Augmented Box Replay: Overcoming Foreground Shift for Incremental Object Detection

Yuyang Liu[1,2,3]   Yang Cong[4,†]   Dipam Goswami[5]   Xialei Liu[6]   Joost van de Weijer[5,7]

[1]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[2] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences   [4]South China University of Technology
[5]Computer Vision Center, Barcelona   [6]VCIP, CS, Nankai University
[7]Department of Computer Science, Universitat Autònoma de Barcelona

liuyuyang@sia.cn, congyang81@gmail.com {dgoswami,joost}@cvc.uab.es, xialei@nankai.edu.cn

Abstract

In incremental learning, replaying stored samples from previous tasks together with current task samples is one of the most efficient approaches to address catastrophic forgetting. However, unlike incremental classification, image replay has not been successfully applied to incremental object detection (IOD). In this paper, we identify the overlooked problem of foreground shift as the main reason for this. Foreground shift only occurs when replaying images of previous tasks and refers to the fact that their background might contain foreground objects of the current task. To overcome this problem, a novel and efficient Augmented Box Replay (ABR) method is developed that only stores and replays foreground objects and thereby circumvents the foreground shift problem. In addition, we propose an innovative Attentive RoI Distillation loss that uses spatial attention from region-of-interest (RoI) features to constrain current model to focus on the most important information from old model. ABR significantly reduces forgetting of previous classes while maintaining high plasticity in current classes. Moreover, it considerably reduces the storage requirements when compared to standard image replay. Comprehensive experiments on Pascal-VOC and COCO datasets support the state-of-the-art performance of our model [1].

1. Introduction

The field of deep learning has witnessed remarkable progress recently, and state-of-the-art object detection models [5, 19, 23, 50, 51] have been developed that performs exceptionally well on benchmark datasets. However, these

---

[1]Code is available at https://github.com/YuyangSunshine/ABR_IOD.git
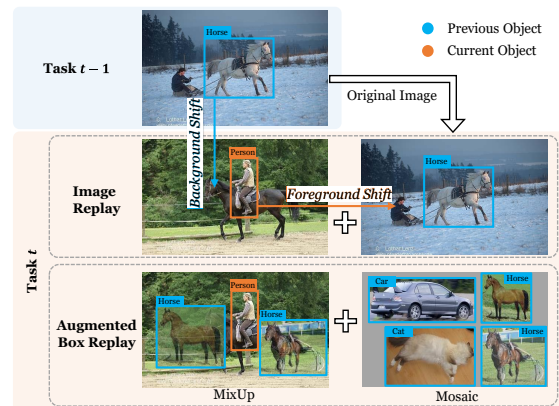[†]The corresponding author is Prof. Yang Cong.

Figure 1: *Background Shift* and *Foreground Shift* for image replay settings. For each task, only the new classes are annotated while the other objects are considered as background (bkg). Moving from task $t-1$ to task t, the definition of the bkg changes, referred to as *background shift* [8]. When current task samples are trained with exemplars from previous tasks, another critical problem-*Foreground Shift* occurs due to varying annotations of *new classes* between new samples (person as foreground) and exemplars (person as bkg) in the same task. Our augmented box replay method resolves these problems by mixing previous objects in the bkg of new images or fusing together for training.

models are typically designed to learn from data in a static manner, assuming that all object classes are available at once during training. In real-world scenarios, new object classes may emerge over time, making it necessary to update the model with new data. The inability to learn incrementally is a significant limitation for object detectors, particularly in cases of limited data storage capacity or data privacy concerns [10, 45]. Therefore, developing incremental object detection (IOD) methods has become an essential and challenging task in real-world applications.

SOTA object detectors experience a phenomenon known as catastrophic forgetting [46], where their performance on previous classes degrades after learning new classes. This issue is commonly observed in incremental settings [10] and can be mitigated by balancing model stability (retaining previous information) and plasticity (learning new information without forgetting previous knowledge). While most studies in incremental learning are based on image classification [2, 30, 33, 49], recently it has been studied in the context of object detection [7, 9, 22, 47, 54] and semantic segmentation [13, 20, 63]. A critical aspect in IOD is the background shift, also known as missing annotations [7, 47] which occurs due to the presence of multiple class objects in an image. Objects belonging to previous or future tasks in incremental object detection are often not annotated and assigned to the background class, as annotations are only available for classes in the current task.

One of most efficient approaches in incremental classification is rehearsal-based strategy with storing images [6, 49]. However, directly applying the replay images into IOD will cause the unlabelled objects of current classes in the replay images to be treated as background by the model. Consequently, the new objects will be background in replay images, while regarded as foreground in the new images. This leads to a contradiction between the foreground annotations in the exemplars and the current images as illustrated in Fig. 1. We refer to this problem as foreground shift which affects the plasticity of the current model.

To overcome the foreground shift for image replay in IOD, we propose a novel method called Augmented Box Replay (ABR). ABR uses mixup and mosaic box augmentation strategies to replay previous objects as an alternative to image replay for training in the current task. Compared to storing images in memory, ABR stores approximately four times as many object instances with the same storage requirements. To more effectively address catastrophic forgetting, we introduce a novel Attentive RoI Distillation loss that utilizes spatial attention from region-of-interest (RoI) features to align the most informative features of the previous and new models and correct the anchor position deviations of proposal pairs.

