

# Purified Distillation: Bridging Domain Shift and Category Gap in Incremental Object Detection

Anonymous Authors

## A IMPLEMENTATION DETAILS ON FASTER R-CNN AND DINO

This section describes the details of implementing our PD on Faster R-CNN [6] and DINO [1].

**Faster R-CNN.** Faster R-CNN [6] is a classic two-stage object detector comprising a backbone, RPN, and detection head. Due to its detection head directly predicting the position deviation relative to the Region of Interest (RoI), we modify this localization branch to model the output as a discrete distribution to meet the requirements of PD. On the other hand, RPN is a crucial component of Faster R-CNN, which is responsible for generating proposals most likely to contain objects. Therefore, we also distill RPN by selecting the top 512 proposals with the highest confidence. The implementation is detailed in Fig. 1a.

**DINO.** As a state-of-the-art transformer-based detector, DINO [8] utilizes a set of trainable queries to detect objects in images. Similar to Faster R-CNN, we also modify DINO’s localization head. Moreover, considering the special structure of DINO, we add an additional attention distillation to the encoder, as shown in Fig. 1b.

## B ABLATION FOR THE MULTI-SCALE CROSS ATTENTION DISTILLATION

Here, we perform more detailed ablation experiments on the proposed Multi-scale Cross Attention Distillation (MCAD). We reckon that multi-scale cross attention can effectively integrate information from different scales. To validate this hypothesis, we replace the MCAD with distillation using: (1)  $L_2$  loss [5], (2) Selective Feature Distillation (SFD) strategy in Faster ILOD [4], and (3) the original Cross Attention (CA) [7]. All experimental results are presented in Tab. 1. We can see that simply forcing the student to mimic the teacher’s behavior by  $L_2$  loss significantly interferes with learning old and new tasks. In contrast, SFD only calculates distillation loss with higher activation values in the feature map, thereby alleviating conflicts between old and current tasks to some extent. Although cross-attention further mitigates catastrophic forgetting, our MCAD effectively integrates multiscale information, resulting in a higher mAP.

## C ABLATION FOR THE DISCRIMINATOR

In this section, we analyze the discriminator in feature space distillation. To investigate the impact of the discriminator on PD, we conduct experiments with different discriminators, as presented in Tab. 2. We compare our discriminator to PixelDiscriminator, NLayerDiscriminator, and a discriminator composed of fully connected layers. The PixelDiscriminator and NLayerDiscriminator, proposed in [9], are widely employed in generative adversarial networks and domain adaptive object detection, and the experimental results demonstrate their effectiveness in reducing feature differences. The PixelDiscriminator predicts each pixel using  $1 \times 1$  convolutions, thus it overlooks

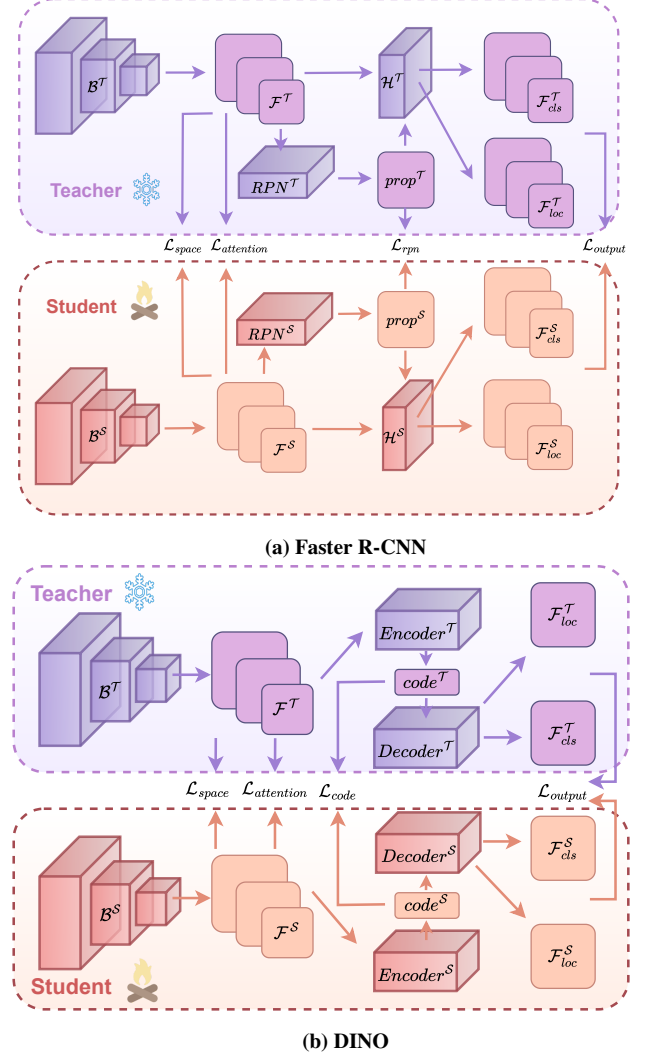
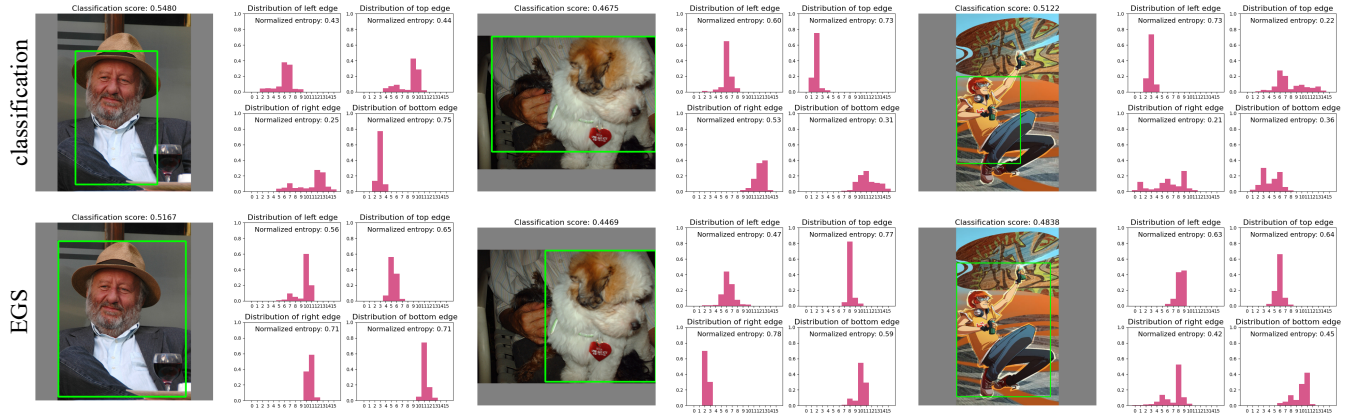


