CrossKD: Cross-Head Knowledge Distillation for Dense Object Detection Supplementary Material

1. Details of Distillation Losses

According to the task of detection heads, *i.e.*, classification, and regression, we apply different distance functions $\mathcal{D}_{\text{pred}}$ to transfer task-specific information in different branches. In this section, we introduce the details of distance functions $\mathcal{D}_{\text{pred}}$ applied in CrossKD.

Regression Branch. There are mainly two types of regression branches that existed in dense detectors. The first regression branch directly regresses the bounding boxes from the anchor boxes (*e.g.*, RetinaNet [5], ATSS [10]) or points (*e.g.*, FCOS [8]). In this case, we directly use GIoU [7] as $\mathcal{D}_{\text{pred}}$, which is defined as:

$$\mathcal{D}_{\text{pred}}(\mathcal{B}, \mathcal{B}') = \frac{|\mathcal{B} \cap \mathcal{B}'|}{|\mathcal{B} \cup \mathcal{B}'|} - \frac{|\mathcal{C} \setminus (\mathcal{B} \cup \mathcal{B}')|}{|\mathcal{C}|}, \tag{1}$$

where $\mathcal{B}$ and $\mathcal{B}'$ represent the predicted and ground-truth bounding boxes and $\mathcal{C}$ is the smallest enclosing convex object for $\mathcal{B}$ and $\mathcal{B}'$.

In the other situation, the regression branch predicts a vector to represent the distribution of box location (*e.g.*, GFL [4]), which contains richer information than the Dirac distribution of the bounding box representation. To efficiently distill the knowledge of location distribution, we employ the same $\mathcal{D}_{\text{pred}}$ like LD [11], which is defined as:

$$\mathcal{D}_{\text{pred}}(\boldsymbol{p}, \boldsymbol{p}') = \text{KL}(s(\boldsymbol{p}/\tau), s(\boldsymbol{p}'/\tau)), \tag{2}$$

where KL means KL divergence, $s(\cdot)$ indicates the Softmax function, and τ is a factor to smooth the distribution.

Classification Branch. Distillation in the classification branch severely suffers from the imbalance of the foreground and background instances problem. To avoid training crash, previous prediction mimicking methods usually design complicated region selection principle to choose effective areas. In contrast, without selecting effective regions, we regard the classification scores predicted by the teacher as the soft labels and directly use Quality Focal Loss (QFL) proposed in GFL [4] to pull close the teacher-student distance. We define $\mathcal{D}_{\text{pred}}$ in the classification branch as:

$$\mathcal{D}_{\text{pred}}(\boldsymbol{p}, \boldsymbol{p}') = (|\sigma(\boldsymbol{p}) - \sigma(\boldsymbol{p}')|)^{\gamma} \cdot \text{BCE}(\sigma(\boldsymbol{p}), \sigma(\boldsymbol{p}')), \tag{3}$$

Loss	Region	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
-	-	35.8	53.1	38.2	18.9	38.9	47.9
BCE	P	36.3	53.8	39.1	19.1	39.6	48.3
BCE	N	36.2	53.5	38.9	19.3	40.0	48.2
BCE	P+N	36.9	54.3	39.5	20.0	40.7	48.4
QFL	P+N	38.7	56.3	41.6	21.1	42.2	51.5

Table 1. Effectiveness of different distillation losses in classification branch. 'BCE' and 'QFL' means the binary cross entropy loss and quality focal loss, respectively. 'P' and 'N' refer to the positive and negative regions. The teacher-student pair is GFL with ResNet-50 and ResNet-18 backbones.