The proposed method is experimentally evaluated on Pascal-VOC and COCO datasets, and significantly outperforms SOTA methods in multiple settings. Our main contributions are three-fold:

- This paper is the first to identify the critical foreground shift issue which has hampered the usage of replay methods for IOD. We propose Augmented Box Replay as a solution that reduces the memory requirements, eliminates the foreground shift, and improves the model stability and plasticity.

- We propose an innovative Attentive RoI Distillation loss to focus the current model on important location
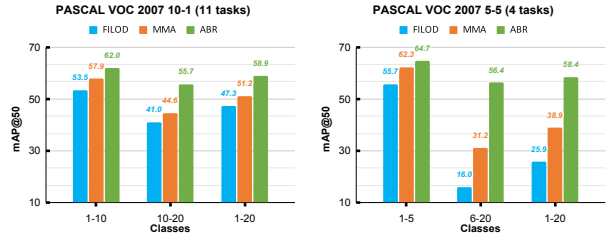

Figure 2: Our ABR method is especially well on the challenging longer sequences (10-1) and when starting with a small initial task (5-5). We compare here with state-of-the-art methods FILOD [47] and MMA [7].

and feature information from the previous model and further reduce catastrophic forgetting.

- Our method outperforms state-of-the-art methods across multiple datasets and settings, showcasing its practicality and effectiveness. Especially, on the more challenging longer task sequences and the difficult scenario with a small initial task, our method obtains significant performance gains (see Fig. 2).

2. Related Work

Object Detection: Detector networks can be categorized into one-stage [5, 35, 38, 50, 55, 56] and two-stage [19, 23, 34, 51] detectors. One-stage detectors which directly predict the output objects are comparatively faster while the two-stage detectors are generally superior in performance. The two-stage methods first extract regions of interests (RoIs) using a network [51] and then obtain the final classification and regression outputs using a multi-layer network on the RoIs. Since these architectures perform poorly in incremental settings, we extend the two-stage Faster R-CNN [51] network such that it can learn new object classes over time.
Incremental Learning: Class-incremental learning [10, 45] and catastrophic forgetting [46] has been explored extensively for image classification [6, 33, 49] problems. The previous works can be categorized into rehearsal-based, parameter-isolation and regularization-based methods. Rehearsal-based methods store training samples [6, 32, 49] from previous tasks or generates training data [29, 53, 58]. Parameter-isolation methods [40, 43, 44, 60] modify the initial network to accommodate new classes. Prior-focused regularization methods constrain learning on new classes and penalizing updating on weights [2, 30] or gradients [42] while data regularization methods perform distillation [25] between the intermediate features [14, 15, 26, 33] or attention maps [12] of the teacher model and the current student model to reduce forgetting. Other methods use embedding networks [61] or classifier drift correction [3] to address the changing class distributions. In our work, we focus on rehearsal-based and regularization-based methods.
Incremental Object Detection: Most of the recent works on incremental object detection use the Faster R-CNN [51]
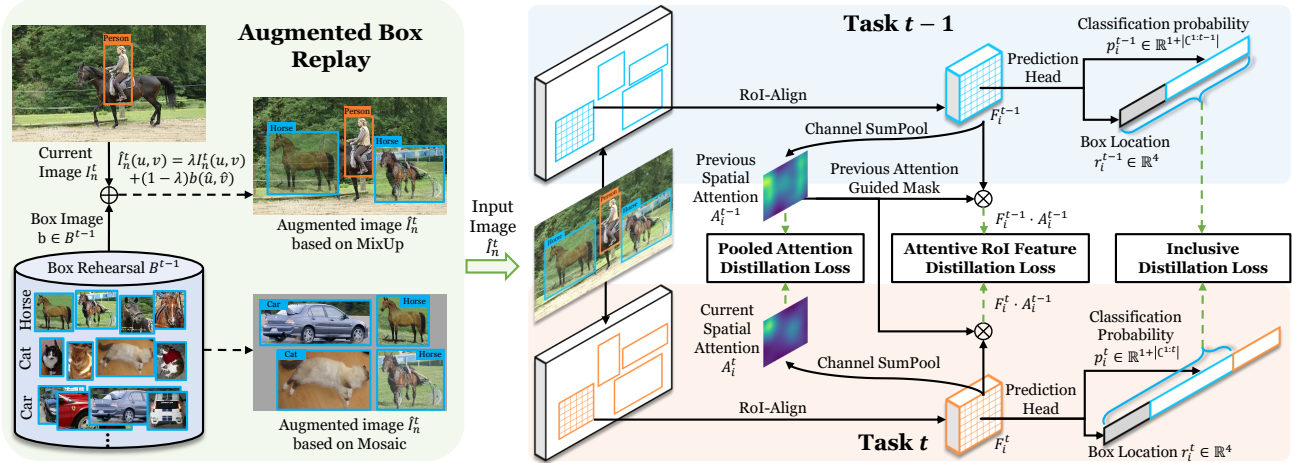
Figure 3: Illustration of our proposed framework, which highlights the key novelties of Augmented Box Replay (ABR) and Attentive RoI Distillation. ABR fuses prototype object b from Box Rehearsal B^{t-1} into the current image I_n^t using mixup or mosaic. Attentive RoI Distillation uses pooled attention A_i and masked features $F_i \cdot A_i^{t-1}$ to constrain the model to focus on important information from previous model. Inclusive Distillation Loss overcomes catastrophic forgetting based on ABR.

