RESPONSE-BASED DISTILLATION FOR INCREMENTAL OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional object detection are ill-equipped for incremental learning. However, fine-tuning directly on a well-trained detection model with only new data will leads to catastrophic forgetting. Knowledge distillation is a straightforward way to mitigate catastrophic forgetting. In Incremental Object Detection (IOD), previous work mainly focuses on feature-level knowledge distillation, but the different response of detector has not been fully explored yet. In this paper, we propose a fully response-based incremental distillation method focusing on learning response from detection bounding boxes and classification predictions. Firstly, our method transferring category knowledge while equipping student model with the ability to retain localization knowledge during incremental learning. In addition, we further evaluate the qualities of all locations and provides valuable response by adaptive pseudo-label selection (APS) strategies. Finally, we elucidate that knowledge from different responses should be assigned with different importance during incremental distillation. Extensive experiments conducted on MS COCO demonstrate significant advantages of our method, which substantially narrow the performance gap towards full training.

1 INTRODUCTION

In the natural world, the visual system of creatures could constantly acquire, integrate and optimize knowledge. Learning mode is inherently incremental for them. In contrast, currently, the classic training paradigm of the object detection model (Tian et al., 2019; Li et al., 2021b) does not have such capability. Supervised object detection paradigm relies on accessing pre-defined labeled data. This learning paradigm implicit assumes data distribution is fixed or stationary, while data from real world is represented by continuous and dynamic data flow, whose distribution is non-stationary. When the model continuously obtains knowledge from non-stationary data distribution, new knowledge would interfere with the old one, triggering catastrophic forgetting (Goodfellow et al., 2015; Mccloskey & Cohen, 1989).

A straightforward way in incremental object detection is based on knowledge distillation (Hinton et al., 2015). Peng et al. (2021) stressed that the Tower layers could reduce catastrophic forgetting significantly. They implemented incremental learning on an anchor-free detector and selectively performed distillation on non-regression outputs. In knowledge distillation for object detection where incremental learning was not introduced, previous work extracted knowledge from the combined distillation of different components. For example, Chen et al. (2017) and Sun et al. (2020) distilled all components of the detector. Nevertheless, the nature of these methods are designed using feature-based knowledge distillation, fully response-based method (Gou et al., 2021) has not been explored in incremental object detection yet. Besides, since different components in the detection make different contributions to incremental distillation, an elaborate design for different responses is essential.

This paper focused on a practical and challenging problem concerning incremental object detection: *how to learn response from detecting bounding boxes and classification predictions*. Responses in object detection contain logits together with the offset of bounding box (Gou et al., 2021). Firstly, since the number of ground truth on each new image is uncertain, one of the foremost considerations is that validate the object of all samples, determining which object is positive or negative and which ground truth each object should regress towards. A troublesome issue is that the output of the regression branch may be substantially different from that of the ground truth. Furthermore, the localization knowledge of each edge in the detection bounding boxes is also response that should be taken seriously. To sum up, we use the response on the location where teacher detector generates high-quality predictions as the ground truth to guide the student detector following the behavior of teacher on the old object. In this case, it is of great significance to use the old detector to provide valuable incremental information from detection bounding boxes and classification predictions.

To tackle the above problems, this paper rethinks response-based knowledge distillation method, finding that distillation at proper locations is crucial in facilitating incremental object detection. We believe that student detector can acquire high-quality knowledge from the teacher detector's high-quality predictions. Driven by this inspiration, we proposed an incremental distillation scheme that learns specific responses from the classification head and regression head respectively. Unlike previous work, we introduce incremental localization distillation (Zheng et al., 2021) in regression response to equip student detector with the ability to learn location ambiguity during incremental learning. Besides, we propose adaptive pseudo-label selection (APS) strategies to automatically select distillation nodes based on statistical characteristics from different responses, which evaluates the qualities of all locations and provides valuable response. We alleviate catastrophic forgetting greatly and significantly narrow the gap with full training by distilling the response alone. Extensive experiments on the MS COCO dataset support our analysis and conclusion.

