A. HOI-NMS & HOI-SoftNMS

NMS [6] and its variants [1] can obtain the better performance when mean-Average-Precision (mAP) is used as an evaluation metric and are therefore employed in state-of-the-art object detectors [2,4,5]. However, theses approaches cannot be applied directly to the HOI detection since there are a pair of bounding boxes of the human and the object in one HOI prediction. Motivated by the previous [1, 6], we present HOI-NMS and HOI-SoftNMS methods to further improve the performance by reducing the number of duplicate HOI predictions in pair-wise level. The pseudo code and the comparison are illustrated as the following.

Algorithm 1 HOI-NMS & HOI-SoftNMS

Input: $H = \{h_i =< b_i^H, b_i^O, c_i^o, c_i^I > |h_i\}^N, S = \{s_i^I\}^N$
H is the list of original HOI predictions;
b_i^H, b_i^O are the bounding boxes of the human and the object of the i-th HOI prediction;
c_i^O, c_i^I are the categories of the object and the interaction of the i-th HOI prediction;
s_i^I is the interaction score of the i-th HOI prediction.

1: **while** $H \neq empty$ **do**
2: $m \leftarrow argmax\ S$
3: $D \leftarrow D \cup h_m; H \leftarrow H - h_m$
4: **for** h_i in H **do**
5: $Calculate\ IoU_{hoi}(h_i, h_m)$ based on Formula 2;
6: **if** NMS and $IoU_{hoi} > t_{iou}$ **then**
7: $s_i = 0$
8: **else if** NMS and $IoU_{hoi} > t_{iou}$ **then**
9: $s_i = s_i \cdot e^{-\frac{IoU_{hoi}^2}{0.5}}$
10: **else**
11: $s_i = s_i$
12: **end if**
13: **end for**
14: **end while**

Output: D, S

Based on iou 1 used in the object detectors, we calculate IoU_{hoi} in line 5 to indicate the extent of overlap between two HOI predictions where the Formula 2 shows the calculating process.

$$iou(b_i, b_m) = \frac{b_i \cap b_m}{b_i \cup b_m} \tag{1}$$

$$IoU_{hoi}(h_i, h_m) = \begin{cases} 0, & If\ (c_i^O \neq c_m^O\ or\ c_i^I \neq c_m^I); \\ min(iou(b_i^H, b_m^H), iou(b_i^O, b_m^O)), & Else; \end{cases} \tag{2}$$

	Method	NMS	Full	Rare	Non-Rare
1	QPIC [7]	-	29.07	21.85	31.23
2		HOI-NMS	29.91	21.77	32.32
3		HOI-SoftNMS	30.00	21.78	32.43
4	**CATN (Ours)**	-	31.62	24.28	33.79
5		HOI-NMS	32.38	25.14	34.52
6		HOI-SoftNMS	**32.40**	**25.15**	**34.54**

Table 1. Comparison against different NMS strategies with their own best performance. Results (Row1/4 vs. Row2/3/5/6) indicate that the performance could be further improved when NMS strategy is employed. Compared with the original CATN with fastText [3] in line 4, HOI-SoftNMS improves mAP-full from 31.62 to 32.40, which obtains the best performance.

t_{iou}	0.2	0.3	0.4	0.5	0.6	0.7
HOI-NMS	31.784	32.079	32.211	32.317	**32.381**	32.372
HOI-SoftNMS	32.379	32.395	**32.397**	32.389	32.389	32.358

Table 2. The effect of different settings of t_{iou} on the HICO-DET dataset. We conduct experiments based on the CATN with fastText [3] and evaluate them on the mAP-full metric.

B. Complexity & Effectiveness

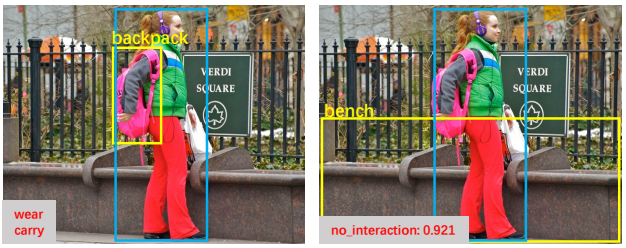
Our innovative method can be regarded as a 'Plug-and-Play' module that could be readily implemented to effectively promote the performance of transformer-based HOI detection models. Though the complexity of our method may slightly increase compared to baseline models, it is controllable in practical scenes. As shown in Table 3, the inference time only increases 2.4ms per image when a lightweight detector (e.g. Yolov5m) and the parallel architecture are adopted.

Method	Detector	mAP	Inf (ms)	Inf in Parallel (ms)
QPIC (baseline)	-	29.07	**43.7**	**43.7**
CATN (Ours)	Faster RCNN	**31.62**	81.3	58.4 (+14.7)
	Yolov5m	31.49	57.2	46.1 (+2.4)

Table 3. Our method could obtain significant improvement with the controllable increase of complexity. The inference time only increases 2.4ms per image on one RTX3090 when a lightweight detector, e.g. Yolov5m, is adopted in parallel.

C. Visualization

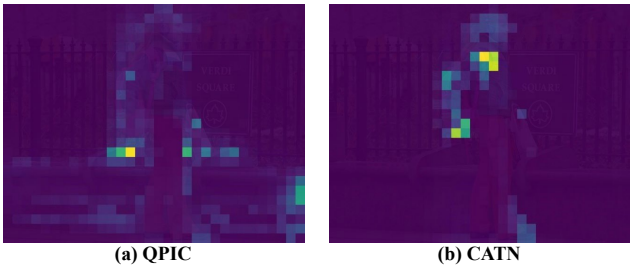
Figure 1 visualizes an HOI sample selected from the HICO-DET dataset. As seen from Figure 1(a, b, c), our CATN outperform the baseline [7] on both object detection and interaction classification. For object detection, our CATN accurately predicts the pair of human and object instances (both categories and bounding-boxes) while there is a false positive sample generated by QPIC (<human, bench, no_interaction(0.92)> is not contained in the image). For interaction classification, the annotated interactions have higher confidence scores, while non-annotated interactions have lower scores.



(a) GT (b) QPIC

(c) CATN (d) attention map of CLAM

Figure 1. Visualization of a sample from HICO-DET. The human bounding boxes, object bounding boxes, object classes, and verb classes are drawn with blue boxes, yellow boxes, yellow characters, and red characters respectively.



(a) QPIC (b) CATN

Figure 2. Visualization of decoder attention maps of QPIC (Baseline) v.s. CATN (Ours). As shown in Figure 2(b), our method can lead to more accurate attention so that the aggregated features are more meaningful and more beneficial to interaction classification.

C.1. The effectiveness of CLAM.

To analyze the effectiveness of the proposed category-level attention module (CLAM), we also visualize the attention map of the category of 'backpack' in Figure 1(d). The relative regions of both the 'backpack' and the interacting human are highlighted in the image, which demonstrates that our CLAM can automatically aggregate the features with rich information of the corresponding category.

C.2. The effectiveness of Q_{CA}.

We have conducted experiments on what initialization (all-zeros, random values, and category information) is useful and experiments on where (Object Query, CLAM, and Verb-Classifier) to leverage such information as shown in Tab.5 and Fig.4 in the original manuscript respectively. This research shows that using category info to initialize the Object Query can effectively promote the performance of HOI detection. To clearly explain how the category-aware information contributes to the HOI detection, we also visualize attention maps in transformer-decoder to show that this information and initialization would lead to more accurate attention, as shown in Figure 2.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 1

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1

[3] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*, 2017. 1

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1

[6] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers*, 100(5):562–569, 1971. 1

[7] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1