Supplementary Material for
Category Query Learning for Human-Object Interaction Classification

Chi Xie[1†] Fangao Zeng[2‡] Yue Hu[3‡] Shuang Liang[1*] Yichen Wei[2‡]
[1]Tongji University [2]MEGVII Technology [3]Shanghai Jiao Tong University
[1]{chixie, shuangliang}@tongji.edu.cn
[2]zfg472988436@163.com, wei_yi_chen@hotmail.com [3]18671129361@sjtu.edu.cn

1. Overview

In this supplemental file, we provide more details of our work to supply the main paper. Specifically,

- **Score integration technique** used in our paper are explained in Sec. 2;
- **Implementation details** are summarized in Sec. 3;
- **Additional ablations** are presented in Sec. 4, which includes the ablations on the score integration technique;
- **Additional qualitative results** are presented in Sec. 5.

2. Score Integration Technique

We introduce the score integration step briefly in Sec. 3.2 of the main paper, which leverages the image-level classification scores to stress or suppress certain categories during instance-level interaction categories. As the score integration step is not the major contribution of the proposed method, and brings minor improvement (as in Tab. 4 of the paper), we did not elaborate on its details in the paper.

Before applying this score integration step, based on Eq. (4) in the paper, we can compute the classification scores for the i-th human-object instance over K interaction categories as

$$s_i = \text{sigmoid}\left(\left[\frac{(F_i, \overline{Q'}_1)}{\|F_i\| \, \|\overline{Q'}_1\|}, \cdots, \frac{(F_i, \overline{Q'}_K)}{\|F_i\| \, \|\overline{Q'}_K\|}\right]\right), \tag{1}$$

where the sigmoid operation is applied on the vector element-wise.

Next, we provide the detailed design of this score integration step. It includes a hard integration and a soft one.

2.1. Hard Score Integration

This hard score integration is motivated by the rank-adaptive pixel classification in RankSeg [4]. It consists of two steps: the first is to use the image classification results to sort and select some interaction categories, and perform H-O pair classification only on the selected categories, namely, category selection; the second is to adopt a series of temperature parameters that ranks the interaction classification results of sorted and selected categories, namely, category ranking.

Category selection. Instead of choosing the labels for an H-O pair from all K predefined categories, based on the previous multi-label image classification prediction $\{p_k\}$ for the image, we perform a selected-label classification. First, the top κ of the classification weights $\{Q'_k\}$ is selected according to the descending order of image classification predictions as

$$\left[\overline{Q'}_1, \cdots, \overline{Q'}_\kappa\right] = \text{Top}-\kappa\left(\left[Q'_1, \cdots, Q'_K\right], \{p_k\}\right), \tag{2}$$

[*] Corresponding author.
[†] Work done during internship at MEGVII technology.
[‡] Work done while worked at MEGVII.

and H-O pair classification is performed as

$$s_i^h = \text{sigmoid}\left(\left[\frac{(F_i, \overline{Q'}_1)}{\|F_i\| \, \|\overline{Q'}_1\|}, \cdots, \frac{(F_i, \overline{Q'}_\kappa)}{\|F_i\| \, \|\overline{Q'}_\kappa\|}\right]\right), \tag{3}$$

where $[\overline{Q'}_1, \cdots, \overline{Q'}_\kappa]$ denotes the top κ selected category queries (classification weights) associated with the largest κ image classification scores, s_i^h denotes the classification scores with hard score integration, and κ represents the number of selected category queries, chosen as a much smaller value than K.

Category ranking. On top of category selection, we apply a set of learnable temperature parameters $[\tau_1, \tau_2, \cdots, \tau_\kappa]$ to adjust the classification scores over the selected top κ categories, so Eq. (3) is changed to

$$s_i^h = \text{sigmoid}\left(\left[\frac{(F_i, \overline{Q'}_1)}{\|F_i\| \, \|\overline{Q'}_1\| \, \tau_i}, \cdots, \frac{(F_i, \overline{Q'}_\kappa)}{\|F_i\| \, \|\overline{Q'}_\kappa\| \, \tau_\kappa}\right]\right). \tag{4}$$

We analyze the influence of κ choices and the benefits of such a ranking adjustment in the ablation study. Note that this is similar to the rank-adaptive pixel classification performed in RankSeg [4] for image and video segmentation tasks, though their classification is a single-label problem and softmax is applied while ours are multi-label and sigmoid is used.