The main contributions of this work can be summarized,

1. To the best of our knowledge, this paper is first work to explore the fully response-based distillation method in incremental object detection.
2. We propose a novel distillation scheme elaborate designed for incremental detection focusing on detection bounding boxes and classification predictions.
3. We propose adaptive pseudo-label selection strategies to automatically select distillation nodes based on statistical characteristics from the different responses.

2 Related work

Incremental Learning. Catastrophic forgetting is the core challenge for incremental learning. Incremental learning based on parameter constraints is a candidate solution for such problem, which protects the old knowledge by introducing an additional parameter-related regularization term to modify the gradient. EWC (Kirkpatrick et al., 2016) and MAS (Aljundi et al., 2018) are two typical representatives of such method. Another solution is incremental learning based on knowledge distillation, as well as the topic of the study. This kind of method mainly projects old knowledge by transferring knowledge in old tasks to new tasks through knowledge distillation. LwF (Li & Hoiem, 2018) is the first algorithm that introduces the concept of knowledge distillation into incremental learning, in the purpose of making predictions of the new model on new tasks similar to that of the old model and thereby protecting the old knowledge in the form of knowledge transfer. However, it would cause knowledge confusion when the correlation between new and old tasks is low. iCaRL (Rebuffi et al., 2017) algorithm uses knowledge distillation to avoid excessive deterioration of knowledge in the network, while BiC (Wu et al., 2019) algorithm added a bias correction layer after the FC layer to offset the category bias of new data when using the distillation loss.

Incremental Object Detection. Compared with incremental classification, achievements on incremental object detection is much less. Meanwhile, the high complexity of the detection task also adds the difficulty of incremental object detection. Shmelkov et al. (2017) proposed to apply LwF to Fast RCNN detector (Girshick, 2015), which is the first work on incremental object detection. Thereafter, some researchers move this area forward. Peng et al. (2021) proposed SID approach for incremental object detection on anchor-free detector and conducted experiments on FCOS (Tian et al., 2019) and CenterNet (Zhou et al., 2019). Li et al. (2021a) studied object detection based on class-incremental learning on Faster RCNN detector with emphasis given to few-shot scenarios, which is also the focus of ONCE algorithm (Perez-Rua et al., 2020). Li et al. (2019) designed an incremental object detection system with RetinaNet detector (Lin et al., 2020) under the scenario of edge device. the latest work, Joseph et al. (2021) introduced the concept of incremental learning when defining the problems of Open World Object Detection (OWOD).

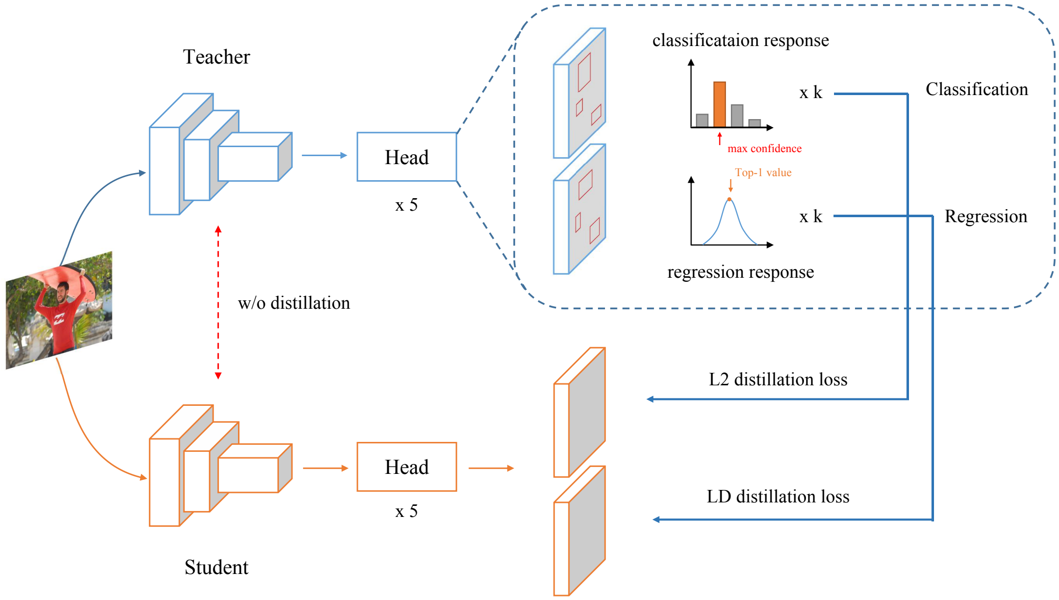


Figure 1: Overview of response-based incremental distillation.

