Supplementary Material for Distillation Using Oracle Queries for Transformer-based Human-Object Interaction Detection

Xian Qu[1] Changxing Ding[1,2*] Xingao Li[1] Xubin Zhong[1] Dacheng Tao[3]
[1] South China University of Technology [2] Pazhou Lab, Guangzhou [3] The University of Sydney
eequxian.scut@mail.scut.edu.cn, chxding@scut.edu.cn
{eexingao,eexubin}@mail.scut.edu.cn, dacheng.tao@gmail.com

This supplementary material includes seven sections. Section A illustrates the structure and convergence curve of applying our methods to CDN-S [37] and HOTR [26], respectively. We provide more experimental results about initial decoder embeddings of DOQ in Section B. Section C shows the convergence curves of applying DOQ and CCS to QPIC [23] one by one. Section D conducts ablation study on the value of some hyper-parameters in our model. The complete results of performance comparisons on HICO-DET [48] are presented in Section E. Section F visualizes the HOI detection results of our model.

A. Application to CDN-S and HOTR

We will release the code of the application of our methods to CDN-S [37] and HOTR [26] to verify that DOQ and CCS are both portable and flexible.

CDN-S consists of a CNN backbone, a transformer encoder, a human-object pair decoder, an interaction decoder, and interaction detection heads. The two decoders are organized in a cascaded manner, which are responsible for human-object pair predictions and interaction predictions, respectively. When applying DOQ to CDN-S during training, we construct a teacher network for the interaction decoder only. The same as the application of DOQ to QPIC [23], we adopt the ground-truth positions and object word embeddings of labeled human-object pairs to construct a set of oracle HOI queries $\boldsymbol{Q}_t$ and initial decoder embeddings $\boldsymbol{D}_{t_0}$, respectively. As illustrated in Figure 1, there is clear performance gain for CDN-S with our proposed methods.

HOTR consists of a CNN backbone, a transformer encoder, an instance decoder, an interaction decoder, and interaction detection heads. The two decoders are organized in a parallel way, which are responsible for object detection and interaction detection, respectively. Following the original settings in HOTR, we fix the network parameters of CNN backbone, encoder, and instance decoder during training. When applying DOQ to HOTR during training, we construct a teacher network for the interaction decoder only. The way to construct $\boldsymbol{Q}_t$ and $\boldsymbol{D}_{t_0}$ are the same as those for QPIC and CDN-S. The HOI detection results and convergence curve are shown in Figure 2. Our model achieves better mAP accuracy with considerably fewer training epochs.

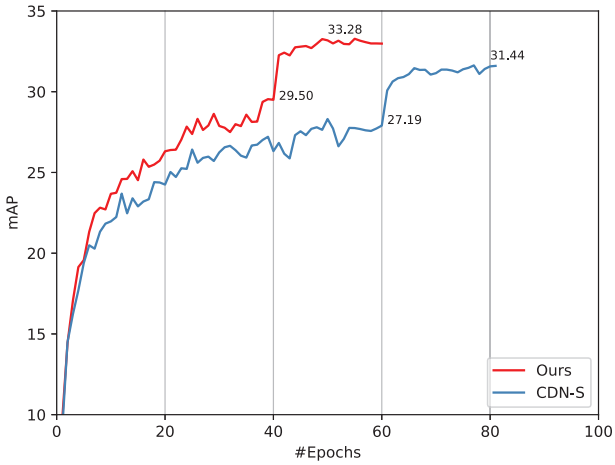

Figure 1. The mAP and convergence curves for CDN-S [37] and our model on HICO-DET [48]. Our model achieves better mAP accuracy with fewer training epochs.

B. More Experimental Results on $\boldsymbol{D}_{t_0}$

In DOQ, we generate the initial decoder embedding $\boldsymbol{D}_{t_0}$ according to the word embedding of the ground-truth object category involved in each labeled human-object pair. Table 3b in the main paper shows that both the object embeddings and verb-class vectors promote the performance compared to the usage of zero vectors. We here try to employ both of them via vector concatenation and the results are shown in Table 1.

It is shown that their combination brings in only slight performance gain. Therefore, we here only adopt object embeddings in order to reduce network parameters.

*Corresponding author.

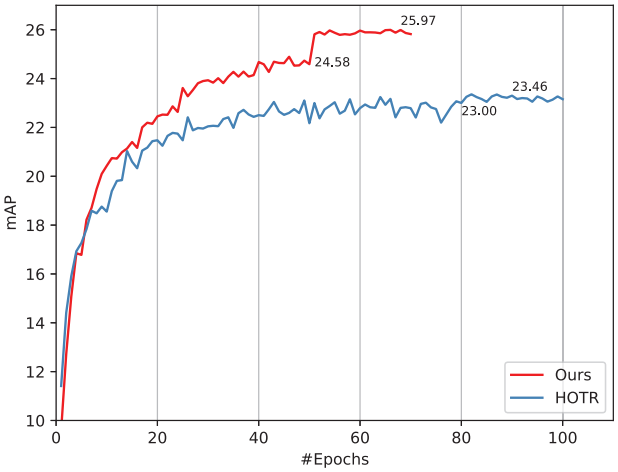

Figure 2. The mAP and convergence curves for HOTR [26] and our model on HICO-DET [48]. Our model achieves better mAP accuracy with fewer training epochs.

