DETRDistill: A Universal Knowledge Distillation Framework for DETR-families —Supplementary Material—

Jiahao Chang[1*] Shuo Wang[1*] Hai-Ming Xu[2*] Zehui Chen[1] Chenhongyi Yang[3] Feng Zhao[1†]
[1]University of Science and Technology of China [2]University of Adelaide [3]University of Edinburgh
{changjh,shuowang2323,lovesnow}@mail.ustc.edu.cn hai-ming.xu@adelaide.edu.au
chenhongyi.yang@ed.ac.uk fzhao956@ustc.edu.cn

In this supplementary material, we first present the performance of our approach for distillation on SOTA detectors and other dataset. Next, we conduct more ablation studies for our approach to validate its effectiveness.

1. Distillation on SOTA Detectors

In the main paper, we have explored knowledge distillation on three baseline detectors. In this section, we perform KD on DINO [2] which is a stronger variant of DN-DETR. Results in Table. 1 show that the student model's performance is consistently improved with our proposed DETRDistill method. Furthermore, such a KD design can also be applicable when the number of queries in teacher and student is inconsistent, as shown in bottom block of Table. 1.

Setting	Query	Backbone	AP	AP_S	AP_M	AP_L
Teacher	900	R-101	50.6	32.7	53.8	65.0
Student	900	R-18	46.1	28.7	48.5	60.0
Ours	-	-	**47.9**(+1.8)	**30.5**	**51.2**	**62.0**
Teacher	900	R-101	50.6	32.7	53.8	65.0
Student	300	R-50	48.6	31.5	51.4	63.1
Ours	-	-	**50.1**(+1.5)	**32.3**	**53.4**	**64.8**

Table 1. Results of our DETRDistill on DINO detector.

2. Distillation on Other Dataset

We also provide experiment on the PASCAL VOC dataset [1] and the performance in Table. 2 shows that our DETRDistill gains 3.50 mAP and 2.36 mAP over the baselines, which indicate that DETRDistill is applicable across different datasets.

3. More Ablation Studies

Apart from the ablation studies presented in the main paper, we further provide more for our proposed approach.

Setting	Backbone	AP50	mAP
Teacher	R-101	81.00	80.99
Student	R-18	75.50	75.49
Ours	-	**79.00**(+3.50)	**78.99**(+3.50)
Teacher	R-101	81.00	80.99
Student	R-50	79.30	79.25
Ours	-	**81.60**(+2.30)	**81.61**(+2.36)

Table 2. Results on PASCAL VOC dataset. Train: trainval set of VOC2007& VOC2012; Eval: VOC2007 test.

Importance of teacher's assignment in Query-prior Assignment Distillation In the proposed Query-prior Assignment Distillation module, the well-trained query set of the teacher model will be fed into the student model as an additional group of prior queries and the teacher's corresponding bipartite assignment will also be used for the distillation loss calculation in Eq. (10) of the main paper. Since the teacher's queries are well-optimized and the number of queries used in the student model is increased, we wonder whether the performance gain simply comes from the use of high-quality queries or the increased number of queries.

To verify the conjectures, we conduct experiments on two kinds of variants of the proposed module, i.e., (1) directly initializing the student's queries with the teacher's query set and the bipartite assignment of the teacher model is not used, we term this variant as Teacher Init. (2) only introducing the teacher's query set as an additional group of queries for the student model without the corresponding teacher's bipartite assignment and we term this variant as Teacher Group.

As the results presented in Table. 4, the Teacher Init variant does not bring performance improvement which proves that the initial query's quality is not the main factor. Meanwhile, the Teacher Group variant only obtains an insignificant performance gain which verifies that the naive increasing query numbers may not be enough. However, our proposed method (Teacher Assigned) achieves better detection

	AdaMixer		Deformable DETR		Conditional DETR	
	Tea-R101	Stu-R50	Tea-R101	Stu-R50	Tea-R101	Stu-R50
Model Params number (M)	153.56	134.57	58.78	39.84	62.13	43.19
Basic Compute Cost (GFLOPs)	178.95	102.88	287.34	192.26	171.4	95.32
KD Compute Cost (GFLOPs)	24.49		51.10		18.35	
Proportion of KD Cost	7.99 %		9.63 %		6.43 %	

Table 3. Comparison of the number of model parameters and computation cost on various detectors. Proportion of KD Cost is defined as KD Compute Cost / (Basic Compute Cost of the Teacher + Basic Compute Cost of the Student + KD Compute Cost).

accuracy which shows the importance of incorporating the teacher's assignment in the Query-prior Assignment Distillation module.

Teacher	Student	Teacher Init	Teacher Group	Teacher Assigned
43.6	42.3	42.3(+0.0)	42.5(+0.2)	**42.9(+0.6)**

Table 4. Comparison of different variants for Query-prior Assignment Distillation module.

Analysis to the computation consumption in training We are interested to know the computational cost of our proposed distillation modules at the training phase. Since various detectors have different model architectures, different Flops are required and we report detector-specific computation consumption in Table. 3. It is clear that our proposed KD module only takes a small proportion of computation consumption in the whole model optimization which verifies the efficiency of our proposed approach.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1

[2] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1