architecture and performs distillation on the intermediate features [7, 22, 39, 47, 59, 65], the region proposal network [7, 47, 65] and head layers [17]. Relatively few works [31, 48, 52] used one-stage architectures for incremental learning. Although the background shift issue was partially addressed in [65] by preventing previous class regions to be sampled as background but it was highlighted recently in [7, 47]. [7] proposed an unbiased classifier training loss and classifier distillation loss to explicitly tackle the background shift. EWC [30] has been adapted by [37] for object detection. While some methods replay images for finetuning [21, 27] after training and for meta-learning [28], very few methods replay whole images [52] or stored feature representations [1] during training. For instance segmentation, [18] explored copying random instances from one image to another. Our work deals with bounding box replay methods to better address the challenges of IOD.

3. Proposed Method

3.1. Problem Formulation and Overview

Object detection is primarily concerned with accurately identifying and localizing objects of interest within an image. Given a set of data $D = \{(I_n, Y_n)\}_{n=1}^N$, an ideal object detector $f_\theta(I_n)$ can predict a series of boxes $\hat{Y}_n$ corresponding to the groundtruth Y_n, where $Y_n = \{y_g = (u_g, v_g, w_g, h_g, c_g)\}_{g=1}^{G_n}$, with (u_g, v_g) denoting the top-left corner coordinates of the bounding box and (w_g, h_g) the width and height of the bounding box, and c_i denotes the class for each of the G_n bounding boxes. Therefore, D has $K = \sum_{n=1}^N G_n$ groundtruth boxes totally. This work focuses on two-stage detectors from the R-CNN family [19, 23, 51] that typically consist of a CNN-based feature extractor, a Region Proposal Network (RPN), and a class-level classi-

cation and bounding box regression network (RCN).

Incremental object detection aims to learn to detect objects in a sequence of T tasks, where each task $D^t = \{(I_n^t, Y_n^t)\}_{n=1}^{N^t}$ corresponds to a new set of classes $\mathcal{C}^t$. The model should be able to detect objects in the new classes $\mathcal{C}^t$ while retaining the ability to detect objects in the previously seen classes $\mathcal{C}^{1:t-1}$ without catastrophic forgetting. However, unlike in the classification tasks where each input has a single label, I_n^t may contain objects from both $\mathcal{C}^t$ and $\mathcal{C}^{1:t-1}$, and the annotations Y_n^t only include the bounding boxes and class labels for $\mathcal{C}^t$. Therefore, $G_n^t \leq$ the number of the real annotations in IOD. The presence of unlabeled previous objects can lead to **Background Shift** during training, where attention of the detector is biased towards the $\mathcal{C}^t$ and it fails to differentiate between the objects from $\mathcal{C}^t$ and $\mathcal{C}^{1:t-1}$. Moreover, misassociations can propagate over time, exacerbating catastrophic forgetting of previous classes.

A straightforward way to a solution is using the original images from $D^{1:t-1}$, as shown in Fig. 1, which provides certain information for $\mathcal{C}^{1:t-1}$. However, the image replay method involves replaying original images from the previous training set during the current one, which can cause **Foreground Shift** due to replay of unlabeled objects from $\mathcal{C}^t$. Thus, the new classes or the foreground in the current images are considered as the background in the replayed images which results in the model failing to generalize to new contexts. Additionally, storing the original images can result in significant memory overhead, since they include a lot of redundant information.

3.2. Augmented Box Replay

To mitigate the foreground shift problem, we propose an Augmented Box Replay (ABR) strategy that selects a subset of informative and representative box images from the

previous task, along with a new set of boxes for the current task t. This method avoids replaying redundant information and optimally employs its storage for the relevant object regions. Specifically, ABR can replay these boxes in an augmented way, which helps the model retain its ability to detect previous objects in new contexts while improving its detection performance on the current task. Fig. 3 illustrates the pipeline of Augmented Box Replay strategy.

At the beginning, we involve a prototype box selection to choose the most representative boxes whose feature maps are close to the mean feature map after training of task $t-1$. The memory buffer is denoted as B^{t-1}, where the memory size M^{t-1} of B^{t-1} is limited. Therefore, the selection is an important factor that affects the performance. The final B^{t-1} can focus on the most relevant information and avoid redundant or irrelevant information. Since box images are smaller than images, the storage cost is reduced, making it scalable to large datasets and complex models. See supplementary material for more details.

To leverage prototype boxes B^{t-1} accumulated from the previous tasks in the current task t, we have designed two types of replay strategies: mixup box replay and mosaic box replay, inspired by [4, 64]. These strategies allow us to effectively transfer knowledge from past tasks to the current one and enhance the performance of the model.

