Overcoming Catastrophic Forgetting in Incremental Object Detection via Elastic Response Distillation

Tao Feng[1], Mang Wang[1*], Hangjie Yuan[2]
[1]Alibaba Group [2]Zhejiang University
fengtao.hi@gmail.com, wangmang.wm@alibaba-inc.com, hj.yuan@zju.edu.cn

Abstract

Traditional object detectors are ill-equipped for incremental learning. However, fine-tuning directly on a well-trained detection model with only new data will lead to catastrophic forgetting. Knowledge distillation is a flexible way to mitigate catastrophic forgetting. In Incremental Object Detection (IOD), previous work mainly focuses on distilling for the combination of features and responses. However, they under-explore the information that contains in responses. In this paper, we propose a response-based incremental distillation method, dubbed Elastic Response Distillation (ERD), which focuses on elastically learning responses from the classification head and the regression head. Firstly, our method transfers category knowledge while equipping student detector with the ability to retain localization information during incremental learning. In addition, we further evaluate the quality of all locations and provide valuable responses by the Elastic Response Selection (ERS) strategy. Finally, we elucidate that the knowledge from different responses should be assigned with different importance during incremental distillation. Extensive experiments conducted on MS COCO demonstrate our method achieves state-of-the-art result, which substantially narrows the performance gap towards full training. Code is available at https://github.com/Hi-FT/ERD.

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 object detection models [19,33] does not have such capability. Supervised object detection paradigm relies on accessing pre-defined labeled data. This learning paradigm implicitly assumes data distribution is fixed or stationary [9, 37], 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 distribution, new knowledge would interfere with the old one, triggering catastrophic forgetting [11, 26]. Based on whether the task identity is provided or must be inferred [34], researchers divide Incremental Learning (IL) into three types: task/domain/class IL. In this paper, we focus on the most intractable scenario for object detection: class incremental object detection.

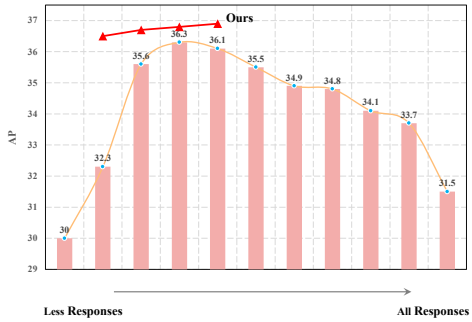


Figure 1. The effect of various responses for IOD.

A flexible way to solve IOD is knowledge distillation [14]. [28] 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. Meanwhile, 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, [5] and [32] distilled all components of the detector. Nevertheless, the nature of these methods are designed using feature-based knowledge distillation [6], response-based method [12] has not been explored in IOD [25] yet. Besides, the advantage of response-based method is that it provides the reasoning information [14, 27] of the teacher detector. Therefore, an elaborate design for different responses is essential [23].

This paper focuses on a practical and challenging problem concerning IOD: *how to learn response from classification predictions and bounding boxes*. Responses in object

*Corresponding author.

detection contain logits together with the offset of bounding box [12]. Firstly, since the number of ground truth on each new image is uncertain, one of the foremost considerations is to validate the response of all samples, determining which response is positive or negative and which response each object should regress towards. Furthermore, as shown in Figure 1, we find that not all responses are important to prevent catastrophic forgetting, thus an appropriate number of response nodes is ideal. [16] also proposed that synaptic consolidation achieves continuous learning by reducing synaptic plasticity critical to previous learning tasks. To sum up, we guide the student detector following the behavior of teacher on the old objects by constraining important responses to stay close to their old values.

To tackle the above problems, this paper rethinks response-based knowledge distillation method, finding that distillation at proper locations is crucial for facilitating IOD. Driven by this inspiration, we proposed an **E**lastic **R**esponse **D**istillation (ERD) scheme that elastically learns responses from classification head and regression head respectively. Unlike previous work, we introduce incremental localization distillation [38] in regression response to equip student detector with the ability to learn location ambiguity [20] during incremental learning. Besides, we propose **E**lastic **R**esponse **S**election (ERS) strategy to automatically select distillation nodes based on statistical characteristics from different responses, which evaluates the qualities of all locations and provides valuable responses. In this paper, we explain how we implement the constraint, and finally how we determine which responses are important. We greatly alleviate catastrophic forgetting problem and significantly narrow the gap with full training. Extensive experiments on the MS COCO dataset support our analysis and conclusion.

The our contributions can be summarized as follows,

(i) To the best of our knowledge, this paper is the first work to explore the response-based distillation method in IOD and dissect the essential differences between feature-based and response-based solutions for IOD. **(ii)** We propose ERD based on statistical analysis, which separately distills selective classification and regression responses using the proposed ERS strategy. **(iii)** Extensive experiments on MS COCO demonstrate that the proposed method achieves state-of-the-art performance and can be easily extended to different detectors.