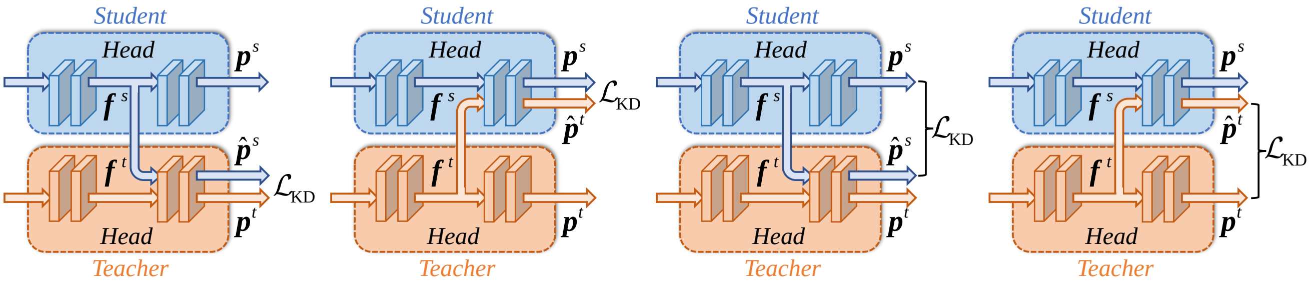
where σ denotes the sigmoid function and BCE indicates binary cross entropy. $(|\sigma(\boldsymbol{p}) - \sigma(\boldsymbol{p}')|)^{\gamma}$ serves as a modulating factor added to the cross entropy function, with a tunable focusing parameter $\gamma \geq 0$. Here, γ is set as 1 in all experiments, which we find is the optimum.

We also compare the performance of QFL with the widely used BCE loss. As shown in Tab. 1, The BCE loss can receive 36.3 and 36.2 AP when separately applied on the positive and negative regions. When we perform distillation on both positive and negative regions, BCE loss can only achieve 36.9 AP, far below 38.7 AP of QFL, which demonstrates the effectiveness of the current distillation losses.

2. The Generalization Ability of CrossKD

CrossKD is adaptable for any detector distillation since the target conflict is a common problem of object detection distillation due to imperfect teacher predictions. To demonstrate the generalization, we apply CrossKD on detectors with various types of backbones and structures.

The results of our CrossKD on a series of lightweight students distilled with GFL with ResNet-18, ResNet-34, and ResNet-50 backbones are presented in Tab. 2. We apply ResNet-101 as the backbone for the teacher detector. As shown in Tab. 2, our method can effectively enhance the performance of all given lightweight detectors. Specifically, CrossKD achieves stable improvements for the students with ResNet-18, ResNet-34, and ResNet-50 back-



(a) CrossKD (b) (c) (d)

Figure 1. Different cross-head strategies. (a) is the original strategy used in CrossKD. (b) delivers the intermediate features of the teacher to the student head and conducts KD between the cross-head predictions of the teacher and the student's predictions. (c) does the same cross-head strategy as (a) but performs KD between the student's original predictions and cross-head predictions. (d) does the same cross-head strategy as (b) but performs KD between the teacher's original predictions and the cross-head predictions.

Student	**CrossKD**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-18		35.8	53.1	38.2	18.9	38.9	47.9
	✓	39.2	57.0	42.2	22.7	43.0	51.3
ResNet-34		38.9	56.6	42.2	21.5	42.8	51.4
	✓	42.4	60.4	45.8	24.4	46.8	55.6
ResNet-50		40.2	58.4	43.3	23.3	44.0	52.2
	✓	43.7	62.1	47.4	26.9	48.0	56.2

Table 2. Quantitative results of CrossKD for lightweight detectors. Standard 1× schedule is applied in all experiments. The teacher detector is GFL with ResNet-101 backbones.

bones, which reach 39.2 AP, 42.4 AP, and 43.7 AP.

Furthermore, we adapt CrossKD to typical Faster R-CNN (two-stage) and Deformable DETR (DETR-like) detectors and report their performance in Tab. 3. In Faster R-CNN, we deliver the student region features to the R-CNN head of the teacher to generate cross-head predictions to accept the teacher's supervision. In Deformable DETR, the cross-head predictions are created by passing the encoder features of the student into each stage of the teacher decoder. As shown in Tab. 3, without finely tuned hyper-parameters, CrossKD boosts the accuracy of ResNet-18 based Faster R-CNN and Deformable DETR to 35.5 (2.0 ↑) and 45.8 (1.7↑) AP, which demonstrates the generalization ability of CrossKD.

Method	Schedule	AP	AP_{50}	AP_{75}
Faster R-CNN R18 (S)	12e	33.5	53.7	35.9
Faster R-CNN R50 (T)	12e	37.4	58.1	40.4
CrossKD	12e	**35.5** (2.0↑)	55.8	38.0
Deform. DETR R18 (S)	50e	44.1	62.8	47.9
Deform. DETR R50 (T)	50e	47.0	66.1	50.9
CrossKD	50e	**45.8** (1.7↑)	63.8	49.9

Table 3. CrossKD for Faster R-CNN and Deformable DETR.