MixUp box replay. This method replays the box images of the previous class in the current data, placed in such a way that the previous box objects blend into the image more naturally, while ensuring that they have minimal overlap with the groundtruth bounding boxes of the new class. It involves assigning a random location in the current image I_n^t to each box image $b \in B^{t-1}$ with size (w_b, h_b), and then mixing it with I_n^t to create a new image $\hat{I}_n^t$. More specifically, $\hat{I}_n^t$ is obtained by overlaying b onto I_n^t at a location with a mixing coefficient λ. For each pixel location in $\hat{I}_n^t$, if (u, v) is not inside the box, then the original pixel value of I_n^t is retained. If (u, v) is inside the box, the mixed pixel value is computed by:

$$\hat{I}_n^t(u,v) = \begin{cases} \lambda I_n^t(u,v) \\ +(1-\lambda)b(\hat{u},\hat{v}), & \text{if } \max_{g \in G_t^n} y_g \cup b \leq th \\ I_n^t(u,v) & \text{otherwise} \end{cases},$$
(1)

where λ is values with the [0, 1] range and is sampled from the Beta distribution [64], $b(\hat{u}, \hat{v})$ is the pixel value of the box image b at location $(\hat{u} = u - w_b, \hat{v} = v - h_b)$, $y_g \cup b$ is the intersection over union (IOU) between each groundtruth annotation $y_g, \forall g \in G_n^t$ and the box image b, and th is a threshold value. If the maximum IOU over union between the groundtruth annotations and the box image $b \leq th$, then the pixel value at (u, v) in the new image $\hat{I}_n^t$ is a mixture of the original pixel value $I_n^t(u, v)$ and the corresponding pixel value in the box image b. Otherwise, the original pixel

Table 1: Influence of different detector components in Faster-ILOD [47] on VOC 10-10 setting.

Frozen Backbone	Feature Distil.	RPN Distil.	RCN Distil.	VOC (10-10)		
				1-10	11-20	1-20
	✓	✓	✓	70.3	53.0	61.7
✓		✓	✓	70.7	53.3	62.0
		✓	✓	70.6	53.7	62.2
	✓		✓	69.8	53.3	61.6
	✓	✓		8.2	62.7	35.5

value $I_n^t(u, v)$ is retained. Note that at most two boxes are mixed up in a single image I_n^t since the boxes are selected randomly and the overlap condition limits the number of boxes that can be mixed up in a single image.

Mosaic box replay. This method involves dividing I_n^t into a grid and randomly selecting a subset of cells. Each cell is then replaced with a box image b from B^{t-1}, and the resulting image $\hat{I}_n^t$ is used for rehearsal. In the mosaic box replay strategy, a composite image is formed by combining four box images into a single mosaic image. To create the composite image, a random location is first selected as the center point of the mosaic image. Then, each of the four boxes is resized to a new size that is proportional to the size of the mosaic image, with the scaling factor μ randomly sampled from the range of [0.4, 0.6]. The resized boxes are arranged in the four quadrants of the mosaic image, and the remaining areas are filled with a fixed color value.

In summary, the Augmented Box Replay offers several advantages for incremental learning in object detection: 1) **Information Richness:** ABR selects the most informative and representative boxes for rehearsal, which preserves the accuracy and diversity of learned model. 2) **Enhanced generalization:** ABR serves as an augmentation method which gives a different background context to both previous and new classes and thus improves the generalization of the model. 3) **Memory efficiency:** ABR replays only a small set of representative box images instead of the entire images, which significantly reduces the memory requirement. 4) **Adaptability:** ABR can easily be integrated with different object detection models to improve their performance.

3.3. Attentive RoI Distillation

Distillation-based methods [7,47,54] are commonly used in IOD, aiming to transfer the knowledge of a model trained on a previous task (teacher) to a current model (student) while simultaneously learning the new task. To further explore the impact of the distillation operation on the forgetting of each detector component, an ablation study is evaluated on the Faster-ILOD model [47] as shown in Table 1. We can find that the feature extractor has a minimal effect on forgetting when either freezing the backbone or applying the feature distillation operation, and the presence or absence of the RPN component only has a 0.1% effect

on forgetting. However, removing the distillation operation of the prediction head (RCN) results in a 26.2% drop in performance. Our obtained analysis and [57] together suggest that forgetting mainly occurs at the classification head. However, a limitation of RPN distillation lies in its focus solely on extracting RPN modules, which provide region proposals without considering features within each proposal. Consequently, the distilled model may overlook informative features within the proposals, leading to sub-optimal performance. To address this, we propose the Attentive RoI Distillation (ARD) loss, which allows the student model to selectively focus on the most important features from the teacher model by aligning the spatial attention maps and masked features of each proposal. Moreover, ARD supports more inclusive RoI features for the final prediction and helps to overcome the forgetting problem in the classification head.

To enable the model to focus on the most informative parts of an image, we calculate a spatial attention map A_i^t for each $F_i^t, \forall i \in P_n^t$, where P_n^t is the number of proposals. The spatial attention map is obtained by raising the absolute value of each feature plane $F_{i,d}^t$ to a power p (in the experiments, $p = 2$) based on [62] and summing them up:

$$A_i^t = \sum_{d=1}^{C} \left| F_{i,d}^t \right|^p, \quad p > 0, \tag{2}$$

Our method employs spatial attention maps from previous and current models to emphasize the most informative features and suppress the less informative ones. More superficially, the pooled attention distillation (PAD) loss is:

$$\mathcal{L}_{PAD} = \left\| A_i^{t-1} - A_i^t \right\|, \tag{3}$$

where A_i^{t-1} and A_i^t are the spatial attention maps for the ith proposal in the previous and current models, respectively. PAD can transfer knowledge from a previously trained model to a new one in a progressive learning setting. The key difference with existing distillation methods in IOD is that here we explicitly distill the knowledge on the location of the relevant features (this is encoded in the attention map). Furthermore, ours applies the attentive distillation into the aligned bounding boxes, which contain the very relevant both location and feature information. Specifically, the Attentive RoI Feature Distillation (AFD) loss is employed:

