in the bottom rows of both datasets, the instance decoder attents to the object extremities. The different attention maps indicate that our instance and interaction decoders indeed capture clear disentangled representations, demonstrating the efficacy of our proposed method.

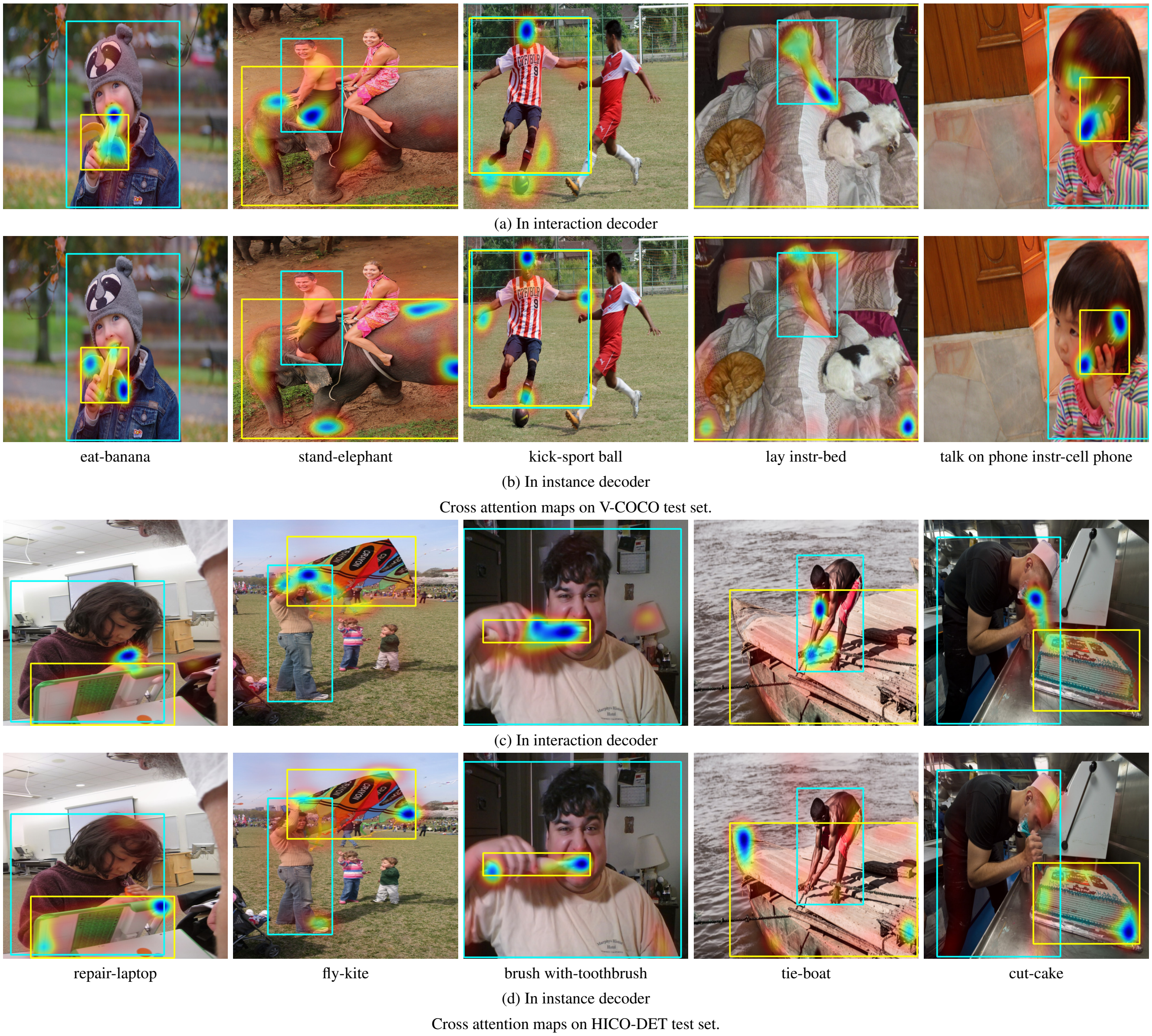
1.2. Visualization of predicted samples

We then show some predicted HOI triplets on both benchmarks. As shown in Fig. 2, from the first row of each dataset, we can observe that our model is able to predict high confident HOI triplets under challenging scenarios with cluttered backgrounds. From the second row of each dataset, we can further observe that our transformer network is able to localize and classify tiny objects accurately even under occlusion. We guess that it's because our unified representation provides a joint configuration of HOI triplets and hence the instance decoder is able to take an interactive prior to predict objects.

In addition, since our proposed disentangled strategy allows the disentangled representations to be implicitly associated with each other via coarse-to-fine manner, and no further grouping process is required, our method is able to accurately predict human-object interactions with multiple persons and objects in the scene. As shown in Fig.2, from the last row of each dataset, we can see that our method can associate human-objects well in multi-person cases.

*Equal contribution.

†Work done when Zhichao and Leshan were interns at VIS, Baidu.



(a) In interaction decoder

eat-banana stand-elephant kick-sport ball lay instr-bed talk on phone instr-cell phone

(b) In instance decoder

Cross attention maps on V-COCO test set.

(c) In interaction decoder

repair-laptop fly-kite brush with-toothbrush tie-boat cut-cake

(d) In instance decoder

Cross attention maps on HICO-DET test set.