Knowledge Distillation for Object Detection. Knowledge distillation (Bucila et al., 2006)is an effective way to transfer knowledge between models. Widely applied in image classification tasks in previous researches, knowledge distillation is now used in object detection tasks more and more frequently. Chen et al. (2017) implemented distillation in all components on Faster RCNN detector (including backbone, proposals in RPN, and head). To imitate the high-level feature response of the teacher model with the student model, Wang et al. (2019) proposed a distillation method based on fine-grained feature imitation. By synthesize category-conditioned objects through inverse mapping, Chawla et al. (2021) proposed a data-free knowledge distillation technology applicable for object detection, but the method would trigger dream-image. Guo et al. (2021) believing that foreground and background both play an unique role in object detection, proposed an object detection distillation method that could decouple foreground and background. Zheng et al. (2021) proposed a localization distillation method introducing knowledge distillation into the regression branch of the object detector, so as to enable the student network to solve the localization ambiguity in object detection as the teacher network. However, existing object detection distillation framework does not pay enough attention to the significant role of the head. In this study, we found the head has its particularly significant.

3 METHOD

3.1 OVERALL STRUCTURE

In general, a one-stage object detector is composed of three components: (i.) backbone for feature extraction; (ii.) neck for fusion of multi-level features; (iii.) head for classification and regression. The purpose of incremental distillation is to transfer old knowledge to the student detector, and this knowledge could be the features of the intermediate layer in the backbone or neck or the soft predictions in the head. Here, we incrementally learn a strong and efficient student object detector by the distillation of incremental knowledge from responses of the different heads. The overall incremental detection framework is shown in Figure 1. Firstly, knowledge distillation is applied to learn incremental response from the classification head and regression head of the teacher detector. Secondly, incremental localization distillation loss is also applied to enhance the localization information extraction ability of the student detector. Notably, the adaptive pseudo-label selection strategies are proposed to gain more meaningful incremental responses from the teacher detector, that is, selective calculation of the distillation loss from the pseudo label provided by the teacher detector. The overall learning target of the student detector is therefore defined as,

$$\mathcal{L}_{total} = \mathcal{L}_{model} + \lambda_1 \mathcal{L}_{dist\_cls}(\mathcal{C}_{\mathcal{S}}, \mathcal{C}_{\mathcal{T}}) + \lambda_2 \mathcal{L}_{dist\_bbox}(\mathcal{R}_{\mathcal{S}}, \mathcal{R}_{\mathcal{T}}) \tag{1}$$

where λ is the parameters that balances the weights of different loss terms. The loss term $\mathcal{L}_{model}$ is standard loss function used in GFocal (Li et al., 2020) to train object detector for the new object class. The second loss term $\mathcal{L}_{dist\_cls}$ is the L2 incremental distillation loss for classification branch. The third loss term $\mathcal{L}_{dist\_bbox}$ is the incremental localization distillation loss for regression branch. In the above, we set $\lambda_1 = \lambda_2 = 1$.

3.2 Distillation at Classification-based Response

The soft predictions from the classification head contains the knowledge of various categories discovered by the teacher model. Through the learning of soft prediction, the student model can inherit hidden knowledge, which is intuitive for classification tasks. Let $\mathcal{T}$ be the teacher model, we use SoftMax function to transform logits $\mathcal{Z}_{\mathcal{T}}$ in final score output, responding probability distribution $\mathcal{P}_{\mathcal{T}}$ is defined as,

$$\mathcal{P}_{\mathcal{T}} = \mathrm{SoftMax}\left(\frac{\mathcal{Z}_{\mathcal{T}}}{t}\right) \tag{2}$$

Similarly, we define $\mathcal{P}_{\mathcal{S}}$ for the student model $\mathcal{S}$,

$$\mathcal{P}_{\mathcal{S}} = \mathrm{SoftMax}\left(\frac{\mathcal{Z}_{\mathcal{S}}}{t}\right) \tag{3}$$

where t is temperature to soften the probability distribution for $\mathcal{P}_{\mathcal{T}}$ and $\mathcal{P}_{\mathcal{S}}$.