2.2. Soft Score Integration

Another way to utilize the image-level classification scores is to directly multiply the instance classification scores s_i with the image classification probabilities $\{p_k\}$, as

$$s_i^s = \left[\sqrt{s_{i,1} * p_1}, \cdots, \sqrt{s_{i,K} * p_K}\right], \tag{5}$$

where s_i^s denotes the interaction classification scores of the i-th H-O instance, with soft sore integration.

Compared with the hard score integration, no interaction class is deprecated during instance classification. They are just stressed or suppressed in a soft way. Therefore, we call this soft score integration.

Note that hard and soft score integration can be applied together, as

$$s_i^{s,h} = \left[\sqrt{s_{i,1}^h * \overline{p}_1}, \cdots, \sqrt{s_{i,\kappa}^h * \overline{p}_\kappa}\right], \tag{6}$$

where $[\overline{p}_1, \cdots, \overline{p}_\kappa]$ is the top κ in $\{p_k\}$. Through experiments in Tab. 1, we find both soft and hard integration bring a small improvement and the best result is achieved when both is used.

3. Implementation Details

Most of the implementation details have been provided in the paper. Here we summarize these details. In the proposed category query learning, transformer decoder with 2 layers is used by default. The structure of each decoder layer in the proposed decoder consists of a cross-attention module, a self-attention module and a FFN in order. The weights of the existing losses in the baselines are not changed, and an image loss with loss weight $\lambda = 1.0$ is added to the final loss. For the asymmetric loss in image classification, we adopt $\gamma+ = 0$, $\gamma- = 4$ and $m = 0.05$. Both hard and soft score integration are used. For category selection and ranking in hard score integration, we set κ as 70 for HICO-DET [1]. Hyper-parameters like learning rate, weight decay, batch size and input image size follow the baseline settings by default.

Following the baseline detectors, the feature extractor is frozen for SCG [7], and updated for QPIC [6] and GEN-VLKT [5]. For the experiments on GEN-VLKT, we change its classification classes from 600 HOI categories to 117 interaction categories for HICO-DET and from 263 to 29 for V-COCO, following most HOI detection methods. For the experiments on SCG, the detection boxes are from a fine-tuned detector provided by DRG [2] for HICO-DET and a fine-tuned DETR for V-COCO [3]. The experiment is conducted on 8 Tesla V100 GPUs.

4. Additional Ablations

In this part, we perform some additional studies on technical details.

hard integration		soft integration	Default		
selection	ranking		Full	Rare	Non-Rare
-	-	-	34.98	31.73	35.95
✓	-	-	35.09	32.98	35.72
✓	✓	-	35.24	32.67	36.01
-	-	✓	35.18	32.23	36.06
✓	✓	✓	**35.36**	**32.97**	**36.07**

Table 1. Ablation on the techniques (soft and hard score integration) that we elaborate in Sec. 2 to utilize image-level classification scores. The best results are marked in **bold**.

κ	-	30	50	70	90	117
mAP	34.98	35.04	35.19	**35.24**	35.08	35.03

Table 2. Ablation on the number of interaction categories in the hard score integration step, i.e., κ. The metric for comparison is the full mAP under *default* setting on HICO-DET dataset. "-" denotes the hard score integration is not used. The best results are marked in **bold**.

4.1. Integration of Image-level Classification Scores

As mentioned in Sec. 2, the score integration process is proposed to utilize the image-classification scores in the proposed method, with two strategies: the **hard score integration**, consisting of category selection and category ranking, and the **soft score integration**, which is a score multiplication operation between instance-level and image-level classification scores. As shown in Tab. 1, each of them brought a marginal improvements: the model with hard score integration achieves 35.24 mAP while the one with score integration achieves 35.18 mAP. Together, a performance of 35.36 is obtained. We use this two techniques together by default.