$$\mathcal{L}_{AFD} = \frac{1}{P_n} \sum_{i=1}^{P_n} \left(F_i^{t-1} - F_i^t \right)^2 A_i^{t-1}, \tag{4}$$

where P_n^t is the number of proposals for I_n^t, F_i^{t-1} and F_i^t are the features extracted from the previous and new models, respectively. The squared difference $(F_i^{t-1} - F_i^t)^2$ penalizes larger deviations between the previous and new features, which further encourages the new model to reproduce

informative features from the previous model. By using the attention maps to weight the MSE term, AFD ensures new model to focus on reproducing the most important features from the previous model, while allowing for some flexibility in reproducing the less informative features. The overall ARD loss function is defined as:

$$\mathcal{L}_{ARD} = \mathcal{L}_{AFD} + \gamma \mathcal{L}_{PAD} \tag{5}$$

where γ is a hyperparameter that controls the strength of the regularization.

ARD loss not only aligns the features of each proposal but also has an effect on the position deviation of each anchor point. This spatial attention feature alignment reduces the impact of background shift caused by the imbalance between new and previous classes and promotes knowledge transfer from the previous model to the new model.

3.4. Inclusive Loss with Background Constraint

To avoid forgetting in classification head, we followed the unbiased classification and distillation losses proposed by [7,8]. However, due to our Augmented Box Replay strategy, the input image $\hat{I}_n^t$ contains many annotations about previous objects. This means that using unbiased losses directly in this situation is not feasible, as it would ignore the positive influence of the B^{t-1} on the previous categories during the training phase. Therefore, we involve Inclusive Loss with Background Constraint to adapt the ABR based on the unbiased classification and distillation losses. In detail, the Inclusive Classification Loss is defined as follows:

$$\mathcal{L}_{IC} = \frac{1}{P_n^t} \sum_{i=1}^{P_n^t} c_i \begin{cases} \log(p_i^b + \sum\limits_{c=1}^{\mathcal{C}^{1:t-1}} p_i^c), & c_i = \mathcal{C}^b \\ \sum\limits_{c=1}^{\mathcal{C}^{1:t}} c_i \log p_i^c, & c_i \in \mathcal{C}^{1:t} \end{cases} \tag{6}$$

where c_i is the label of proposal i, p_i^b is the probability as $\mathcal{C}^b$, p_i^c is the probability as $\mathcal{C}^t$. For positive RoI of $\mathcal{C}^{1:t}$ in ABR, the standard RCN loss based on cross-entropy is maintained. However, for negative RoI, the sum of the probabilities of $\mathcal{C}^{1:t-1}$ is treated as $\mathcal{C}^b$, ensuring that the model does not learn to predict $\mathcal{C}^{1:t-1}$ for unlabeled objects.

Moreover, the inclusive distillation (ID) loss maintains the performance of task $t-1$ by aligning the probabilities of the previous model for the background class with the probabilities of the new model for both $\mathcal{C}^b$ and $\mathcal{C}^t$. The training data for ABR includes grountruth annotations from box rehearsal, and the teacher model can detect previous objects. Therefore, we only need to focus on each proposal of $\mathcal{C}^t$:

$$\mathcal{L}_{ID} = \frac{1}{\Omega} \begin{cases} p_i^{b,t-1} \log(p_i^{b,t} + \sum\limits_{c=1}^{\mathcal{C}^t} p_i^{c,t}), & c_i = \mathcal{C}^b \\ \sum\limits_{c=1}^{\mathcal{C}^{t-1}} p_i^{c,t-1} \log(p_i^{c,t}), & c_i \in \mathcal{C}^{1:t} \end{cases} \tag{7}$$

where $\Omega = |\mathcal{C}^{1:t-1}| + 1$ is the number of previous and background classes, $p_i^{b,t-1}$ and $p_i^{c,t-1}$ are the classification probabilities of the background class and previous classes in task $t-1$, respectively, $p_i^{b,t}$ and $p_i^{c,t}$ are the classification probabilities of the background class and new classes in task t, respectively, for the proposal i, $p_i^{c,t}$ is the classification probability of previous classes and new classes for the proposal i in the current task t.

4. Experiments

4.1. Experimental Settings

Datasets: We evaluate the proposed method on two publicly available datasets namely PASCAL VOC 2007 [16] and MS COCO 2017 [36]. PASCAL VOC 2007 contains 20 object classes and 9,963 images, 50% of which is used for training and validation and the remaining 50% for testing following [16]. MS COCO 2017, as a challenging dataset, has 80 different object classes and provides 83,000 images for training, 40,000 for validation and 41,000 for testing.

IOD Protocols: Following previous works on this topic [7, 28, 54], we obey the same experimental protocols. Each training task contains all images which have at least one bounding box from a new class. The annotations are available only for the new classes while the previous and future classes are not annotated. This setting is practical and can also have repetitions of images across tasks.