Previous works usually directly use all the prediction responses in the classification head and treat each position equally. If there is any inappropriate balance, the response generated by the background category may overwhelm the response generated by the foreground category, thereby interfering with the retention of old knowledge. To tackle this problem, the L2 incremental distillation loss for the classification-based response is as follows,

$$\mathcal{L}_{dist\_cls}\left(\mathcal{C}_{\mathcal{S}}, \mathcal{C}_{\mathcal{T}}\right) = \sum_{i=1}^{m} \left({\mathcal{P}_{\mathcal{T}}}^i - {\mathcal{P}_{\mathcal{S}}}^i\right)^2 \tag{4}$$

where ${\mathcal{P}_{\mathcal{T}}}^i$ is the category response of the frozen teacher detector from m pseudo object classes using the new data, and ${\mathcal{P}_{\mathcal{S}}}^i$ is the category response of the student detector for the old object classes. By distilling the selected response, the student detector inherits the knowledge of the positive object category to a greater extent.

3.3 Distillation at Regression-based Response

The bounding box response from the regression branch is also quite important for incremental detection. Contrary to the discrete class information, there is a possibility that the output of the regression branch may provide a regression direction that contradicts the ground truth. That's because, even if the image does not contain any objects of the old category, the regression branch will still predict the bounding box, although the confidence is relatively low. That poses a challenge for learning the knowledge of the old model to correctly predict the bounding box of the old object. On the other hand, in previous works, only the bounding box of a relatively high-confidence object was learned as the knowledge of the teacher detector, ignoring the localization information.

Benefit from the general distribution of bounding box $\mathcal{B}$ from GFocal detector, each edge of $\mathcal{B}$ can be represented by probability distribution through SoftMax function (Zheng et al., 2021). Further, the probability matrix of bounding box $\mathcal{B}$ is defined as,

$$\mathcal{B} = [p_t, p_b, p_l, p_r] \, \varepsilon \mathbb{R}^{n \times 4} \tag{5}$$

Therefore, we can extract the incremental localization knowledge of bounding box $\mathcal{B}$ from teacher detector $\mathcal{T}$ and transfer it to student detector $\mathcal{S}$ by using KL-Divergence loss,

$$\mathcal{L}_{LD}^{e} = \mathcal{L}_{KL}\left(\mathcal{P_S}^{j}, \mathcal{P_T}^{j}\right) \tag{6}$$

Finally, incremental localization distillation loss for the regression-based response is defined as,

$$\mathcal{L}_{dist\_bbox}\left(\mathcal{R_S}, \mathcal{R_T}\right) = \sum_{j=1}^{J} \sum_{e \in \mathcal{B}} \mathcal{L}_{LD}^{e} \tag{7}$$

where $\mathcal{R_T}^{j}$ is the regression response of the frozen teacher detector from J pseudo bounding box using the new object, and $\mathcal{R_S}^{j}$ is the regression response of the student detector for the old bounding box. Compared to only use the bounding box in previous works, incremental localization distillation can provide extra localization response.

3.4 Adaptive Pseudo-label Selection

When an incremental object detector is trained, the gap of knowledge between the teacher detector and the student detector is obvious. For a new sample, it's preferable for the teacher detector to provide the high-quality knowledge, as the student detector will benefit from positive response. To this end, a basic problem related to incremental object detection has been thoroughly studied: *how to select distillation nodes as positive response.* Traditional selection strategies depend on sensitive hyper-parameters such as setting confidence and Top-K . Those empirical practices in which rules are fixed have such consequences that too small thresholds lead to the ignoring of some objects while too large ones probably result in the introduction of negative response.

To solve this problem, the adaptive pseudo-label selection (APS) strategy is proposed. Algorithm 1 describes how the proposed strategy works for an input image. We obtain positive response from the category and bounding box as distillation nodes respectively.