4.2. Different κ in Hard Score Integration

In this part, we study to influence of the number of the selected categories, denoted as κ, in the hard score integration step. As shown in Tab. 2, the selection and ranking on interaction categories works best when $\kappa = 70$. For a smaller κ, some categories may be filtered by mistake, like when $\kappa = 30$, the performance is only 35.04 mAP, falls behind the optimal setting by 0.15 mAP. When $\kappa = 117$, none of the categories are filtered and only ranking operation is still effective. This results in a little performance drop of 0.16 mAP. We use $\kappa = 70$ by default.

5. Additional Qualitative Results

In Fig. 1, we provide more qualitative results in addition to the cases in the paper.

The first case shows the proposed method uses the feature of other instances in the image to help the recognition of a small and challenging instance. The image contains multiple instances of person directing and inspecting an airplane. The TP instance visualized is associated with a small and occluded person, which the baseline fails to discover (the score is denoted as "-"). The proposed method successfully predict this instance with a high score of 0.46. This is consistent with the quantitative discussion on Fig. 5 in the paper.

In the second case, the proposed method discovers an interaction category "repair" neglected by the baseline, possibly with the help of correlations between categories ("inspect" and "repair"). The "repair" interaction is semantically abstract, but the existence of "inspect" may help. This may explain why removing the self-attention from our decoder with cause performance drop in Tab. 6 of the paper: it may learns the dependencies between different interaction categories. In the forth case, the learning of "cut with" interaction may also benefit from the recognition of "hold". Additionally, there is an obvious annotation mistake in the second case: interactions like "ride" and "sit on" are not labeled though they exists (in the background). The proposed method still produces relatively high image-level classification scores for these two categories. Actually, such annotation mistakes exists widely in HICO-DET dataset, and the increase on mAP may not fully show the effectiveness of the proposed method.

In the third and forth cases, our method shows its ability to distinguish whether instances belonging to an interaction category existing in the image. In the third image, though it produce a relatively high "hold" score of 0.36 at image level, it