Implementation Details: Similar to [7, 9, 22, 28, 39, 47, 54, 59, 65], we use the Faster R-CNN [51] architecture with a Resnet-50 [24] backbone pretrained on ImageNet [11]. We train the network with SGD optimizer, momentum of 0.9 and weight decay of 10^{-4}. We use a learning rate of 5 $\times 10^{-3}$ for the initial task and 2×10^{-3} for the subsequent tasks. We used 15K iterations for 5 or 10 class increments in a task and 5K iterations when adding 1 or 2 new classes. We set the memory size as 2,000 for all the experiments on PASCAL VOC 2007, 10,000 for 70-10 and 5,000 for 40-40 settings on MS COCO 2017 respectively. Our method uses a stack to store boxes, which are randomly selected and placed (while considering overlap criteria) during each iteration. To balance the number of old and new objects, we determine the 1:1:2 ratio for mixup, mosaic, and new images based on comparisons across different settings.

Evaluation: We evaluate the methods in terms of mean average precision at 0.5 IoU threshold for PASCAL VOC 2007. For MS COCO 2017, we also report the mAP at different IoU ranging from 0.5 to 0.95 IoU (mAP@[50:95]), at 0.50 IoU (mAP@50) and at 0.75 IoU (mAP@75).

4.2. Quantitative Evaluation

Following previous works [7, 28, 47, 54, 59, 65], we evaluate our method on settings with different number of initial classes and one or more incremental tasks. We compare our method with two baselines, the Fine-tuning when the model is trained with the data incrementally without any regularization or data replay, and the Joint training when the model is trained on the entire dataset with all the annotations. All results are obtained after training of the last task.

4.2.1 PASCAL VOC 2007

For PASCAL-VOC 2007, we perform our experiments on 19-1, 15-5, 10-10 and 5-15 single incremental task settings adding 1, 5, 10, 15 classes respectively. For multi-step incremental settings, we evaluate on 10-5, 5-5, 10-2, 15-1 and 10-1 settings where we add 5, 5, 2, 1 and 1 classes respectively at every step till all the 20 classes are seen.

Single-step increments: We benchmark our ABR method against the existing methods on Table 2. We notice that Fine-tuning suffers from catastrophic forgetting across all settings. ABR outperforms all other methods across all the settings, significantly improving over MMA on the new classes by 4.5 mAP on 15-5, 8.9 mAP on 10-10 and 9.8 mAP on 5-15. We argue that the enhanced stability and plasticity is due to the augmented box replay of previous classes and our effective attention distillation. Our improvements over the methods storing exemplars [21, 27, 28] confirm the importance of the box replay for IOD.

Multi-step increments: The catastrophic forgetting and the background shift problem is more crucial on the longer incremental settings as seen in the performance from Table 3. Fine-tuning suffers from almost complete forgetting on the initial classes. ABR improves over the closest competitor MMA by 3.9 mAP on 10-5, 3.5 mAP on 10-2, 3.4 mAP on 15-1 and 7.7 mAP on the longest and most challenging setting 10-1. It is interesting to observe that most methods struggle on the 5-5 setting with only 5 initial classes while ABR improves over MMA by 19.5 mAP. This implies that the existing methods require more classes in the initial task to achieve better generalization and thus, fails to adapt to new classes when the first task has lesser classes in 5-5 setting. On the most difficult setting of 10-1 with 10 increments, ABR outperforms MMA by 4.1 mAP on the previous classes and 11.1 mAP on the new classes. Note that for multiple increment settings, the improvement in the performance of incremental classes is not only due to better learning of new classes but also due to lesser forgetting of the intermediate task classes after moving to new tasks.

4.2.2 MS COCO 2017

For MS COCO 2017, we perform experiments on 40-40 and 70-10 settings adding 40 and 10 classes respectively. As shown in Table 4, Fine-tuning suffers from catastrophic forgetting on both settings. While Faster ILOD and MMA has improved over Fine-tuning, our method improves average mAP@[50:95] over MMA by 1.5 on 40-40 setting and by

Table 2: mAP@0.5% results on settings with single increments on Pascal-VOC 2007. Best among columns in **bold** and second best among columns are underlined. Methods with * use exemplars. †: results from re-implementation.

#Method	19-1			15-5			10-10			5-15		
	1-19	20	1-20	1-15	16-20	1-20	1-10	11-20	1-20	1-5	6-20	1-20
Joint Training	70.1	75.7	74.3	76.4	67.8	74.3	75.5	73.0	74.3	70.1	75.7	74.3
Fine-tuning	11.8	64.7	14.4	15.9	54.2	25.5	2.6	63.4	32.9	6.9	63.1	49.1
ILOD (FasterRCNN)† [54]	69.8	64.5	69.6	72.5	58.5	68.9	69.8	53.7	61.7	61.0	37.3	43.2
Faster ILOD† [47]	70.9	63.2	70.6	**73.1**	57.3	69.2	70.3	53.0	61.7	62.0	37.1	43.3
PPAS [65]	70.5	53.0	69.2	-	-	-	63.5	60.0	61.8	-	-	-
MVC [59]	70.2	60.6	69.7	69.4	57.9	66.5	66.2	66.0	66.1	-	-	-
MMA† [7]	70.9	62.9	70.5	72.7	60.6	69.7	69.8	63.9	66.8	**66.8**	57.2	59.6
ORE* [27]	69.4	60.1	68.9	71.8	58.7	68.5	60.4	68.8	64.6	-	-	-
OW-DETR* [21]	70.2	62.0	69.8	72.2	59.8	69.1	63.5	67.9	65.7	-	-	-
Meta-ILOD* [28]	70.9	57.6	70.2	71.7	55.9	67.8	68.4	64.3	66.3	-	-	-
ABR (Ours)	**71.0**	**69.7**	**70.9**	73.0	**65.1**	**71.0**	**71.2**	**72.8**	**72.0**	64.7	**71.0**	**69.4**

Table 3: mAP@0.5% results on settings with multiple increments on Pascal-VOC 2007. Best among columns in **bold** and second best among columns are underlined. †: results from re-implementation.