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 [16] and MAS [1] are two typical representatives of such method. Another solution is incremental learning based on knowledge distillation. This kind of method mainly projects old knowledge by transferring knowledge in old tasks to new tasks through knowledge distillation. LwF [21] is the first method 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 [30] algorithm uses knowledge distillation to avoid excessive deterioration of knowledge in the network, while BiC [36] 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, IOD is less explored. Meanwhile, the high complexity of the detection task also adds the difficulty of incremental object detection. [31] proposed to apply LwF to Fast RCNN detector [10], which is the first work on incremental object detection. Thereafter, some researchers move this area forward. [28] proposed SID approach for IOD on anchor-free detector and conducted experiments on FCOS [33] and CenterNet [39]. [18] studied object detection based on class-incremental learning on Faster RCNN detector with emphasis on few-shot scenarios, which is also the focus of ONCE algorithm [29]. [17] designed an incremental object detection system with RetinaNet detector [24] on edge devices. the latest work, [15] introduced the concept of incremental learning when defining the problems of Open World Object Detection (OWOD). However, existing IOD distillation framework does not pay enough attention to the significant role of head. In this study, we found head has its great significance in the area of IOD.

Knowledge Distillation for Object Detection. Knowledge distillation [2, 4] 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 frequently [8]. [5] implemented distillation for all components of Faster RCNN (including backbone, proposals in RPN, and head). To imitate the high-level feature response of the teacher model with the student model, [35] proposed a distillation method based on fine-grained feature imitation. By synthesize category-conditioned objects through inverse mapping, [3] proposed a data-free distillation technology applicable for object detection, but the method would trigger dream-image. [13] believing that foreground and background both play an unique role in object detection, proposed an object detection distillation method that decoupled foreground and background. [38] proposed a localization distillation method introducing knowledge distillation into the regression branch of the detector, so as to enable the student network to solve the localization ambiguity as the teacher network.

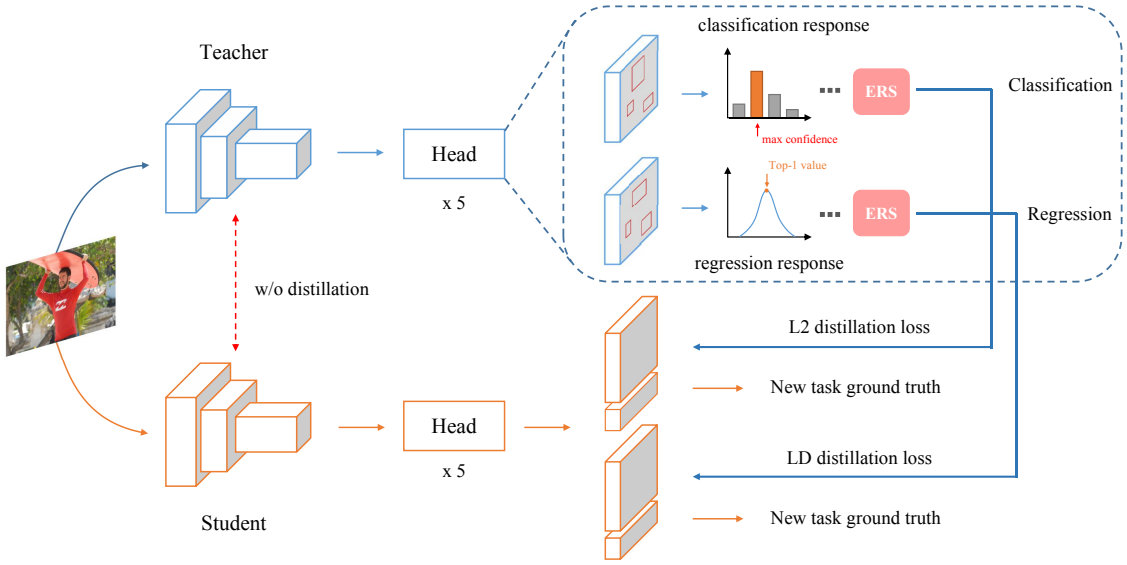


Figure 2. Overall structure of elastic response distillation for incremental object detection.

3. Method

3.1. Motivation

The purpose of IOD is to transfer old knowledge to student detector, and this knowledge could be the features of intermediate layers in backbone or neck, or the soft targets in head. Unlike feature-based method, response-based method can provides the reasoning information of teacher detector [14,27]. Therefore, we incrementally learn a strong and efficient student object detector by the distillation of incremental knowledge from responses of different heads.