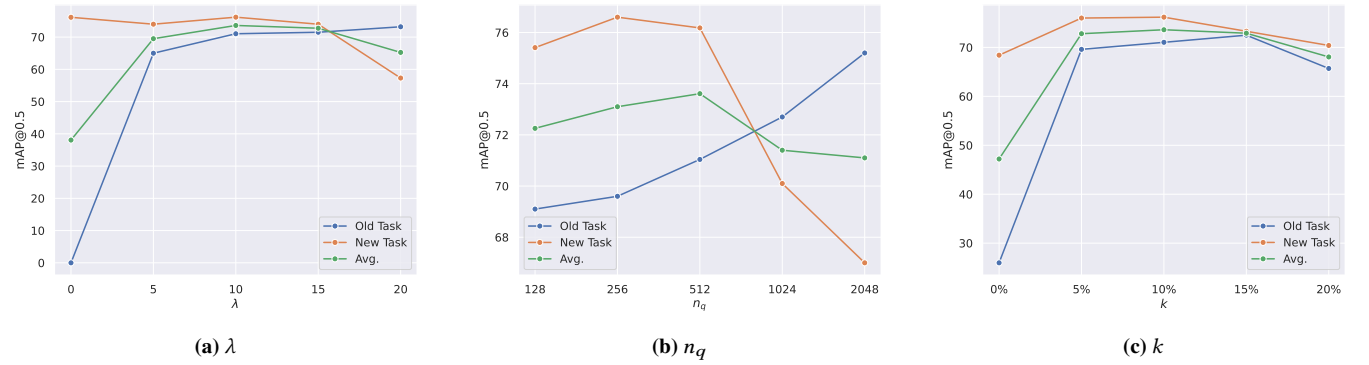
Figure 1: PD implementation details on (a) Faster R-CNN and (b) DINO.

Table 1: Ablation results for the multi-scale cross attention distillation. Here, SFD stands for selective feature distillation strategy in Faster ILOD [4], and CA is the original cross attention [7].

Distillation	Kitchen → KITTI			Comic → Parasites		
	Old	New	Avg.	Old	New	Avg.
$L_2$	73.92 (-20.81)	45.30	59.61	47.79 (- 8.65)	30.66	39.22
SFD	87.22 (- 7.51)	69.68	78.45	49.65 (- 6.79)	80.21	64.93
CA	87.95 (- 6.78)	71.37	79.66	51.03 (- 5.41)	79.58	65.30
MCAD	89.70 (- 5.03)	71.22	<b>80.46</b>	52.39 (- 4.05)	81.07	<b>66.73</b>



**Figure 2: Comparison of different output nodes selection strategy. The first row represents the bounding box with the highest classification in each image, while the second row represents those with the highest product of classification and normalized entropy of the location distribution in each image.**



**Figure 3: Experiments on the influence of different hyperparameters. Notice that we average the detection performance under scenarios (Kitchen  $\rightarrow$  KITTI) and (Comic  $\rightarrow$  Parasites) for better visualization.**

**Table 2: Incremental learning performance with different discriminators. The PixelDiscriminator consists of three  $1 \times 1$  convolutional layers, while the NLayerDiscriminator is composed of several  $4 \times 4$  convolutional layers with different strides. The FCDiscriminator is designed with three fully connected layers, and the dimension of the intermediate hidden layer is set to 512.**

