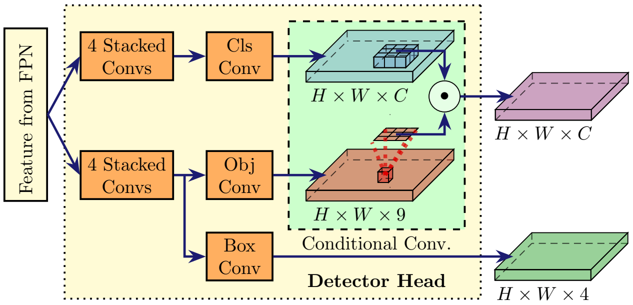


Figure 1: PAA head amended with Conditional Objectness Prediction (COP)

Appendices

A. Conditional Objectness Prediction

A.1. Motivation and Network Architecture

Since object detection performs classification and localization concurrently, their quality must be consistent. For example, a prediction with a high classification probability but low IoU box yields a false positive, while the reverse induces false negative. However, single-stage detectors implement the branches independently, typically each with 4 stacked convolutions. During training and inference, there is no connection between them.

Furthermore, although the two branches have the same computation and feature extraction capacity, the localization receives significantly less training feedback than the classification. This is because most of the samples are negative, which have no box targets for training localization, hence are discarded during gradient backward. Recent methods also add auxiliary branches to predict the localization quality, such as IoU [12, 5], Centerness [6, 14, 11], but is trained only on positive samples. In addition, the positive samples of the same object are often connected and appear in a small local window. However, they are treated independently during non-maxima suppression.

Therefore, we propose adding an auxiliary Conditional Objectness Prediction (COP) to the localization branch. It is similar to the Regional Proposal Network (RPN) of two-stage detector [8] but with renovations, as shown in Fig. 1. Concretely, at each anchor a_i, we predict the objectness scores $\{o_i^k\}_{k=1}^{3\times 3}$ of its 3×3 nearest neighbors to capture their spatial correlation. The final classification probability is the dot product of the objectness $\{o_i^k\}$ and the corresponding 3×3 local window of the classification prediction $\{p_i^k\}_{k=1}^{3\times 3}$

$$p(a_i) = \frac{1}{9}\sum_{k=1}^{3\times 3} o_i^k p_i^k, \tag{1}$$

where o_i and p_i are the confidence score (*i.e.* after Sigmoid) of the objectness branch and classification branch, which are supervised implicitly and mutually through COP product during gradient back-propagation. Therefore, we can fuse and jointly train the branches, make the training consistent with inference. Consequently, all samples in the localization now receive gradient feedback.

Our COP shares a common with the Implicit Object recently introduced in [16, 2], as they are both trained jointly with the classification branch. However, our motivation and implementation are different: (i) We believe features in the regression branch are also helpful to predict objects in the class-agnostic manner, similar to the RPN head in Faster-RCNN, and should not be discarded. COP is introduced to distribute gradient feedback to all samples in the localization branch. (ii) We implement COP as Conditional Convolution [13, 10], where the weights are generated dynamically for each sample. Hence, we can embed the local relationship between the samples to reduce false-positive prediction.

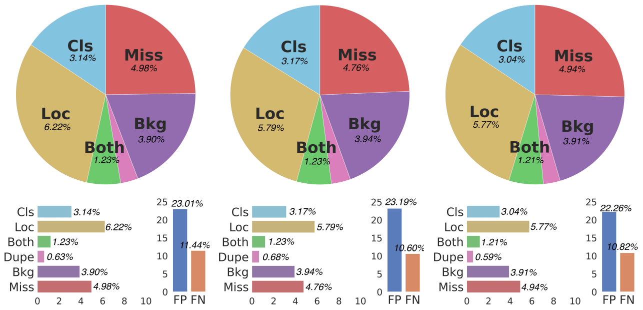
A.2. Ablation Study

We investigate the effectiveness of the Conditional Objectness (COP) with different backbones, including EfficientNet-B0 (Eff-B0) [9], RepVGG-A0 (A0)[3], ResNet18 (R18), and ResNet50 (R50)[4], and compare it with IoU prediction and Implicit Object prediction (IOP). For easy comparison, we use the baseline PAA method, that has IoU prediction by default. Table 1 summarizes the results.

Table 1: Compare different auxiliary predictions: IoU, Implicit Object Prediction (IOP), and Conditional Objectness Prediction(COP) with different backbones. (*) denotes the branch is trained but not used during inference.