3.2. Overall Structure

The overall framework of the proposed method is shown in Figure 2. Firstly, ERD is applied to learn elastic response from the classification head and regression head of the teacher detector. Secondly, incremental localization distillation loss is applied to enhance the localization information extraction ability of the student detector. Notably, the ERS strategies are proposed to gain more meaningful incremental responses from the teacher detector, that is, selective calculation of the distillation loss from the response provided by the teacher detector. The overall learning target of the student detector is therefore defined as,

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{model} &+ \lambda_1 \mathcal{L}_{ERD\_cls}(\mathcal{C}_{\mathcal{T}}, \mathcal{C}_{\mathcal{S}}) \\ &+ \lambda_2 \mathcal{L}_{ERD\_bbox}(\mathcal{B}_{\mathcal{T}}, \mathcal{B}_{\mathcal{S}})\end{aligned} \tag{1}$$

where λ_i is the parameters that balances the weights of different loss terms, and the subscript $\mathcal{T}$ and $\mathcal{S}$ separately represents teacher and student. The loss term $\mathcal{L}_{model}$ is the detector-specific classification and localization loss to train student detector for detecting new objects. The second loss term $\mathcal{L}_{ERD\_cls}$ is the incremental L2 distillation loss for the classification branch. The third loss term $\mathcal{L}_{ERD\_bbox}$ is the incremental localization distillation loss for the regression branch. Both $\mathcal{L}_{ERD\_cls}$ and $\mathcal{L}_{ERD\_bbox}$ are used for the outputs of old classes. We use $\lambda_1 = \lambda_2 = 1$ by default.

In the following subsection, we mainly present ERD and ERS for GFLV1 [20] while we generalize our method to FCOS in Table 7, which illustrates the effectiveness of our method.

3.3. ERD at Classification Head

The soft predictions from the classification head contains the knowledge of various categories discovered by the teacher detector. Through the learning of soft predictions, the student model can inherit hidden knowledge, which is intuitive for classification tasks [14]. Let $\mathcal{T}$ be the teacher model, we use SoftMax to transform logits $\mathcal{C}_{\mathcal{T}}$ into distribution, then the outputting probability distribution $\mathcal{P}_{\mathcal{T}}$ is defined as,

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

Similarly, we define $\mathcal{P}_{\mathcal{S}}$ for the student model $\mathcal{S}$ as $\mathcal{P}_{\mathcal{S}} = \mathrm{SoftMax}\left(\mathcal{C}_{\mathcal{S}}/t\right)$, where t is the temperature factor to soften the probability distribution for $\mathcal{P}_{\mathcal{T}}$ and $\mathcal{P}_{\mathcal{S}}$.

Previous work usually directly utilizes all the predicted responses in classification head and treat each position equally, e.g. $\mathcal{L}_{cls} = \sum_{i=1}^{N} \mathcal{L}_{KL}\left(\mathcal{P}_{\mathcal{T}}, \mathcal{P}_{\mathcal{S}}\right)$. 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. Here, we selectively calculate the distillation loss from response, thus the incremental distillation loss at classification head is as follows,

$$\mathcal{L}_{ERD\_cls}\left(\mathcal{C}_{\mathcal{T}}, \mathcal{C}_{\mathcal{S}}\right) = \sum_{i=1}^{m} \left(\mathcal{C}_{\mathcal{T}}{}^{i} - \mathcal{C}_{\mathcal{S}}{}^{i}\right)^2 \tag{3}$$

where ${\mathcal{C}_{\mathcal{T}}}^i$ is one of the m selected category responses from the teacher detector using the new data. ${\mathcal{C}_{\mathcal{S}}}^i$ is the corresponding category responses of the student detector. By distilling the selected responses, the student detector incrementally inherits the old knowledge of the teacher detector.

3.4. ERD at Regression Head

The bounding box responses from the regression branch are also important for IOD. Contrary to the discrete class information, the output of regression branch may provide a regression direction contradicting the real direction. Even if an image does not contain any objects of old categories, the regression branch would still predict bounding boxes, though the confidence is low. This poses a challenge for transferring regression knowledge from teacher detector to student detector. Furthermore, in previous work, only the bounding boxes of objects with high classification confidence are utilized as the regression knowledge from the teacher detector, which ignores the localization information of regression branch.

Benefitting from the general representation of distributions for bounding boxes from GFLV1 detector, each edge e of a bounding box can be represented as a probability distribution through SoftMax function [38]. Thus, the probability matrix of each bounding box $\mathcal{B}$ can be defined as,

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

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

$$\mathcal{L}_{LD}^j = \sum_{e \in \mathcal{B}} \mathcal{L}_{KL}^e \left({\mathcal{B}_{\mathcal{T}}}^j, {\mathcal{B}_{\mathcal{S}}}^j \right) \tag{5}$$

Finally, the incremental localization distillation loss at regression head is defined as,

$$\mathcal{L}_{ERD\_bbox} \left(\mathcal{B}_{\mathcal{T}}, \mathcal{B}_{\mathcal{S}} \right) = \sum_{j=1}^{J} \mathcal{L}_{LD}^j \tag{6}$$

where ${\mathcal{B}_{\mathcal{T}}}^j$ is the regression response of the teacher detector from J selected bounding boxes using the new data, and ${\mathcal{B}_{\mathcal{S}}}^j$ is the corresponding regression response of the student detector. Notably, the incremental localization distillation provides extra localization information.