Classification head. The statistical characteristics of the category information are utilized to determine the response of classification, as described in L-2 to L-12. We first calculate the classification confidence of each position. After that, we calculate the mean μ_C and standard deviation σ_C in L-6 and L-7. With these statistical, the threshold τ_C is obtained in L-8. Finally, we select these candidates whose confidence are greater than the threshold τ_C in L-9 to L-12.

Regression head. The statistical characteristics of the distribution information are utilized to determine the response of regression, as described in L-14 to L-23. For the GFocal detector, the author points out that a certain and unambiguous bounding box, whose distribution is usually sharp. Therefore, the Top-1 value is usually very large if the distribution is sharp. Based on these statistical characteristics, the top-1 is used to measure the confidence of the bounding box. We first calculate the Top-1 of each distribution. After that, we calculate the mean μ_B and the standard deviation σ_B of all Top-1 in L-17 and L-18. Then, the threshold τ_B is obtained in L-19. Finally, we select these candidates whose confidence are greater than the threshold τ_B in L-20 to L-23.

The proposed APS strategy has the following advantages: 1. guaranteeing fair selection of pseudo labels of different objects. 2. using statistical characteristics of different branches to adaptively select pseudo labels to provide the incremental response.

4 Experiments and Discussion

In this section, we perform experiments on several incremental scenarios on the MS COCO dataset using baseline detector GFocal to validate our method. Then, we perform ablation studies to prove the effectiveness of each component of our method. Finally, we discuss a question: What are the bottlenecks in our method?

Algorithm 1 Adaptive Pseudo-label Selection (APS)

Input: Unlabeled image $\boldsymbol{I}$, image-level labels $\boldsymbol{c}$, $\boldsymbol{b}$, teacher detector $\boldsymbol{\theta'}$
Output: Sampled pseudo-label sets $\boldsymbol{C'}$, $\boldsymbol{B'}$

```
Inference I with θ' yields the classification score C and predicted distribution B

Classification branch:
for k = 1 to C do
    G_C ⟵ confidence(C_k)
Compute μ_C = mean(G_C)
Compute σ_C = std(G_C)
Compute threshold τ_C = μ_C + σ_C
for each candidate c in C do
    if G_{C_k} ≥ τ_C then
        Add candidate c to C'
return C'

Regression branch:
for k = 1 to B do
    G_B ⟵ Max(B_k)
Compute μ_B = mean(G_B)
Compute σ_B = std(G_B)
Compute threshold τ_B = μ_B + σ_B
for each candidate b in B do
    if G_{S_b} ≥ τ_B then
        Add candidate b to B'
return B'
```

Implementation Details. We build our method on top of the GFocal detector using their public implementations. The teacher and student detectors defined in our experiments are standard GFocal architectures. For the GFocal detector, ResNet-50 is used as its backbone, FPN (Lin et al., 2017) is used as its neck. We trained our detector to follow the same parameters described in their paper. All the experiments are performed on 8 NVIDIA Tesla V100 GPU, with batch size of 8.

Datasets and Evaluation Metric. MS COCO 2017 (Chen et al., 2015) is a challenging benchmark in object detection which contains 80 object classes. For experiments on the COCO dataset, we use train and validation set for training and test set for testing. The standard COCO protocols are used as an evaluation metric, i.e. AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L.

Experiment Setup for MS COCO. The detector is trained by 12 epochs (1x mode) for each incremental step for the MS COCO dataset. The setting is consistent for all the detectors in the different scenarios. We set up experiments in the following scenarios:

- **40 + 40:** we train a base detector with the first 40 classes and then the last 40 classes are learned incrementally as new object classes.
- **75 + 5:** we train a base detector with the first 75 classes and then the last 5 classes are learned incrementally as new object classes.
- **Last 40 + First 40:** we specially train a base detector with the last 40 classes and then the first 40 classes are learned incrementally as new object classes.