Figure 1. Visualization of cross attention maps of the same triplet prediction in our interaction decoder and instance decoder. The first two rows are from V-COCO and the last two rows are from HICO-DET. we can observe that our interaction decoder attends to the interactive regions of human and objects or informative human parts, while the instance decoder attends to the object extremities. The different attention maps imply that interaction and instance decoders indeed capture disentangled representations of images.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

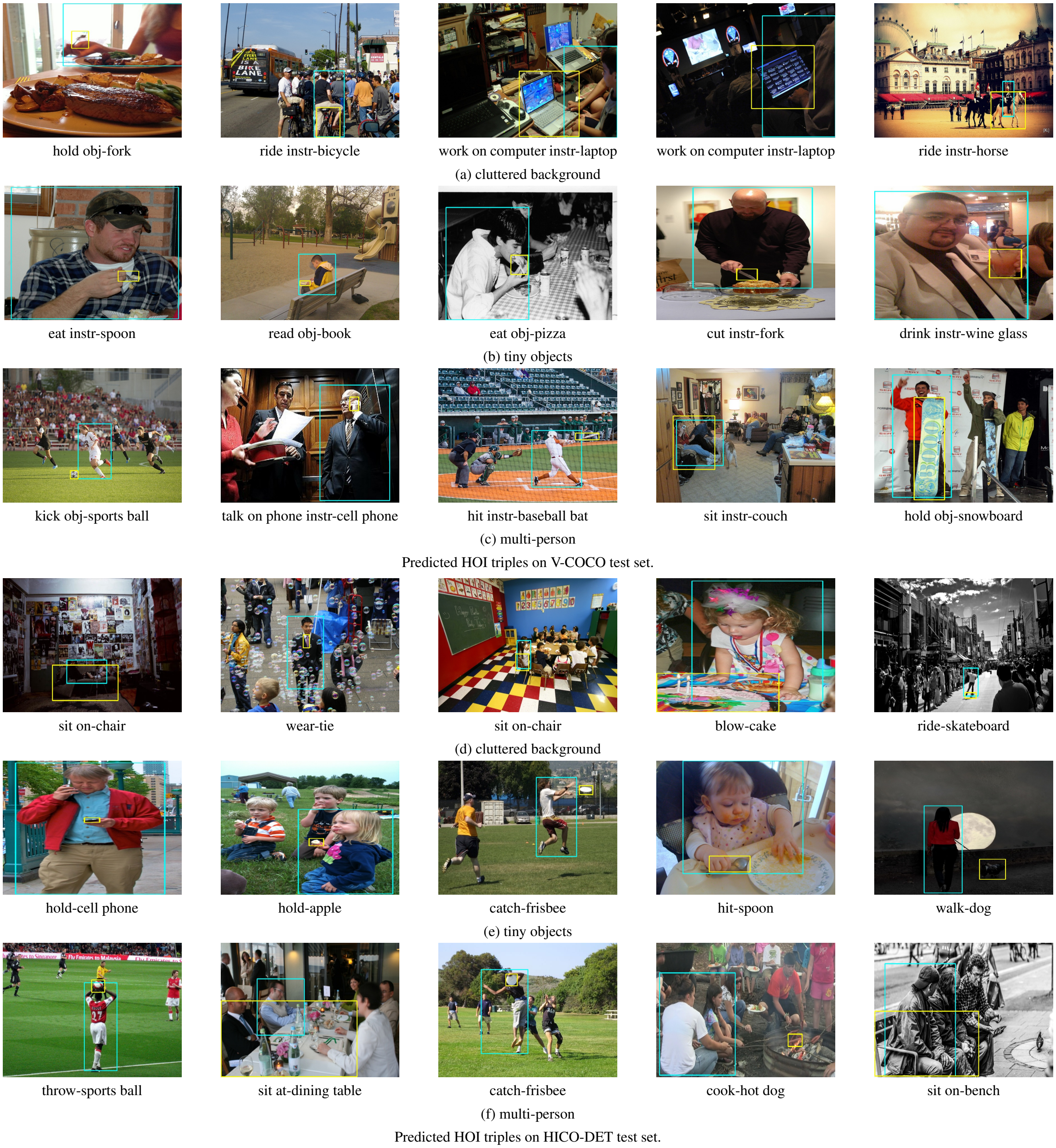
[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1

[3] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 1

[4] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1

[5] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 1

[6] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object



hold obj-fork | ride instr-bicycle | work on computer instr-laptop | work on computer instr-laptop | ride instr-horse

(a) cluttered background

eat instr-spoon | read obj-book | eat obj-pizza | cut instr-fork | drink instr-wine glass

(b) tiny objects

kick obj-sports ball | talk on phone instr-cell phone | hit instr-baseball bat | sit instr-couch | hold obj-snowboard

(c) multi-person

Predicted HOI triples on V-COCO test set.

sit on-chair | wear-tie | sit on-chair | blow-cake | ride-skateboard

(d) cluttered background

hold-cell phone | hold-apple | catch-frisbee | hit-spoon | walk-dog

(e) tiny objects

throw-sports ball | sit at-dining table | catch-frisbee | cook-hot dog | sit on-bench

(f) multi-person

Predicted HOI triples on HICO-DET test set.

Figure 2. Highly confident HOI predictions on V-COCO(first three rows) and HICO-DET(last three rows) test set. As shown above, our method works well under challenging scenarios with cluttered backgrounds(first rows), and is able to predict tiny-objects even under occlusions(second rows), and also works well with multi-person cases(third rows), demonstrating the effectiveness of our proposed disentangled strategy.

interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 1