3.5. Elastic Response Selection

As shown in Figure 1, choosing all the responses leads to bad performance, thus response selection is important to prevent catastrophic forgetting. Then a natural question arises: *how to select responses as the distillation nodes.* Common selection strategies depend on sensitive hyperparameters such as setting confidence thresholds or selecting Top-K scores. These empirical practices may result in a consequence that small thresholds ignore several old objects while large ones bring negative responses.

To solve the above problem, we propose the ERS strategy as illustrated in Algorithm 1. We respectively select responses from the classification head and regression head as the distillation nodes.

Classification head. Statistical characteristics are utilized to select responses of the classification head, as described in L-3 to L-11. Specifically, We first calculate the confidence score of each node. After that, we calculate the mean μ'_C and standard deviation σ'_C in L-5 and L-6. With these statistics, the elastic threshold τ'_C can be obtained in L-7. Finally, we select response nodes whose confidence scores are greater than the threshold τ'_C in L-8 to L-11 as the distillation nodes.

Regression head. Statistical distribution information is utilized to select responses of the regression head, as described in L-13 to L-22. For GFLV1, a certain and unambiguous bounding box usually has a sharper distribution. Therefore, the Top-1 value is relatively larger if the distribution is sharp. Based on the above statistical properties, the Top-1 value is used to measure the confidence of each bounding box. Specifically, we first select the Top-1 value of each distribution. After that, we calculate the mean μ'_B and the standard deviation σ'_B of all Top-1 values in L-15 and L-16. Then, the threshold τ'_B is obtained in L-17. Finally, we select these candidates whose confidence are greater than the threshold τ'_B in L-18 to L-20. The *nms* operator returns a sampled set that is filtered by NMS in L-21.

The motivations behind ERS are explained as follows:

Maintain fairness among different responses. In a normal distribution, approximately 16% and 2.5% of the samples are separately distributed in the interval $[\mu + \sigma, +\infty]$ and $[\mu + 2\sigma, +\infty]$. In our case, the number of positive responses are distributed from 100 to 1000 per image. In contrast, the strategy of selecting all or top-k responses leads to unfairness for different responses.

Elastic selection by statistical characteristics. In the IOD task, responses generated by background objects may overwhelm the responses generated by foreground objects. Thus a high μ indicates high-quality candidates, while a low one indicates low-quality candidates. ERS can elastically select enough positive responses following the statistical characteristics of different branches.

4. Experiments and Discussions

In this section, we perform experiments on MS COCO 2017 [7] using the baseline detector GFLV1 to validate our method. Then, we perform ablation studies to prove the effectiveness of each component. Finally, we discuss the

Table 1. Incremental results (%) based on GFLV1 detector on COCO benchmark under different scenarios. ("Δ" represents an improvement over Catastrophic Forgetting. "∇" represents the gap towards the Upper Bound.)

Scenarios	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Full data	Upper Bound	40.2	58.3	43.6	23.2	44.1	52.2
40 classes + 40 classes	Catastrophic Forgetting	17.8	25.9	19.3	8.3	19.2	24.6
	LwF [21]	17.2 ($\Delta-0.6/\nabla23.0$)	25.4	18.6	7.9	18.4	24.3
	RILOD [17]	29.9 ($\Delta12.1/\nabla10.3$)	45.0	32.0	15.8	33.0	40.5
	SID [28]	34.0 ($\Delta16.2/\nabla6.2$)	51.4	36.3	18.4	38.4	44.9
	ERD	**36.9** ($\Delta19.1/\nabla3.3$)	**54.5**	**39.6**	**21.3**	**40.4**	**47.5**
50 classes + 30 classes	Catastrophic Forgetting	14.1	20.6	15.2	7.0	14.5	19.2
	LwF [21]	5.0 ($\Delta-9.1/\nabla35.2$)	9.5	4.6	5.0	6.7	5.7
	RILOD [17]	28.5 ($\Delta14.4/\nabla11.7$)	43.2	30.2	15.4	31.6	38.0
	SID [28]	33.8 ($\Delta19.7/\nabla6.4$)	51.0	36.1	17.6	38.1	45.1
	ERD	**36.6** ($\Delta22.5/\nabla3.6$)	**54.0**	**38.9**	**19.4**	**40.4**	**48.0**
60 classes + 20 classes	Catastrophic Forgetting	9.8	14.0	10.6	4.3	14.1	13.5
	LwF [21]	5.8 ($\Delta-4.0/\nabla34.4$)	10.8	5.3	4.0	8.5	7.7
	RILOD [17]	25.4 ($\Delta15.6/\nabla14.8$)	38.8	26.8	13.9	29.0	33.7
	SID [28]	32.7 ($\Delta22.9/\nabla7.5$)	49.8	34.6	17.2	37.6	43.5
	ERD	**35.8** ($\Delta26.0/\nabla4.4$)	**52.9**	**38.4**	**20.6**	**39.4**	**46.5**
70 classes + 10 classes	Catastrophic Forgetting	4.3	6.5	4.5	2.1	5.1	6.8
	LwF [21]	7.1 ($\Delta2.8/\nabla33.1$)	12.4	7.0	4.8	9.5	10.0
	RILOD [17]	24.5 ($\Delta20.2/\nabla15.7$)	37.9	25.7	14.2	27.4	33.5
	SID [28]	32.8 ($\Delta28.5/\nabla7.4$)	49.0	35.0	17.1	36.9	44.5
	ERD	**34.9** ($\Delta30.6/\nabla5.3$)	**51.9**	**37.4**	**18.7**	**38.8**	**45.5**