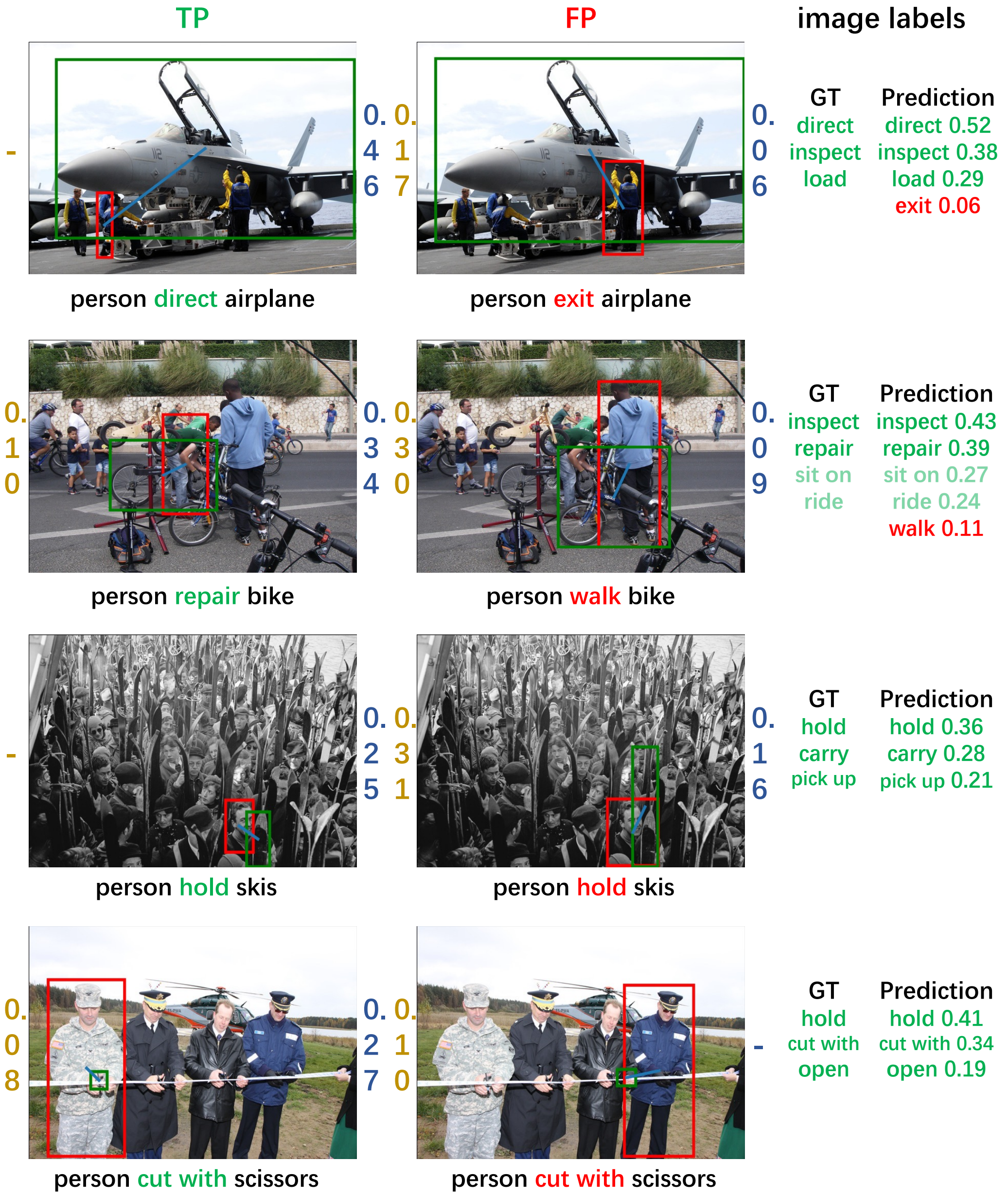

Figure 1. More qualitative comparison between the baseline and the proposed method on HICO-DET. From left to right, column 1: true positive (TP) detection results, whose interaction score is increased by the proposed method; column 2: false positive (FP) detection results, whose interaction score is decreased by the proposed method; column 3: corresponding image-level GT and predictions by the proposed method. Scores on the left and right of an image are the interaction classification scores of the visualized instance in the image from the baseline and the proposed method. "-" for score denotes a instance not discovered (thus no scores). Best viewed in color.

does not take all the H-O pairs in the image as "hold", which would be very wrong. It successfully discovers the TP "hold" instance that the baseline missed, and suppresses the FP "hold" from 0.31 to 0.16. This is consistent with the quantitative results in Tab.4 of the paper that shows the proposed method benefits more from the *adaptive* instance classification weight rather than simply an image classification task. Notably, these two are challenging images with "dense" interaction instances, especially the third case, which corresponds to the discussion in Fig. 5 and Sec 5.5 of the paper.

6. Potential Limitation and Social Impact

The proposed method focuses on the interaction classification sub-task in HOI detection. It does not improve H-O pair detection directly. In the future, we will try to extend this idea to the classification of human and objects in HOI to improve H-O pair detection.

The proposed algorithm has no evident negative impact to society. However, someone might use this method for malicious usage, e.g., to attack people in military usage or invasion of privacy with surveillance. Therefore, we encourage well-intended application of the proposed method.

References

[1] Y.-W Chao, Y Liu, X Liu, H Zeng, and J Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2

[2] C Gao, J Xu, Y Zou, and J.-B Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 2

[3] S Gupta and J Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[4] H He, Y Yuan, X Yue, and H Hu. Rankseg: Adaptive pixel classification with image category ranking for segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 682–700. Springer Nature Switzerland Cham, 2022. 1, 2

[5] Y Liao, A Zhang, M Lu, Y Wang, X Li, and S Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 2

[6] M Tamura, H Ohashi, and T 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. 2

[7] F. Z Zhang, D Campbell, and S Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 2