4.1 Overall Performance

We reported the incremental results under the first 40 classes + last 40 classes scenario in Table 1. In this scenario, we observed that if the old detector and new data were directly used to conduct fine-tuning process, then the AP dropped to 17.8% as compared to the 40.2% in full data training. This is because the fine-tuning made the detector's memory of old object classes close to 0, resulting in catastrophic forgetting (ref to Figure 2(b)). Our method far outperformed fine-tuning across various IoUs evaluation criteria from 0.5 to 0.95. The experimental results show that when IoU is 0.5, 0.75

Table 1: Incremental results based on GFocal detector on COCO benchmark under first 40 classes + last 40 classes. (”Δ” represents an improvement over Catastrophic Forgetting. ”∇” represents the gap with Upper Bound.)

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Upper Bound	40.2	58.3	43.6	23.2	44.1	52.2
Catastrophic Forgetting	17.8	25.9	19.3	8.3	19.2	24.6
LwF (Li & Hoiem, 2018)	17.2($\Delta-0.6/\nabla 23.0$)	25.4	18.6	7.9	18.4	24.3
RILOD (Li et al., 2019)	29.9($\Delta 12.1/\nabla 10.3$)	45.0	32.0	15.8	33.0	40.5
SID (Peng et al., 2021)	34.0($\Delta 16.2/\nabla 6.2$)	51.4	36.3	18.4	38.4	44.9
Ours	**36.9**($\Delta 19.1/\nabla 3.2$)	**54.5**	**39.6**	**21.3**	**40.4**	**47.5**

Table 2: Incremental results based on GFocal detector on COCO benchmark under last 40 classes + first 40 classes. (”Δ” represents an improvement over Catastrophic Forgetting. ”∇” represents the gap with Upper Bound.)

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Upper Bound	40.2	58.3	43.6	23.2	44.1	52.2
Catastrophic Forgetting	22.6	32.7	24.2	15.1	25.0	27.6
LwF (Li & Hoiem, 2018)	20.5($\Delta-2.1/\nabla 19.7$)	29.9	22.1	13.0	22.5	25.3
RILOD (Li et al., 2019)	34.1($\Delta 11.5/\nabla 6.1$)	51.1	36.8	19.1	38.0	43.9
SID (Peng et al., 2021)	33.5 ($\Delta 10.9/\nabla 6.7$)	50.9	36.3	19.0	37.7	43.0
Ours	**37.5**($\Delta 14.9/\nabla 2.8$)	**55.1**	**40.4**	**21.3**	**41.1**	**48.2**

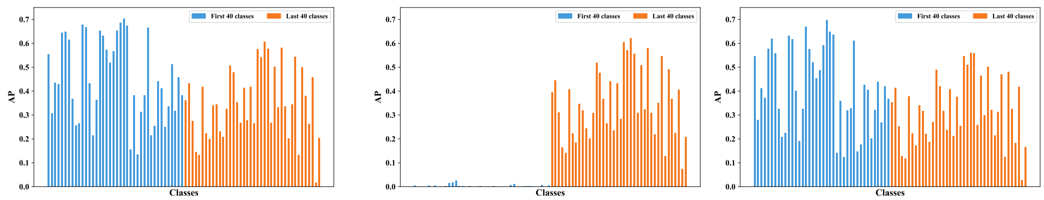
and 0.95, the AP improves by 19.1%, 28.6% and 20.3%, respectively. This indicates that our method can well address catastrophic forgetting. Notably, even compared with the full data training where the entire dataset was used, our method only had a gap of 3.2%. This indicates that the student detector maintained a good memory of the old objects while learning new objects. To put it more intuitively, we visualized the incremental results of all object classes, as shown in Figure 2. The blue column denotes the AP of the first 40 classes, while the orange column denotes the AP of the last 40 classes. As can be seen, our method has produced significant outcomes. In Figure 3, we further visualized the AP of all objects of the first 40 classes and the last 40 classes.

Considering the long-tail problem of the COCO dataset, we particularly configured an incremental experiment under the last 40 classes + first 40 classes scenario. In this scenario, the first 40 classes object contain more memories that should be retained, which means that more incremental responses can be obtained. As can be seen from Table 2, the incremental performance of our method has been further improved, with the gap against full data training reduced to 2.8% and the improvement on catastrophic forgetting increased to 14.9%. This also validates our inference that the method we propose benefits from more incremental responses.