application scenario of our method.

Implementation Details. We build our method on top of the GFLV1 detector. The teacher and student detectors defined in our experiments are standard GFLV1 architectures. For GFLV1 detector, ResNet-50 is used as its backbone, FPN [22] is used as its neck. We train the detector to following the same settings as the original paper. All the experiments are performed on 8 NVIDIA Tesla V100 GPUs, with a batch size of 8. For the parameter α, we use $\alpha_1 = \alpha_2 = 2$ by default.

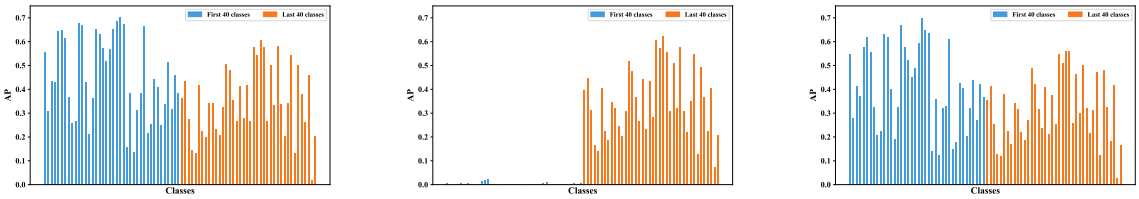
Datasets and Evaluation Metric. MS COCO 2017 is a challenging dataset in object detection which contains 80 object classes. For experiments on this dataset, we use the train set for training and the minival set for testing. The standard COCO protocols are used as the evaluation metrics, i.e. AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L.

Experiment Setup. The detector is trained by 12 epochs (1x mode) for each incremental step. The settings are consistent for all the detectors in the different scenarios. Specifically, we conduct experiments in the following Class Incremental Learning scenarios with different splits:

(i) One-step: 40 + 40 to 70 + 10 with a step size of 10 classes, increasing base class numbers and decreasing new class numbers. **(ii)** Multi-step: two-step and four-step settings with 20 new classes and 10 new classes respectively added each time. **(iii)** Last 40 + First 40: last 40 classes as the base classes and first 40 classes as new classes.

4.1. Overall Performance

One-step. We reported the incremental results under the first 40 classes + last 40 classes scenario in Table 1. In this case, we observe that if the old detector and the new data are directly used to conduct fine-tuning process, then the AP drops to 17.8% as compared to the 40.2% in full data training (Upper Bound). This is because the fine-tuning process makes the detector's memory of old objects close to 0, resulting in catastrophic forgetting (ref to Figure 3b). Our method far outperformed fine-tuning across various evaluation criteria. Concretely, when the IoU is 0.5, 0.75 and 0.95, the AP respectively improves by 19.1%, 28.6% and 20.3%, which indicates that our method can well address the catastrophic forgetting problem. Notably, even compared with the upper bound where the entire dataset is used for training, our method only has a performance gap of 3.3%. It indicates that the student detector maintains a good memory of old objects while is able to learn knowledge of new objects. Remarkably, as shown in Table 1, the performance of fine-tuning decreases drastically as the number of new



(a) Upper Bound. (b) Catastrophic Forgetting. (c) Elastic Response Distillation.

Figure 3. AP of per-class among different learning schemes. (a) Detector is trained with all data.(b) Student detector is fine-tuned with new classes.(c) Student detector is learned via ERD.

Table 2. Incremental results (AP/AP_{50}, %) based on GFLV1 detector on COCO benchmark under the four-step setting. A(a-b) is the one-step normal training for categories a-b and +B(c-d) is the incremental training for categories c-d.

Method	A(1-40)	+B(40-50)	+B(50-60)	+B(60-70)	+B(70-80)	A(1-80)
Catastrophic Forgetting	45.7/ 66.3	5.8/ 8.5	5.7/ 8.3	6.3/ 8.5	3.3/ 4.8	40.2/ 58.3
RILOD [17]		25.4/ 38.9	11.2/ 17.3	10.5/ 15.6	8.4/ 12.5	
SID [28]		34.6/ 52.1	24.1/ 38.0	14.6/ 23.0	12.6/ 23.3	
ERD		**36.4/ 53.9**	**30.8/ 46.7**	**26.2/ 39.9**	**20.7/ 31.8**	

Table 3. Incremental results (AP/AP_{50}, %) based on GFLV1 detector on COCO benchmark under the two-step setting, where the meanings of A(a-b) and +B(c-d) are similar to Table 2.

