Supplementary Material for Towards Hard-Positive Query Mining for DETR-based Human-Object Interaction Detection

Xubin Zhong[1], Changxing Ding[1,2]⋆, Zijian Li[1], and Shaoli Huang[3]

[1] School of Electronic and Information Engineering, South China University of Technology , Guangzhou, China
[2] Tencent AI-Lab, Shenzhen, China
eexubin@mail.scut.edu.cn, chxding@scut.edu.cn,
eezijianli@mail.scut.edu.cn, shaolihuang@tencent.com

This supplementary material includes four sections. Section A conducts ablation study on the value of some hyper-parameters in our approach. Section B illustrates the structure and convergence curve of applying our methods to HOTR [1] and CDN-S [2]. Section C presents the complete comparison results. Section D provides qualitative comparison results for QPIC [3] and QPIC + HQM.

A Ablation Study on Hyper-parameters

A.1 Ablation Study on the Value of IoUs in GBS

Experiments are conducted on the HICO-DET database [4]. Results are summarized in Table 1. It is shown that GBS achieves the best performance when the IoUs are randomly sampled within [0.4, 0.6].

Table 1. Ablation study on the value of IoUs in GBS in DT Mode of HICO-DET.

IoUs	Full	Rare	Non-rare
[0.3, 0.5]	30.30	24.63	32.01
[0.4, 0.6]	**30.57**	**24.64**	**32.34**
[0.5, 0.7]	30.35	24.55	32.08

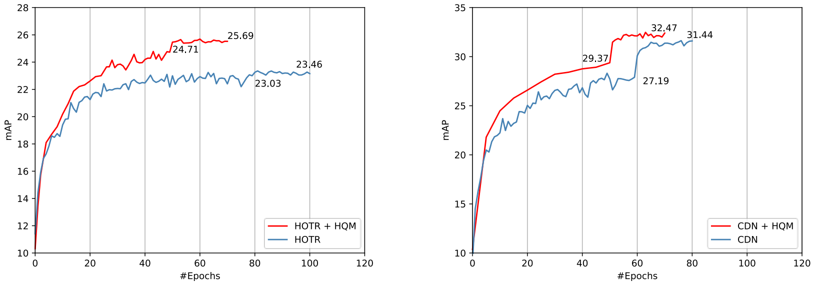
Table 2. Ablation study on the value of K and γ in DT Mode of HICO-DET.

K	γ	Full	Rare	Non-rare
100	0.2	30.46	25.30	32.00
100	0.4	**30.58**	**25.48**	**32.10**
100	0.6	30.38	24.74	32.07
80	0.4	30.39	25.34	31.90
100	0.4	**30.58**	**25.48**	**32.10**
120	0.4	30.24	24.30	32.00

A.2 Ablation Study on the Value of K and γ in AMM

Experiments are conducted on the HICO-DET database and the results are tabulated in Table 2. We observe that AMM achieves the best performance when K and γ are set as 100 and 0.4, respectively.

⋆ Corresponding author.



(a) (b)

Fig. 1. The mAP and training convergence curves on HICO-DET. (a) Results for HOTR and HOTR + HQM. (b) Results for CDN-S and CDN-S + HQM.

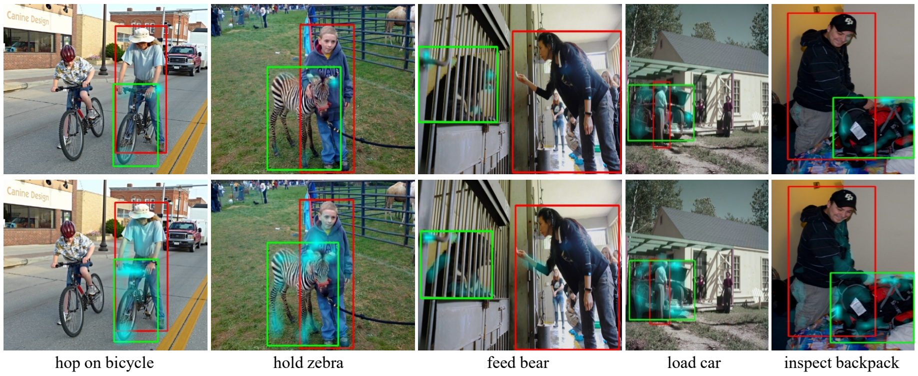


Fig. 2. Visualization of HOI detection results and the cross-attention maps in one decoder layer on HICO-DET. Images in the first and second rows represent results for QPIC and QPIC+HQM, respectively. Best viewed in color.

B Application to HOTR and CDN-S

HQM is plug-and-play and can be readily applied to many DETR-based HOI detection methods. We here present more results on applying HQM to HOTR [1] and CDN-S [2].

HOTR is composed of a CNN backbone, a transformer encoder, an instance decoder, an interaction decoder, and interaction detection heads. HOTR performs human-object pair detection and interaction prediction in parallel branches with independent queries. We mainly apply HQM to its interaction detection branch to overcome the weight-fixed query problem.