#Method	10-5 (3 tasks)			5-5 (4 tasks)			10-2 (6 tasks)			15-1 (6 tasks)			10-1 (11 tasks)		
	1-10	11-20	1-20	1-5	6-20	1-20	1-10	11-20	1-20	1-15	16-20	1-20	1-10	11-20	1-20
Joint Training	75.5	73.0	74.3	70.1	75.7	74.3	75.5	73.0	74.3	76.4	67.8	74.3	75.5	73.0	74.3
Fine-tuning	5.3	30.6	18.0	0.5	18.3	13.8	3.79	13.6	8.7	0.0	10.47	5.3	0.0	5.1	2.55
ILOD (FasterRCNN)† [54]	67.2	59.4	63.3	58.5	15.6	26.3	62.1	49.8	55.9	65.6	47.6	60.2	52.9	41.5	47.2
Faster ILOD† [47]	68.3	57.9	63.1	55.7	16.0	25.9	64.2	48.6	56.4	66.9	44.5	61.3	53.5	41.0	47.3
MMA† [7]	67.4	60.5	64.0	62.3	31.2	38.9	65.7	52.5	59.1	67.2	47.8	62.3	57.9	44.6	51.2
ABR (Ours)	68.7	67.1	67.9	64.7	56.4	58.4	67.0	58.1	62.6	68.7	56.7	65.7	62.0	55.7	58.9

Table 4: mAP results on MS COCO 2014 at different IoU, where the best among columns in **bold**.

#Method	40-40 mAP@			70-10 mAP@		
	[50 : 95]	50	75	[50 : 95]	50	75
Joint Training	35.9	60.5	38.0	35.9	60.5	38.0
Fine-tuning	19.0	31.2	20.4	5.6	8.6	6.2
Faster ILOD [47]	20.6	40.1	-	21.3	39.9	-
MMA [7]	33.0	56.6	34.6	30.2	52.1	31.5
ABR (Ours)	**34.5**	**57.8**	**35.2**	**31.1**	**52.9**	**32.7**

0.9 on 70-10 setting. These results signify lesser forgetting and better adaptation to new classes with our method.

4.3. Analysis and Ablation Study

We investigate the role of the network components, replay selection strategies, augmentation types in Table 5 on the VOC 10-10 and 10-5 settings. We take the baseline model with the RCN classification and distillation loss proposed by [7]. We show that our attentive RoI distillation improves over the RPN distillation used by [7, 47] owing to better exploitation of location and feature information of the RoIs. In replay strategies, we implemented the herding strategy [49] for selecting boxes to replay. Our method improves 1%~1.5% mAP over the herding strategy. We can observe that our proposed prototype box selection can better capture more representative prototype samples for previous classes. Further, we add mixup and mosaic replay individually and observe that both strategies improve the perfor-

mance on previous and new classes. The best performance is achieved when both mixup and mosaic replay are performed with the new images.

We investigate the role of the memory size and train ABR with different memory size of previous class boxes. Fig. 4 plots the mAP@50 results with increasing memory size. It is observed that the performance increases with increasing memory size or replay of more previous objects. It can be observed that after the memory size > 2000, the growth rate of mAP tends to be more stable. Therefore, in the main experiments, we use a memory size of 2000.

Table 6 presents a comparison between image replay and our proposed ABR method. The same number of objects ensures that the original information about the previous categories stored in the memory buffer is consistent, and the same storage space controls practicality in real-world applications. As shown in Table 6, despite having the same number of objects, image replay performs worse than augmented box replay in recognizing new classes. This confirms that replaying original images can lead to foreground shifts and limit the adaptation of new classes. On the other hand, our memory buffer contains about 4 times as many original objects for previous classes as image replay.

4.4. Visualization

Fig. 5 shows some examples of images generated by mixup replay in VOC 10-10 setting. It can be seen intu-

Table 5: Ablation study highlighting contribution from different components, where the best among columns in **bold**.

RCN $\mathcal{L}_{IC},\mathcal{L}_{ID}$	RPN Distil.	RoI $\mathcal{L}_{ARD}$	PBS	Selection Herding	Random	AugmentedType Mixup	Mosaic	VOC (10-10) 1-10	11-20	1-20	VOC (10-5) 1-10	11-15	16-20	1-20
✓								43.5	75.9	59.4	65.1	31.3	59.8	55.3
✓	✓							45.2	75.6	60.4	67.1	30.5	59.3	55.9
✓		✓						47.9	**76.2**	62.0	67.0	35.6	58.4	57.0
✓		✓	✓			✓		68.9	72.6	70.7	67.4	72.8	**63.5**	67.7
✓		✓	✓				✓	70.6	71.2	70.9	67.0	70.7	61.8	66.6
✓	✓		✓			✓	✓	69.7	72.4	71.0	67.4	**72.9**	61.1	67.2
✓		✓			✓	✓	✓	68.7	71.5	70.1	67.0	71.2	62.8	67.0
✓		✓		✓		✓	✓	69.4	71.6	70.5	67.4	72.3	61.1	67.2
✓		✓	✓			✓	✓	**71.2**	72.8	**72.0**	**68.7**	71.5	62.8	**67.9**

Table 6: Rehearsal alternative on Pascal VOC 2007 in mAP@50. All experiments are done in our proposed method with image replay or augmented box replay (ABR).