Method	A(1-40)	+B(40-60)	+B(60-80)	A(1-80)
Catastrophic Forgetting	45.7/ 66.3	10.7/ 15.8	9.4/ 13.3	40.2/ 58.3
RILOD [17]		27.8/ 42.8	15.8/ 4.0	
SID [28]		34.0/ 51.8	23.8/ 36.5	
ERD		**36.7/ 54.6**	**32.4/ 48.6**	

classes decreases (17.8% to 4.3%) under different incremental conditions (50 classes + 30 classes, 60 classes + 20 classes, and 70 classes + 10 classes), while our method still remains a high level performance (36.9% to 34.9%). To sum up, our method has great robustness for overcoming catastrophic forgetting.

In addition, we compare our method with LwF [21], RILOD [17] and SID [28] as well. Table 1 shows that although LwF works well in incremental classification, it has even lower AP than direct fine-tuning in detection task, which reveals naively borrowing methods from incremental classification area would generate negative influence to the IOD task. For the typical IOD approaches (i.e. RILOD and SID), in order to fairly compare with them, we replicate them based on the GFLV1 detector. For RILOD, we completely follow their implementations. For SID, we use the component with the greatest improvement in the paper. When compared with the aforementioned approaches, the proposed method achieves state-of-the-art performance in four incremental scenarios. Notably, the performance improvements are all significant.

To put it more intuitively, we visualize the AP of all classes in first 40 classes and last 40 classes in Figure 4. Furthermore, the per-class results are visualized in Figure 3, where the blue columns denote the per-class AP in first 40 classes, and the orange columns denote the per-class AP in last 40 classes. As Figure 3 shows, the proposed method reserves a majority of information for the old classes while learns knowledge from newly coming classes.

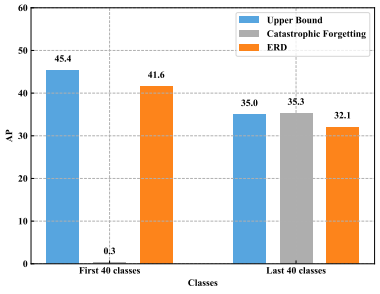


Figure 4. AP of all classes in first 40 classes vs. last 40 classes.

Multi-step. We reported the incremental results under multi-step settings to illustrate the continual learning ability of the proposed method. In Table 3 (two-step) and Table 2 (four-step), our method outperforms fine-tuning by a large margin for each incremental step on both multi-step settings. This is because, the detector continuously obtains knowledge from the dynamic data flow, new knowledge interferes with the old one, triggering catastrophic forgetting, while ERD provides valuable responses in each step to alleviate the problem. In addition, ERD performs favorably well on each incremental step against the previous state-of-the-art. Remarkably, the AP of RILOD and SID decreases drastically as the number of new classes increases (27.8% to 15.8% and 34.0% to 23.8%, 25.4% to 8.4% and 34.6% to 12.6%) under two multi-step settings, while our method still remains a high performance (36.7% to 32.4% and 36.4% to 20.7%). ERD is able to restore the previous class perfor-

Table 4. Ablation study (%) based on GFLV1 detector using the COCO benchmark under first 40 classes + last 40 classes. ("Δ" represents an improvement over Catastrophic Forgetting. "∇" represents the gap towards the 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(Δ13.7/∇8.7)	48.3	33.4	17.7	35.3	41.3
KD:all cls	23.8(Δ6.0/∇16.4)	36.6	24.9	11.8	27.2	32.9
KD:all reg	13.0($\Delta-4.8$/∇27.2)	21.1	13.4	5.0	14.7	18.6
ERD:cls + ERS	33.2(Δ15.4/∇7.0)	51.2	35.2	18.5	37.8	43.8
ERD:cls + reg + ERS	**36.9**(Δ19.1/∇3.3)	**54.5**	**39.6**	**21.3**	**40.4**	**47.5**

Table 5. Varying α for ERS (%).

Threshold	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$\alpha_{1,2}=1,1$	36.5	54.2	39.2	20.6	40.3	46.9
$\alpha_{1,2}=1,2$	36.8	54.4	39.6	21.5	40.4	47.5
$\alpha_{1,2}=2,1$	36.7	54.3	39.6	21.5	40.4	47.6
$\alpha_{1,2}=2,2$	**36.9**	**54.5**	39.6	21.3	40.4	47.5

mance to a respectable level. It indicates that the proposed method has a significant ability to alleviate catastrophic forgetting.

