Exploring Structure-aware Transformer over Interaction Proposals for Human-Object Interaction Detection — CVPR 2022 Supplementary Material*

Yong Zhang†, Yingwei Pan‡, Ting Yao‡, Rui Huang†, Tao Mei‡, and Chang-Wen Chen§
† The Chinese University of Hong Kong, Shenzhen ‡ JD Explore Academy
§ The Hong Kong Polytechnic University
yongzhang@link.cuhk.edu.cn, {panyw.ustc, tingyao.ustc}@gmail.com, ruihuang@cuhk.edu.cn, tmei@jd.com, changwen.chen@polyu.edu.hk

The supplementary material contains: (1) training time analysis of the proposed STIP [5]; (2) more ablation studies on the type of HOI feature representation; (3) cross-attention visualization of our STIP for HOI detection.

A. Training Time Analysis

In this section, we include the detailed training time comparison between our proposed STIP and several existing Transformer-style HOI detectors [2–4, 6]. As shown in Table 1, existing Transformer-style HOI detectors commonly suffer from slow convergence as in DETR [1], which usually require over 100 training epochs. We speculate that this may be the result of HOI set prediction driven from the parametric interaction queries with randomly initialized embeddings. Instead, our STIP starts HOI set prediction from non-parametric interaction queries (i.e., high-quality interaction proposals), thereby leading to more efficient Transformer learning with only 30 epochs.

Method	Batch Size	Epochs (HICO-DET dataset)
AS-Net [2]	64	90
HOTR [3]	16	100
QPIC [4]	16	150
HOITrans [6]	16	250
STIP (Ours)	8	30

Table 1. Training time comparison among Transformer-style HOI detectors. All runs are trained with the same backbone (ResNet-50) and optimizer (AdamW) for fair comparison.

B. Ablation Study on HOI Features

Here we conduct additional ablation studies to examine how HOI detection performance is affected when capitalizing on different types of HOI feature representations in our Interaction Proposal Network. Table 2 details the performances by exploiting different HOI features in STIP. In particular, the use of only appearance feature (A) in general achieves superior performances by clearly outperforming existing HOI detectors (see Table 1 and 2 in main paper). As expected, by integrating appearance feature with spatial or linguistic feature, consistent performance gains are attained. The results indicate that both spatial and linguistic cues are complements to the visual appearance cues of humans and objects. Finally, the combination of all the three types of HOI features reaches the highest performances, which basically demonstrate the complementarity in between.

Feature	V-COCO		HICO-DET (Default)		
	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Full	Rare	Non-Rare
A	65.09	69.66	29.76	26.94	30.61
A+S	65.64	70.47	30.31	27.85	31.04
A+L	65.60	70.33	30.01	26.50	31.06
A+S+L	**66.04**	**70.65**	**30.56**	**28.15**	**31.28**

Table 2. An ablation study on the use of different HOI features. The letters in Feature column indicate the input features: **A** (Appearance/Visual features), **S** (Spatial features), **L** (Linguistic feature of label semantic embeddings).

C. Cross-attention Visualization

In order to better qualitatively evaluate the structure-aware cross-attention module in our structure-aware Transformer, we visualize the attended image regions according to the learned cross-attention weights for HOTR [3], Base+HM+TR, and our STIP in Figure 1. Note that Base+HM+TR is a degraded version of our STIP by using vanilla Transformer for HOI set prediction. Specifically, as shown in Figure 1 (a-1) and (a-2), both HOTR and Base+HM+TR always focus on similar regions even when

*This work was performed at JD Explore Academy.