Strategy	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
-	35.8	53.1	38.2	18.9	38.9	47.9
(a)	38.7	56.3	41.6	21.1	42.2	51.5
(b)	35.4	52.5	37.8	18.6	38.4	47.1
(c)	34.5	51.9	36.7	17.8	37.6	45.1
(d)	32.5	48.8	35.0	16.6	35.0	42.8

Table 4. Comparisons of different cross-head strategies. The strategies (a), (b), (c), (d) have shown in Fig. 1, where (a) is the current strategy used in CrossKD. The teacher-student pair is GFL with ResNet-50 and ResNet-18 backbones.

3. More Ablations

In this section, we experiment different cross-head strategies to demonstrate the effectiveness of our CrossKD, which are illustrated in Fig. 1. As presented in Tab. 4, strategy (b), which differently reuses the student's detection head, achieved only 35.4 AP, significantly lower than the 38.7 AP obtained by CrossKD. We hypothesize that this difference in performance may be attributed to the suboptimal optimization of the student's blocks in this approach. Fig. 1(c) and Fig. 1(d) minimize the distances between the original predictions and the cross-head predictions. However, these strategies have limited impact on the student's backbones, resulting in 34.5 AP and 32.5 AP for Fig. 1(c) and Fig. 1(d), respectively.

Moreover, Fig. 1(b), (c), and (d) all perform distillation losses and detection losses at the student's detection heads, so the target conflict problem still exists. In contrast, CrossKD separates the distillation losses onto the teacher's branch and hence avoids the target conflict problem. As a result, CrossKD receives the highest AP of 38.7 among all cross-head strategies.

4. Relation to Previous Works

In this section, we describe the differences of our method and some related works which are originally designed for the classification task [1–3, 6, 9]. Here, we compare CrossKD with these works from the aspects of motivation and structure to emphasize the differences.

Motivation. Previous works all concentrate on the classification task. For instance, Bai *et al*. [1] aims to alleviate overfitting in few-shot task. Li *et al*. [3] focuses on using a residual network to help a non-residual network overcome gradient vanishing. Some works [6, 9] target on the general KD scenario in classification. These methods all attempts to solve specific problems in classification and are not specially designed for distilling object detectors.

In contrast, CrossKD, which is specially designed for the object detection task, focuses on the target conflict problem in object detection. To our knowledge, this is the first work to discuss the target conflict problem in distilling object detectors. As presented in Sec. 1 of the main paper, the teacher detector usually predicts inaccurate results, which conflict with the ground-truth targets. The traditional KD methods supervise the student detector with those two controversial labels at the same place, resulting in low distillation efficiency. To alleviate this problem, we propose to deliver the intermediate features of the student to the part of the teacher's detection head and generate new cross-head predictions to accept the distillation losses.

However, without the detection-specific design, those methods can not achieve a promising performance.

Structure. Previous works tend to design a complicated manner to utilize the teacher-student latent features. Typically, Li *et al*. [3] forwards every stage features of the student into the teacher's blocks. Liu *et al*. [6] alternately delivers the intermediate features from the student to the teacher or from the teacher to the student. These strategies significantly increase the computational complexity in training phase.

Instead of applying a complicated design, CrossKD is relatively simple, which only passes the student's latent features through part of the teacher's detection head. Despite its simplicity, extensive experiments demonstrate its effectiveness in object detection KD.

5. Result Visualization

We visualize the detection results of the teacher, the student, and our CrossKD in Fig. 2. As the visualization shows, CrossKD usually receives even better results than the teacher detector, which demonstrates that CrossKD can relieve the influence of the teacher's inaccurate predictions and achieve a better optimization towards ground-truths.

References

[1] Haoli Bai, Jiaxiang Wu, Irwin King, and Michael Lyu. Few shot network compression via cross distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3

[2] Animesh Koratana, Daniel Kang, Peter Bailis, and Matei Zaharia. LIT: Learned intermediate representation training for model compression. In *International Conference on Learning Representations*, 2019.