In addition, we also compared our method with LwF, RILOD, and SID. Both Table 1 and Table 2 show that although LwF works well in incremental classification, it is even lower AP than directly fine-tuning in detection tasks. To a fair comparison with RILOD and SID, we replicated them based on GFocal detector. For RILOD, we completely followed their method. For SID, we used the component with the greatest improvement proposed by the authors. Both tables show that the improvement of our method to catastrophic forgetting is outstanding.

4.2 Ablation Study

As shown in Table 3, we validated the effectiveness of different components of the proposed method on the COCO benchmark to highlight our improvement in performance. “all cls + all reg” denotes that responses from both the classification branch and regression branch are treated equally in the incremental distillation, which is also our baseline performance. “all cls” denotes that only classification responses in the incremental distillation process are treated equally. “all reg” denotes that only regression responses in the incremental distillation process are treated equally. “cls + APS”



(a) Upper Bound (b) Catastrophic Forgetting (c) Response-based Distillation

Figure 2: AP of Per-class among different learning schemes. (a) Detector is trained with all data.(b) Student detector is finetuned with new classes.(c) Student detector is distilled via different response.

denotes that the APS strategy is employed to conduct incremental distillation over classification responses, as shown in Equation 4. “cls + reg +APS” denotes that responses based on regression are also used, as shown in Equation 7. In Table 3, separately distillation all responses from classification and regression, obtained 23.8% and 13.0% of AP. When only all responses from the regression branch are used, AP is even lower than the fine-tuning performance, which also supports our assumption stated in the introduction section. Comparatively, the direct incremental distillation of all responses from classification and regression branches obtains 31.5% of AP. By utilizing APS to decouple classification responses, the student detector obtained higher results. Our decoupling proposal can improve the result from 31.5% of AP to 33.2%. The incremental distillation process further utilized the APS strategy to decouple regression responses, obtaining 36.9% of AP on the COCO benchmark, a 5.4% improvement compared with the baseline performance. All these results clearly point to the advantageous performance of our method.

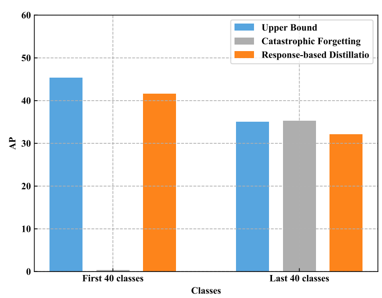


Figure 3: First 40 classes vs. Last 40 classes.

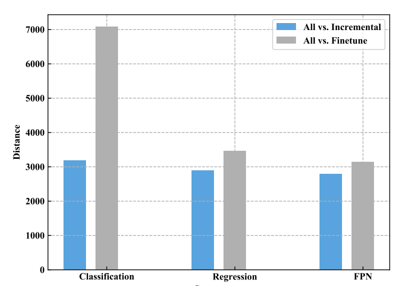


Figure 4: L2 distance analysis.

4.3 Discussion

In this section, we present further insights into response-based incremental distillation. We reveal the contribution of different components for distillation detection and discuss the impact of incremental response in the head.

Distance between different components. We calculate the feature distance between different components to illustrate why response-based distillation can attain higher performance compared to other components. We randomly choose 10 images from COCO minival and calculate the L2 distance of features in different components of different training strategies. As shown in Figure 4, “All” denotes that the detector with full data training; ‘Incremental’ denotes that the detector with incremental data training; “Finetune” denotes that the detector with finetuning training. Distilling student detector via classification-based and regression-based incremental response in the head can substantially narrow the distances with upper bound. However, neither the L2 distance between “All vs. Incremental” and “All vs. Finetune” improves significantly in the FPN representing the feature-based distillation. This also supports our assumption that different response from the head has its particularly significant, especially classification response.

Incremental response helps both learning and generalization. We notice that the incremental response from the head can provide an effective guidance to avoid catastrophic forgetting problems.

Table 3: Ablation study based on GFocal detector using the COCO benchmark under first 40 classes + last 40 classes. ("Δ" represents an improvement over Catastrophic Forgetting. "∇" represents the gap with Upper Bound.)