Discriminators	Parameters	Kitchen $\rightarrow$ KITTI			Comic $\rightarrow$ Parasites		
		Old	New	Avg.	Old	New	Avg.
PixelDiscriminator	75,072	86.26 (- 8.47)	70.50	78.38	48.98 (- 7.46)	80.22	64.60
NLayerDiscriminator	8,166,541	88.01 (- 6.72)	70.83	79.42	50.58 (- 5.86)	81.15	65.86
FCDiscriminator	630,961,079	77.75 (-16.98)	65.29	71.52	40.42 (-16.02)	78.90	59.66
Ours	2,659,330	89.61 (- 5.12)	71.05	<b>80.33</b>	52.39 (- 4.05)	81.07	<b>66.73</b>

the spatial information of features. Also, the FCDiscriminator fails to capture the spatial structure and local patterns of features, significantly disrupting knowledge transfer from the teacher to the student. The NLayerDiscriminator, with its more complex structure and larger receptive field, better maintains the memory of old tasks but comes with higher computational costs. Compared with these

discriminators, our discriminator integrates multi-level features to obtain comprehensive information, thus achieving better incremental learning performance at a relatively lower computational cost.

## D ANALYSIS FOR OUTPUT NODES SELECTION STRATEGIES

In this section, we assess the performance of our selection strategy by comparing it to the classification-based strategy. Incremental Object Detection (IOD) methods mostly utilize output distillation, aiming to transfer knowledge from the teacher’s output to the student. As the classification scores and location offsets in the output contain numerous negative samples lacking meaningful information, a criterion must be defined to filter output nodes. Faster ILOD [4] selects the top- $k$  boxes based on the maximum classification score of bounding boxes to selectively calculate distillation loss. ERD [2] considers the mean and variation to set a threshold to choose output nodes. However, none of these methods consider both the classification and location of bounding boxes simultaneously, which is crucial because inaccurate boxes can interfere with the student’s learning, especially



**Figure 4: Qualitative results in the multi-step incremental object detection. We provide the predictions of different methods after incremental learning on the dataset sequence (VOC → KITTI → Watercolor → Comic → Kitchen → Parasites).**

in incremental object detection with both domain shift and category differences.

To address this issue, we propose Entropy-Guided Selection (EGS), which utilizes the entropy of the distribution of bounding box positions to measure the model’s localization confidence. To unify with the classification score, we normalize the entropy to  $[0, 1]$  and then select the top- $k$  nodes with the highest confidence based on the element-wise product of the classification score and the normalized entropy. In Fig. 2, we visualize the highest-scoring bounding boxes selected by different strategies. Nodes selected solely based on classification exhibit high classification confidence but low position entropy, and fail to denote objects correctly. While EGS balances both classification and localization, resulting in nodes that can more accurately identify objects, with a minor loss in classification confidence.

## E SENSITIVE ANALYSIS FOR HYPERPARAMETERS

In this section, we conduct sensitivity analysis on  $\lambda$ ,  $n_q$ , and  $k$  in two typical incremental object detection scenarios, (Kitchen → KITTI)

and (Comic → Parasites). The trends of mAP changes with different hyperparameters on both old and new tasks are illustrated in Fig. 3.

$\lambda$ . This parameter balances the learning of new and old knowledge, and Fig. 3a presents the experimental results of different values. The larger  $\lambda$ , the model pays more attention to learning the current task, and vice versa.

$n_q$ . The number of queries,  $n_q$ , is a critical parameter to maintain the previous knowledge. As shown in Fig. 3b, a smaller  $n_q$  prevents the student from comprehensively retaining old task-related knowledge, while a larger one can hinder learning new tasks.

$k$ . On the other hand, the results of different  $k$  are reported in Fig. 3c. The impact of  $k$  is similar to  $n_q$ , but very large values of  $k$  lead to significant conflicts in learning new and old tasks, impairing performance in both tasks.

## F MORE QUALITATIVE RESULTS

Here we present qualitative results for multi-step incremental object detection, where the dataset sequence is (VOC → KITTI → Watercolor → Comic → Kitchen → Parasites). Comparative methods include vanilla finetuning, LwF [3], Faster ILOD [4], SID [5], and



ERD [2]. After completing all incremental steps, we validate models on all datasets, and the qualitative results are provided in Fig. 4. It can be observed that even in such a complex scenario, our PD consistently preserves the memory of old tasks, while the comparative methods exhibit more significant forgetting (misidentifying the background as an object or incorrectly classifying objects).

## REFERENCES

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. Springer, 213–229.
- [2] Tao Feng, Mang Wang, and Hangjie Yuan. 2022. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9427–9436.
- [3] Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2017), 2935–2947.
- [4] Can Peng, Kun Zhao, and Brian C Lovell. 2020. Faster ILOD: Incremental learning for object detectors based on faster rcnn. *Pattern Recognition Letters* 140 (2020), 109–115.
- [5] Can Peng, Kun Zhao, Sam Maksoud, Meng Li, and Brian C Lovell. 2021. SID: Incremental learning for anchor-free object detection via Selective and Inter-related Distillation. *Computer Vision and Image Understanding* 210 (2021), 103229.
- [6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [8] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605 [cs.CV]
- [9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.