[3] Guilin Li, Junlei Zhang, Yunhe Wang, Chuanjian Liu, Matthias Tan, Yunfeng Lin, Wei Zhang, Jiashi Feng, and Tong Zhang. Residual distillation: Towards portable deep neural networks without shortcuts. *Advances in Neural Information Processing Systems*, 33:8935–8946, 2020. 3

[4] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21002–21012. Curran Associates, Inc., 2020. 1

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[6] Dongyang Liu, Meina Kan, Shiguang Shan, and CHEN Xilin. Function-consistent feature distillation. In *International Conference on Learning Representations*, 2019. 3

[7] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[9] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021. 3

[10] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[11] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9407–9416, June 2022. 1

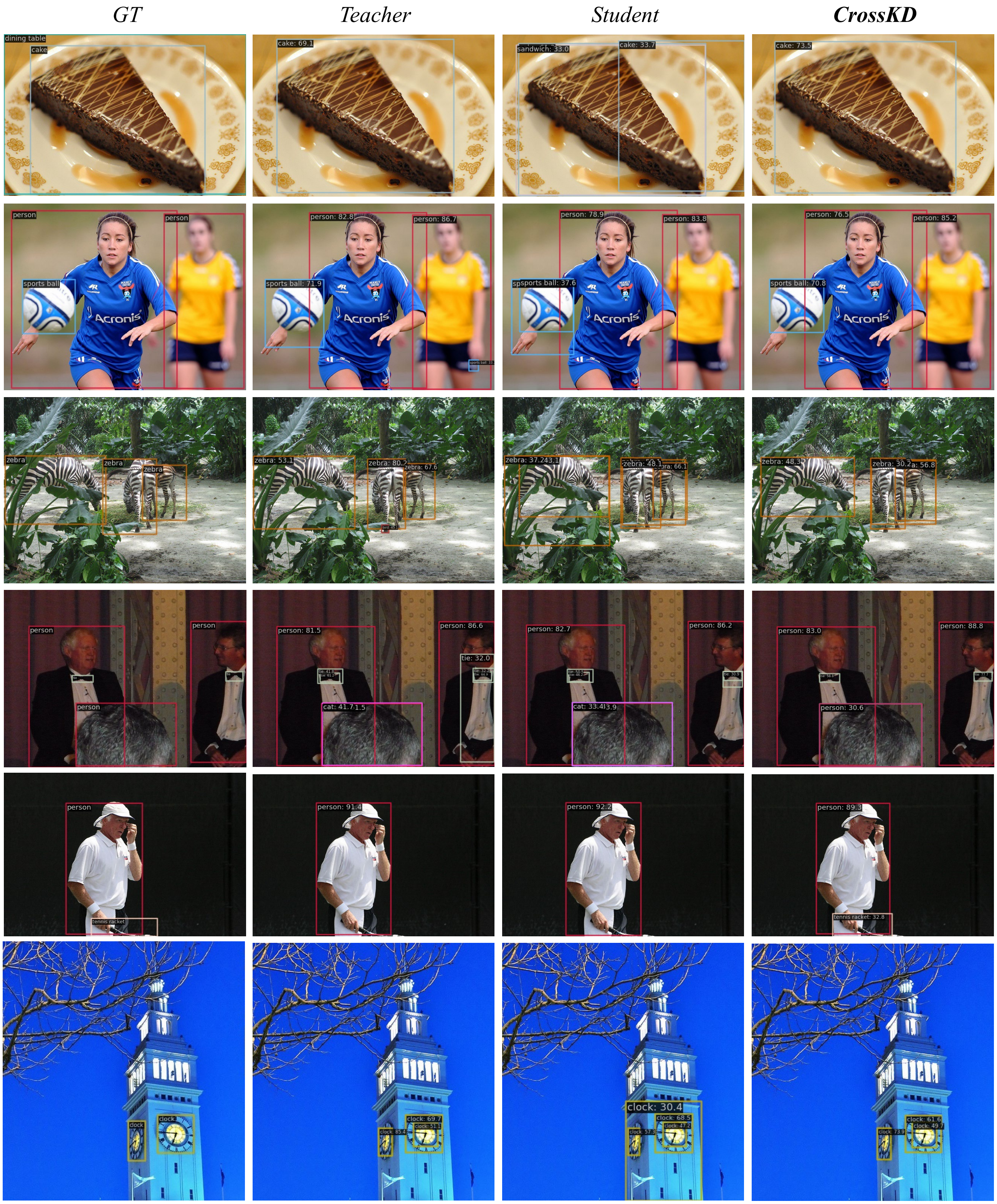


Figure 2. Visualization of detection results of CrossKD. The teacher is GFL-R50 with 40.2 AP and student is GFL-R18 with 35.8 AP.