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Upper Bound	40.2	58.3	43.6	23.2	44.1	52.2
Catastrophic Forgetting	17.8	25.9	19.3	8.3	19.2	24.6
KD:all cls + all reg	31.5($\Delta 13.7/\nabla 8.7$)	48.3	33.4	17.7	35.3	41.3
KD:all cls	23.8($\Delta 10.1/\nabla 16.4$)	36.6	24.9	11.8	27.2	32.9
KD:all reg	13.0($\Delta - 4.8/\nabla 27.2$)	21.1	13.4	5.0	14.7	18.6
KD:cls + APS	33.2($\Delta 15.4/\nabla 7.0$)	51.2	35.2	18.5	37.8	43.8
KD:cls + reg + APS	**36.9**($\Delta 19.1/\nabla 3.2$)	**54.5**	**39.6**	**21.3**	**40.4**	**47.5**

Table 4: Incremental results based on GFocal detector on COCO benchmark under first 75 classes + last 5 classes. ("Δ" represents an improvement over Catastrophic Forgetting.)

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Catastrophic Forgetting	3.8	5.9	3.8	1.9	5.3	6.5
All response	32.5 ($\Delta 28.7$)	48.9	34.7	18.3	35.9	41.1
Adaptive response	28.3 ($\Delta 24.5$)	42.4	30.3	15.0	31.5	37.0

Thereby, the student detector achieves noticeable improvement in different scenarios. In Table 4, our method can still learn new object classes without forgetting old ones even with a little data. But, due to the insufficient incremental response provided in the +5 classes scenario, our method did not achieve a more competitive AP. However, our method still contributes to generalization. In this case, we can degrade the adaptive response to all responses in exchange for a better compromise. Comparatively, when sufficient incremental responses emerge, our method is easy to achieve (near) perfect AP.

5 CONCLUSION

In this paper, we design an entirely response-based incremental object detection paradigm. This method uses only the detection head to achieve incremental detection, which significantly alleviates catastrophic forgetting. We innovatively learn responses from detection bounding boxes and classification predictions, and specifically introduce incremental localization distillation in the regression response. Second, the adaptive selection technique is designed to provide a fair incremental response in the different heads. Extensive experiments validate the effectiveness of our method. Finally, our empirical analysis reveals the contribution of different responses and components in incremental detection, which could provide insights to further advancement in the field.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, pp. 144–161. Springer, 2018.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos (eds.), *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pp. 535–541. ACM, 2006.

Akshay Chawla, Hongxu Yin, Pavlo Molchanov, and Jose M. Alvarez. Data-free knowledge distillation for object detection. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pp. 3288–3297. IEEE, 2021.

Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 742–751, 2017.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

Ross Girshick. Fast r-cnn, 2015.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819, 2021.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 2154–2164. Computer Vision Foundation / IEEE, 2021.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

K. J. Joseph, Salman H. Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Towards open world object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 5830–5840. Computer Vision Foundation / IEEE, 2021.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry P. Heck. RILOD: near real-time incremental learning for object detection at the edge. In Songqing Chen, Ryokichi Onishi, Ganesh Ananthanarayanan, and Qun Li (eds.), *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, SEC 2019, Arlington, Virginia, USA, November 7-9, 2019*, pp. 113–126. ACM, 2019.

Pengyang Li, Yanan Li, and Donghui Wang. Class-incremental few-shot object detection, 2021a.

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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 11632–11641. Computer Vision Foundation / IEEE, 2021b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 936–944. IEEE Computer Society, 2017.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020.

M. Mccloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Can Peng, Kun Zhao, Sam Maksoud, Meng Li, and Brian C. Lovell. SID: incremental learning for anchor-free object detection via selective and inter-related distillation. *Comput. Vis. Image Underst.*, 210:103229, 2021.

Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy Hospedales, and Tao Xiang. Incremental few-shot object detection, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5533–5542. IEEE Computer Society, 2017.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 3420–3429. IEEE Computer Society, 2017.

Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *CoRR*, abs/2006.13108, 2020.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9626–9635. IEEE, 2019.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4933–4942. Computer Vision Foundation / IEEE, 2019.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 374–382. Computer Vision Foundation / IEEE, 2019.

Zhaohui Zheng, Rongguang Ye, Ping Wang, Jun Wang, Dongwei Ren, and Wangmeng Zuo. Localization distillation for object detection, 2021.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019.