4.2. Ablation Study

We validate the effectiveness of each component of the proposed method on MS COCO. In Table 4, "KD" denotes only use the distillation loss without selection, while "ERD" denotes the selection strategy are introduced. "all cls + all reg" denotes responses from both classification and regression branch are treated equally in the incremental process, which is used as our baseline. "all cls" denotes all classification responses in the incremental process are treated equally. "all reg" denotes all regression responses are treated equally in the incremental process. "cls + ERS" denotes that the ERS strategy is employed on the classification branch to conduct incremental distillation, as shown in Equation 3. "cls + reg + ERS" denotes responses on regression branch are added as well, as shown in Equation 6. In Table 4, distillation on either classification or regression branch can merely obtain a low performance (i.e. 23.8% and 13.0% of AP). When all responses from the regression branch are used, AP is even lower than the fine-tuning strategy, which supports our findings shown in Figure 1. Comparatively, when combined responses from classification with regression branch, the AP reaches to 31.5%. When ERS is involved to select responses from classification branch, the student detector can obtain higher results (i.e. 33.2%). Furthermore, when performing ERS on regression branch, the AP continually increases to 36.9%, which is a dramatically improvement (i.e. 5.4%) compared with the baseline. All these results clearly point out the advantages of the proposed method.

Parameter α. We conduct four groups of experiments to investigate the robustness of the proposed method on the parameter α, which is utilized to elastically select positive responses from classification head and regression head. In table 5, different combinations of α_1 and α_2 are chosen from the set ([1,1], [1,2], [2,1], [2,2]) to perform the training process. We observe that the maximum performance gap is merely 0.4%, which indicates the proposed ERS is insensitive to the parameter α. Therefore, the proposed ERS can be regarded as nearly parameter-free.

4.3. Discussions

In this section, we present further insights into response-based IOD.

Generalization on different detectors. We perform extended experiments to validate the generality of the proposed method on the FCOS detector. For FCOS, we only need to replace the LD loss with GIoU loss. For both regression and centerness branches, we employ the statistical characteristics of category information to determine the elastic responses. Other settings are consistent with the proposed method. Results in Table 7 show that our method still brings stable gain regardless of the detector structure. To sum up, we only need to adjust our method slightly for adapting the head of different detectors, which indicates the generalizability of the proposed method.

Elastic response helps both learning and generalization. Considering the long-tail distribution of COCO, we configure an experiment under the last 40 classes + first 40 classes scenario. In this case, objects of the first 40 classes contain more information, which means more responses could be obtained. As shown in Table 6, the performance can be further improved, with a smaller gap 2.7% against the upper bound, which indicates the proposed method benefits from more responses to alleviate catastrophic forgetting.

Distances of different components. In order to verify why the response-based distillation can attain higher performance compared to feature-based solutions, we randomly

Table 6. Incremental results (%) based on GFLV1 detector on COCO benchmark under last 40 classes + first 40 classes. ("Δ" represents an improvement over Catastrophic Forgetting. "∇" represents the gap towards the 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 [21]	20.5 ($\Delta - 2.1/\nabla 19.7$)	29.9	22.1	13.0	22.5	25.3
RILOD [17]	34.1 ($\Delta 11.5/\nabla 6.1$)	51.1	36.8	19.1	38.0	43.9
SID [28]	33.5 ($\Delta 10.9/\nabla 6.7$)	50.9	36.3	19.0	37.7	43.0
ERD	**37.5** ($\Delta 14.9/\nabla 2.7$)	**55.1**	**40.4**	**21.3**	**41.1**	**48.2**

Table 7. Incremental results (%) based on FCOS detector.

Model	Method	Centerness	Elastic	AP	AP_{50}	AP_{75}
FCOS	Upper Bound	✔		38.5	57.5	41.3
	Fine-tuning	✔		16.7	25.6	17.9
	All			31.5	49.6	33.2
		✔		31.7	49.9	33.3
	ERD		✔	**34.4**	**52.8**	36.5
		✔	✔	34.2	52.4	36.6

Table 8. Quantitative results (%) of feature-based and response-based solutions.

Method	Feature	Response	Elastic	AP	AP_{50}	AP_{75}
All		✔		31.5	48.3	33.4
FPN + All	✔	✔		32.5	49.7	34.4
FPN + ERS	✔	✔	✔	36.5	54.0	39.0
ERD		✔	✔	**36.9**	**54.5**	**39.6**

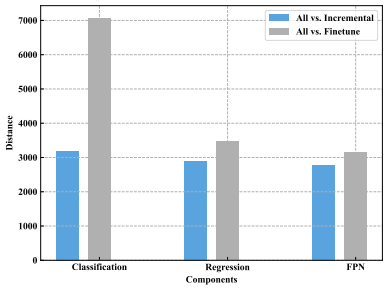


Figure 5. Feature distance analysis of different components.

choose 10 images from COCO minival and calculate the L2 feature distances in varying components using different training strategies. As shown in Figure 5, "All" denotes the full data training strategy; "Finetune" denotes the fine-tuning strategy; "Incremental" denotes the proposed method. When compared "All vs. Incremental" with "All vs. Finetune", the distance of classification head is larger than that of regression head, and distances of the former two are larger than that of FPN (i.e. feature layers). It means that the response-based distillation provides more contributions to alleviate catastrophic forgetting.