Auxiliary Prediction			**mAP**			
IOU	**IOP**	**COP**	**Eff-B0**	**A0**	**R18**	**R50**
✓			32.4	34.0	35.8	40.4
✓	✓		33.4	34.7	36.7	**41.6**
✓		✓	**33.5**	**34.8**	**36.9**	**41.6**
*	✓		33.4	**34.8**	36.7	41.5
*		✓	33.5	**34.8**	**36.9**	**41.6**
	✓		33.3	**34.8**	36.6	41.1
		✓	**33.4**	34.7	**36.9**	**41.2**

At first, we add the IOP or COP to the default PAA head, and observe that both IOP and COP can improve the baseline with considerable margins. For the Eff-B0, A0, R18, R50 backbones, IOP increases $+1, +0.7, +0.9, +1.2$AP, and COP increases $+1.1, +0.8, +1.1, +1.2$AP, respectively. COP and IOP



(a) PAA (b) CoLAD (c) CoLAD-COP

Figure 2: Error analysis using TIDE [1] toolbox of the models PAA, CoLAD, and CoLAD-COP with the same backbone ResNet50 on the MS COCO *minval* set.

perform equally on R50, but COP is slightly better for small backbones Eff-B0($+0.1\mathrm{AP}$), A0 ($+0.1\mathrm{AP}$), and R18 ($+0.2\mathrm{AP}$).

Secondly, we try dropping the IoU prediction during inference and use only IOP or COP, and observe that the results remain almost unchanged (4th and 5th rows).

However, when we train the models without the IoU branch, the performance is dropped more severely for ResNet50 backbone (6th and 7th rows). This proves that IoU is still helpful as deep supervised signal for the regression branch in training, but can be safely omitted during infer.

B. Prediction Error Analysis

Beyond evaluating the mAP metric, we use the TIDE [1] toolbox to analyze the prediction errors of the three models, PAA, CoLAD, and CoLAD-COP, with the same backbone ResNet50.

As shown in Fig. 2, CoLAD and CoLAD-COP help reduce the localization error of the baseline PAA from 6.22% to 5.79% and 5.77%, respectively. CoLAD also reduces the classification error by 3.04%. These indicate that the dynamic mechanism in CoLAD is effective to guide the network to learn a good label assignment, which results in low classification and localization error. In addition, CoLAD can recall more objects, since the false negative percentage is reduced from 11.5% for PAA to 10.85% for both CoLAD and CoLAD-COP. Finally, the introduction of COP can better suppress noisy prediction, as the false positive ratio is reduced from 23.01% to 22.26%.

C. Compare with other distillation methods

Head-to-head comparison of distillation methods for object detection is not easy, since each method is typically developed for a particular detector. Therefore, for reference purpose only, we select LD [15] to compare, since it is based on the SOTA single-stage detector Generalized Focal (GF) [7, 6], which is inline with us but has higher performance than our PAA baseline. However, we emphasize that the two methods address different problems. LD[15] focuses on localization distillation and is applied particularly for GFL detector, while we address the label assignment. Therefore, the two methods can be combined.

Table 2: Compare our LAD techniques to Localization Distillation (LD) [15] for different ResNet backbones. T and S denote teacher and student networks. LAD is based on PAA[5] and LD is based on GF[7]. The results are compared for student networks w.r.t. its baseline on COCO *test* set.

T	S	PAA	LAD	CoLAD	SoLAD	GF	LD
R50	R18	35.8	36.9	36.5	38	36.0	36.1
R101	R18	35.8	36.8	36.6	38.4	35.8	36.5
R101	R50	40.4	41.6	41.3	42.4	40.1	41.1

Table 2 compare the two methods using the same teacher and student's backbones. It is obvious that LAD and CoLAD are superior to LD in all cases. Moreover, our LAD/CoLAD is very simple and can be adapted quickly to any single-stage detectors without architecture modification, and not restricted to Generalized Focal detector [7, 6]. This shows how flexible and effective our method is compared to other distillation methods, such as feature mimicking.

Acknowledgement

We would like to thank Francisco Menendez, Su Huynh, Vinh Nguyen, Nam Nguyen and other colleagues for valuable discussion and helping reviewing the paper.

References

[1] Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *ECCV*, 2020.

[2] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. *arXiv preprint arXiv:2103.09460*, 2021.

[3] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. *arXiv preprint arXiv:2101.03697*, 2021.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. *arXiv preprint arXiv:2007.08103*, 2020.

[6] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2021.

[7] 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. *arXiv preprint arXiv:2006.04388*, 2020.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[9] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[10] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. *arXiv preprint arXiv:2003.05664*, 2020.

[11] 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*, pages 9627–9636, 2019.

[12] Shengkai Wu, Xiaoping Li, and Xinggang Wang. Iou-aware single-stage object detector for accurate localization. *Image and Vision Computing*, 97:103911, 2020.

[13] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *arXiv preprint arXiv:1904.04971*, 2019.

[14] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. *arXiv preprint arXiv:1909.02466*, 2019.

[15] Zhaohui Zheng, Rongguang Ye, Ping Wang, Jun Wang, Dongwei Ren, and Wangmeng Zuo. Localization distillation for object detection. *arXiv preprint arXiv:2102.12252*, 2021.

[16] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020.