Table 1. Ablation study on initial decoder embeddings.

Methods	Full	Rare	Non-rare
verb-class vectors	30.10	24.23	31.85
object embeddings	30.41	25.10	**32.00**
both	**30.43**	**25.56**	31.88

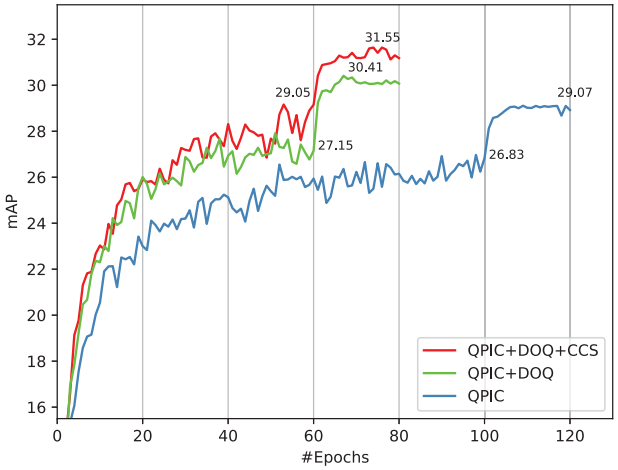

Figure 3. The mAP and convergence curves for the original QPIC model [23], QPIC with DOQ, and QPIC with both DOQ and CCS on the HICO-DET database [48]. Benefited from DOQ, the convergence rate of QPIC can be significantly accelerated.

C. Convergence Curves by DOQ Alone

The mAP and convergence curves for the original QPIC model [23], QPIC with DOQ, and QPIC with both DOQ and CCS are presented in Figure 3. It is shown that the accelerated convergence rate is mainly due to the application of DOQ.

Table 2. Ablation study on the value of α_1 and α_2.

α_1	α_2	Full	Rare	Non-Rare
0.1	10	31.30	26.14	32.84
1	10	**31.55**	**26.75**	**32.99**
10	10	31.41	26.72	32.82
1	5	31.38	25.98	32.99
1	10	**31.55**	**26.75**	**32.99**
1	15	31.15	25.55	32.82

D. Ablation Study on Hyper-parameters

Ablation Study on the Value of α_1 and α_2 in DOQ. Experiments are conducted on the HICO-DET [48] database. The experimental results are summarized in Table 2. We can observe that our model achieves the best performance when α_1 and α_2 are set as 1 and 10, respectively.

Ablation Study on the Value of γ in CCS. Experiments are conducted on the HICO-DET [48] database. The experimental results are summarized in Table 3. We can observe that our model achieves the best performance when γ is set as 0.25.

Table 3. Ablation study on the value of γ.

γ	Full	Rare	Non-Rare
0.15	31.29	26.17	32.81
0.25	**31.55**	**26.75**	**32.99**
0.35	30.81	25.36	32.44

E. Performance Comparisons on HICO-DET

We here present the complete comparisons between our method and state-of-the-arts on both DT and KO modes of HICO-DET [48] in Table 4.

F. Qualitative Visualization Results

Figure 4 presents more qualitative comparisons between our model and QPIC [23] in terms of attention maps and HOI detection results on HICO-DET [48]. It is shown that our method produces more discriminative attention maps.

We also present some failure cases of our method in terms of HOI detection in Figure 5.