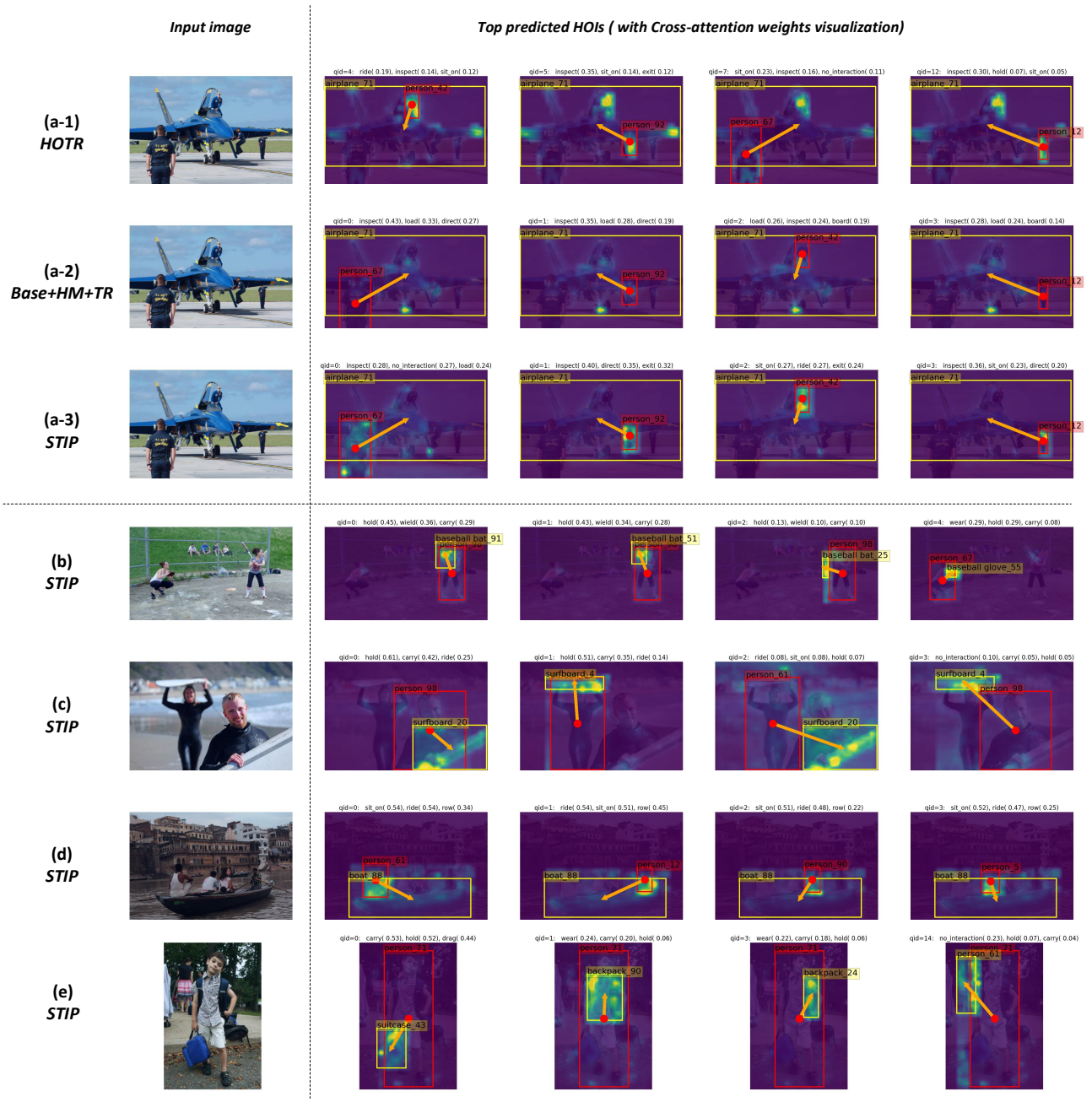

Figure 1. The cross-attention visualization of testing samples on HICO-DET dataset. We highlight the attended image regions according to the cross-attention weights in the last layer of Transformer. (a-1), (a-2), and (a-3) show the cross-attention visualization results of the same image for HOTR [3], Base+HM+TR and our proposed STIP (using the same pre-trained DETR). (b)-(e) showcase more cross-attention visualization results of our STIP. Particularly, for each image, we present the top-4 predicted human-object pairs. For each human-object pair (identified by 'qid'), we also list the top-3 predicted interactions and their classification scores on the top of each image.

predicting different human-object pairs for the same image. We speculate that this may be the result of solely performing cross-attention learning in vanilla Transformer without any prior knowledge, where the estimated cross-attention can be easily overwhelmed with the inherent salient regions in images. As an alternative, our structure-aware Transformer in STIP facilitates cross-attention learning with additional guidance of intra-interaction structure, and thus accurately steers cross-attention over image areas for depicting the target interaction (see Figure 1 (a-3)). Similarly, in Figure 1

(b)-(e), our STIP manages to attend over the relevant regions for recognizing the corresponding target interactions.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[2] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 1

[3] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 1, 2

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

[5] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, 2022. 1

[6] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 1