Compared with HOTR, CDN-S formulates human-object pair detection and interaction prediction as two successive steps. Decoder embeddings produced by the former step are adopted as queries for the latter one. Therefore, queries for the interaction decoder have been adaptive rather than weight-fixed. Accordingly, we apply HQM to CDN's decoder layers for human-object pair detection, which adopt weight-fixed queries.

Table 3. Performance comparisons on HICO-DET. ‘A’, ‘P’, ‘S’, and ‘L’ represent the appearance feature, human pose feature, spatial feature, and language feature, respectively.

					DT Mode			KO Mode	
	Methods	Feature	Backbone	Detector	Full	Rare	Non-Rare	Full	Rare	Non-Rare
CNN-based	InteractNet [11]	A	ResNet-50-FPN	COCO	9.94	7.16	10.77	-	-	-
	UnionDet [10]	A	ResNet-50-FPN	HICO-DET	17.58	11.72	19.33	19.76	14.68	21.27
	PD-Net [5]	A+S+P+L	ResNet-152	COCO	20.81	15.90	22.28	24.78	18.88	26.54
	DJ-RN [12]	A+S+P+L	ResNet-50	COCO	21.34	18.53	22.18	23.69	20.64	24.60
	PPDM [8]	A	Hourglass-104	HICO-DET	21.73	13.78	24.10	24.58	16.65	26.84
	GGNet [9]	A	Hourglass-104	HICO-DET	23.47	16.48	25.60	27.36	20.23	29.48
	VCL [7]	A+S	ResNet-101	HICO-DET	23.63	17.21	25.55	25.98	19.12	28.03
DETR-based	HOTR [1]	A	ResNet-50	COCO	23.46	16.21	25.62	-	-	-
	HOI-Trans [14]	A	ResNet-50	HICO-DET	23.46	16.91	25.41	26.15	19.24	28.22
	AS-Net [10]	A	ResNet-50	HICO-DET	28.87	24.25	30.25	31.74	27.07	33.14
	QPIC [3]	A	ResNet-50	HICO-DET	29.07	21.85	31.23	31.68	24.14	33.93
	ConditionDETR [13]	A	ResNet-50	HICO-DET	29.65	22.64	31.75	32.11	24.62	34.34
	CDN-S [2]	A	ResNet-50	HICO-DET	31.44	27.39	32.64	34.09	29.63	35.42
	HOTR [1] + **HQM**	A	ResNet-50	COCO	**25.69**	**24.70**	**25.98**	**28.24**	**27.35**	**28.51**
	QPIC [3] + **HQM**	A	ResNet-50	HICO-DET	**31.34**	**26.54**	**32.78**	**34.04**	**29.15**	**35.50**
	CDN-S [2] + **HQM**	A	ResNet-50	HICO-DET	**32.47**	**28.15**	**33.76**	**35.17**	**30.73**	**36.50**

We will release codes to show more details of applying HQM to HOTR and CDN-S. As shown in Fig. 1, HQM not only promotes the mAP accuracy of both models, but also significantly accelerates their training convergence rates.

C Performance Comparisons on HICO-DET

We here present the complete comparisons between our method and state-of-the-arts on HICO-DET in Table 3.

D Qualitative Visualization Results

Fig. 2 provides more qualitative comparisons between QPIC [3] and QPIC + HQM in terms of cross-attention maps and HOI detection results on HICO-DET. We can observe that HQM enables QPIC to capture more discriminative image areas for HOI prediction.

References

1. B. Kim, J. Lee, J. Kang, E. Kim, and H. Kim. HOTR: End-to-End Human-Object Interaction Detection with Transformers. In: CVPR (2021) 1, 2, 3
2. A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li. Mining the Benefits of Two-stage and One-stage HOI Detection. In: NeurIPS (2021) 1, 2, 3
3. M. Tamura, H. Ohashi, and T. Yoshinaga. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In: CVPR (2021) 1, 3
4. Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In: WACV (2018) 1
5. X. Zhong, C. Ding, X. Qu, and D. Tao. Polysemy deciphering network for human-object interaction detection. In: ECCV (2020) 3
6. B. Kim, T. Choi, J. Kang, and H. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In: ECCV (2020) 3
7. Z. Hou, X. Peng, Y. Qiao, and D. Tao. Visual Compositional Learning for Human-Object Interaction Detection. In: ECCV (2020) 3
8. Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020) 3
9. X. Zhong, X. Qu, C. Ding, and D. Tao. Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection. In: CVPR (2021) 3
10. B. Kim, T. Choi, J. Kang, and H. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In: ECCV (2020) 3
11. G. Gkioxari, R. Girshick, Detecting and recognizing human-object interactions. In: CVPR (2018) 3
12. Y. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu. Detailed 2D-3D Joint Representation for Human-Object Interaction. In: CVPR (2020) 3
13. D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, J. Wang. Conditional DETR for Fast Training Convergence. In: ICCV (2021) 3
14. C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, End-to-end human object interaction detection with hoi transformer. In: CVPR (2021) 3