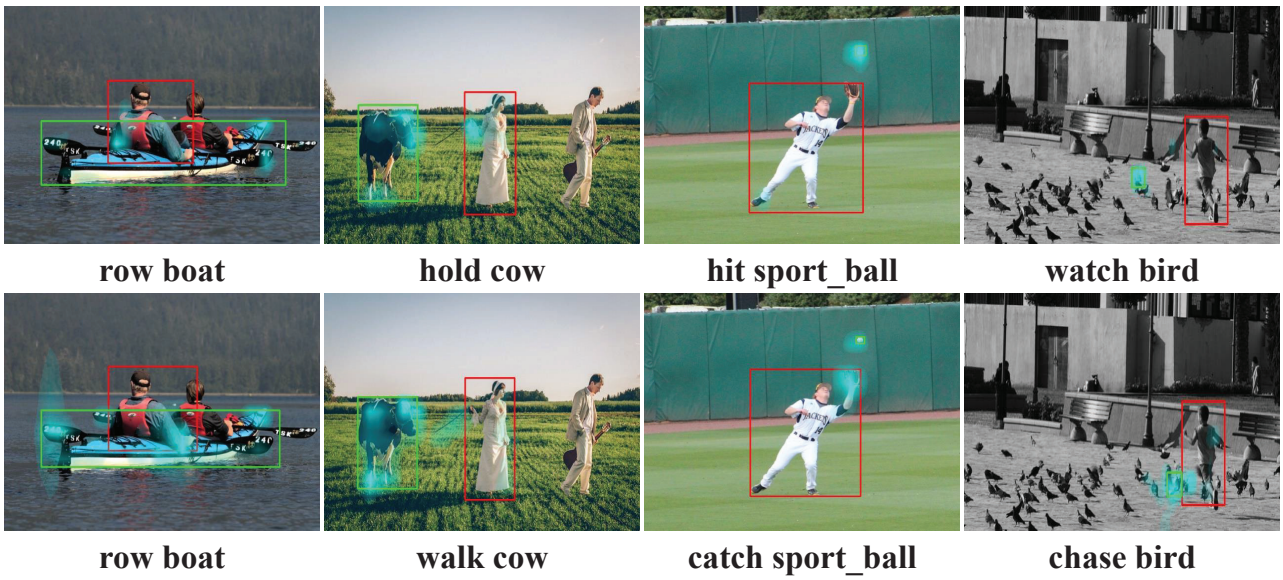

Figure 4. Visualization of HOI detection results and attention maps in decoder layers. The two rows represent results for QPIC [23] and our method, respectively.

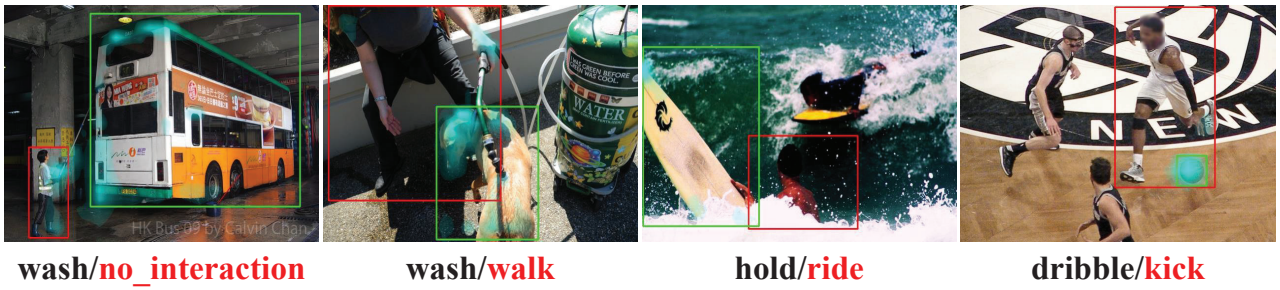

Figure 5. Failure cases of our model for HOI detection on HICO-DET [48]. The ground-truth and the predicted one are typed in black and red, respectively.

Table 4. Performance comparisons on HICO-DET.

	Method	Detector	Backbone	Default Mode			Known object Mode		
				full	rare	non-rare	full	rare	non-rare
Two-Stage	SG2HOI [18]	COCO	ResNet-50	20.93	18.24	21.78	24.83	20.52	25.32
	DJ-RN [33]	COCO	ResNet-50	21.34	18.53	22.18	23.69	20.64	24.60
	SCG [20]	COCO	ResNet-50-FPN	21.85	18.11	22.97	–	–	–
	ConsNet [39]	COCO	ResNet-50-FPN	22.15	17.12	23.65	–	–	–
	PastaNet [35]	COCO	ResNet-50	22.65	21.17	23.09	24.53	23.00	24.99
	IDN [36]	COCO	ResNet-50	23.36	22.47	23.63	26.43	25.01	26.85
	DRG [47]	HICO-DET	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
	IDN [36]	HICO-DET	ResNet-50	24.58	20.33	25.86	27.89	23.64	29.16
One-Stage	IP-Net [30]	COCO	Hourglass-104	19.56	12.79	21.58	22.05	15.77	23.92
	HOTR [26]	COCO	ResNet-50	23.46	16.21	25.62	–	–	–
	ASNet [24]	COCO	ResNet-50	24.40	22.39	25.01	27.41	25.44	28.00
	GGNet [22]	HICO-DET	Hourglass-104	23.47	16.48	25.60	27.36	20.23	29.48
	PST [19]	HICO-DET	ResNet-50	23.93	14.98	26.60	26.42	17.61	29.05
	HOI-Trans [25]	HICO-DET	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57
	ASNet [24]	HICO-DET	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
	QPIC [23]	HICO-DET	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
	QPIC [23]	HICO-DET	ResNet-101	29.90	23.92	31.69	32.38	26.06	34.27
	CND-S [37]	HICO-DET	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
	Ours (HOTR)	COCO	ResNet-50	**25.97**	**26.09**	**25.93**	**28.23**	**28.22**	**28.23**
	Ours (QPIC)	HICO-DET	ResNet-50	**31.55**	**26.75**	**32.99**	**34.15**	**29.62**	**35.50**
	Ours (QPIC)	HICO-DET	ResNet-101	**31.80**	**25.95**	**33.55**	**34.42**	**28.07**	**36.32**
	Ours (CDN-S)	HICO-DET	ResNet-50	**33.28**	**29.19**	**34.50**	**36.11**	**31.61**	**37.45**