Quantitative analysis of feature-based and response-based solution. Besides qualitative analysis in Figure 5, we further analyze the quantitative difference between feature-based and response-based solutions. As shown in table 8, when combined FPN (i.e. feature layers) with all responses in head, it would produce positive effects. The reason is that feature layers provide more capacity for the learning procedure compared with head alone. Nevertheless, when the ERS strategy is added to head, the final performance is dramatically improved (32.5% vs. 36.9%), while the involvement of feature layers brings negative impacts (-0.4% in AP). We guess a feasible explanation could be the optimization directions are changed, as feature layers tend to a global direction while head expects to reserve positive responses after selection.

5. Conclusion

In this paper, we elaborately design a response-based incremental paradigm in object detection field, which significantly alleviates the catastrophic forgetting problem. Firstly, we learn responses from the classification head and regression head, and specifically introduce incremental localization distillation in regression responses. Secondly, the elastic selection strategy is designed to provide suitable responses in different heads. Extensive experiments validate the effectiveness of the proposed method. Finally, elaborate analysis discusses the generalizability of our method and essential differences between response-based and feature-based distillation for incremental detection task, which provides insights for further exploration in this field.

Broader Impact

The study of IOD would make us better understand the formation mechanism of neural networks from the system level, which provides a technical basis for the development of lifelong learning mechanism. The ultimate goal is that detectors can perform continual learning like creatures. However, models after incremental learning may lead to some privacy issues, while we can mitigate it by limiting the accessibility of trained models.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Computer Vision - ECCV 2018 - 15th European Conference*, volume 11207, pages 144–161. Springer, 2018. 2

[2] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006*, pages 535–541. ACM, 2006. 2

[3] 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*, pages 3288–3297. IEEE, 2021. 2

[4] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 3430–3437, 2020. 2

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems NeurIPS 2017,*, pages 742–751, 2017. 1, 2

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5008–5017. Computer Vision Foundation / IEEE, 2021. 1

[7] 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. 4

[8] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7842–7851, 2021. 2

[9] Tao Feng, Kaifan Ji, Ang Bian, Chang Liu, and Jianzhou Zhang. Identifying players in broadcast videos using graph convolutional network. *Pattern Recognit.*, 124:108503, 2022. 1

[10] Ross Girshick. Fast r-cnn, 2015. 2

[11] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015. 1

[12] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819, 2021. 1, 2

[13] 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*, pages 2154–2164. Computer Vision Foundation / IEEE, 2021. 2

[14] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv preprint*, 1503.02531, 2015. 1, 3

[15] 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*, pages 5830–5840. Computer Vision Foundation / IEEE, 2021. 2

[16] 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. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2

[17] 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 *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, SEC 2019*, pages 113–126. ACM, 2019. 2, 5, 6, 8

[18] Pengyang Li, Yanan Li, and Donghui Wang. Class-incremental few-shot object detection, 2021. 2

[19] 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*, pages 11632–11641. Computer Vision Foundation / IEEE, 2021. 1

[20] 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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS 2020*, 2020. 2, 3

[21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018. 2, 5, 6, 8

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 936–944. IEEE Computer Society, 2017. 5

[23] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 2999–3007. IEEE Computer Society, 2017. 1

[24] 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. 2

[25] Liyang Liu, Zhanghui Kuang, Yimin Chen, Jing-Hao Xue, Wenming Yang, and Wayne Zhang. Incdet: In defense of elastic weight consolidation for incremental object detection. *IEEE Trans. Neural Networks Learn. Syst.*, 32(6):2306–2319, 2021. 1

[26] M. Mccloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989. 1

[27] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019,*, pages 4696–4705, 2019. 1, 3

[28] 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. 1, 2, 5, 6, 8

[29] Juan-Manuel Pérez-Rúa, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 13843–13852. Computer Vision Foundation / IEEE, 2020. 2

[30] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 5533–5542. IEEE Computer Society, 2017. 2

[31] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 3420–3429. IEEE Computer Society, 2017. 2

[32] Ruoyu Sun, Fuhui Tang, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Distilling object detectors with task adaptive regularization. *CoRR*, abs/2006.13108, 2020. 1

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 9626–9635. IEEE, 2019. 1, 2

[34] Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning. *CoRR*, abs/1904.07734, 2019. 1

[35] 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*, pages 4933–4942. Computer Vision Foundation / IEEE, 2019. 2

[36] 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*, pages 374–382. Computer Vision Foundation / IEEE, 2019. 2

[37] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7476–7485, 2021. 1

[38] Zhaohui Zheng, Rongguang Ye, Ping Wang, Jun Wang, Dongwei Ren, and Wangmeng Zuo. Localization distillation for object detection, 2021. 2, 4

[39] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2