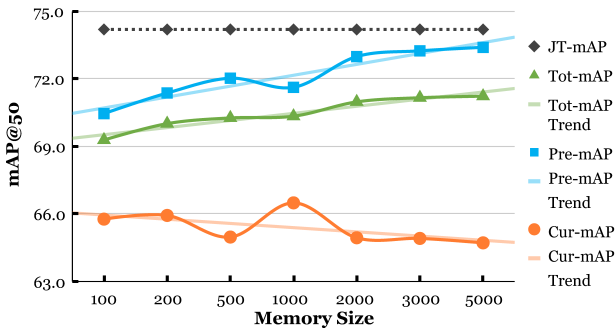| Type	Buffer Size	Objects	Memory↓	VOC (10-10) 1-10	11-20	1-20
-	-	-	-	47.9	76.2	62.0
Image	182	455	15.5Mb	70.2	62.2	66.2
Image	800	2000	68Mb	71.6	57.9	64.7
ABR	2000	2000	15.5Mb	71.2	72.3	72.0

Figure 4: The average mAP@50 of previous, current and total classes in terms of different memory sizes at PASCAL VOC 2007 15-5 setting.

Figure 5: Examples of images generated by mixup augmentation for 10-10 setting on PASCAL VOC 2007. Blue boxes represent previous classes which are replayed in the background of new images. Orange boxes represent the ground truth annotations of current classes.

itively that the mixup strategy makes the box reasonably integrated into the new images and minimizes the occlusion with the new objects. In addition, the background information compared to the new objects is greatly enhanced. The inference results are available in supplementary material.

5. Conclusion

In this paper, we studied the experience replay method for incremental object detection problem and introduced the critical issue of foreground drift during old image replay. We hypothesize that the foreground drift is the reason that replay methods, which are dominant in incremental learning for image classification, have been little studied for IOD.

To tackle this problem, our proposed method ABR stores bounding boxes from old classes and replays them with new images using mixup and mosaic augmentation strategies. This overcomes the foreground drift situation since only the old classes are stored and replayed and not the unlabeled new classes from old images. In addition to box replay, the proposed attentive RoI distillation uses both the location and feature information for the RoIs extracted from the RPN and enables retention of meaningful knowledge of old classes. Further, our method reduces the memory overhead significantly. We demonstrate that ABR outperforms existing methods across all settings on representative datasets.

This work lays the foundation for bounding box replay instead of the traditional image or feature replay methods for object detection tasks. Future research should explore the implications of the foreground shift in incremental semantic segmentation and extend our approach to popular transformer methods [41].

References

[1] Manoj Acharya, Tyler L Hayes, and Christopher Kanan. Rodeo: Replay for online object detection. In *British Machine Vision Conference (BMVC)*, 2020. 3

[2] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[3] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 4

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[6] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[7] Fabio Cermelli, Antonino Geraci, Dario Fontanel, and Barbara Caputo. Modeling missing annotations for incremental learning in object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 5, 6, 7

[8] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 5

[9] Li Chen, Chunyan Yu, and Lvcai Chen. A new knowledge distillation for incremental object detection. In *International Joint Conference on Neural Networks (IJCNN)*, 2019. 2, 6

[10] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 1, 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6

[12] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[13] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[14] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[15] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 2009. 6

[17] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[18] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[19] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3

[20] Dipam Goswami, René Schuster, Joost van de Weijer, and Didier Stricker. Attribution-aware weight transfer: A warm-start initialization for class-incremental semantic segmentation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2

[21] Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6, 7

[22] Yu Hao, Yanwei Fu, Yu-Gang Jiang, and Qi Tian. An end-to-end architecture for class-incremental object detection with knowledge distillation. In *International Conference on Multimedia and Expo (ICME)*, 2019. 2, 3, 6

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[25] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[26] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[27] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6, 7

[28] KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 3, 6, 7

[29] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[30] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017. 2, 3

[31] Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry Heck. Rilod: Near real-time incremental learning for object detection at the edge. In *Symposium on Edge Computing*, 2019. 3

[32] Mingxue Li, Yang Cong, Yuyang Liu, and Gan Sun. Class-incremental gesture recognition learning with out-of-distribution detection. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 2

[33] Zhizhong Li and Derek Hoiem. Learning without forgetting. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 2

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 6

[37] 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. *Transactions on Neural Networks and Learning Systems (TNLSS)*, 2020. 3

[38] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[39] Xialei Liu, Hao Yang, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Multi-task incremental learning for object detection. *arXiv preprint arXiv:2002.05347*, 2020. 3, 6

[40] Yuyang Liu, Yang Cong, Gan Sun, Tao Zhang, Jiahua Dong, and Hongsen Liu. L3doc: Lifelong 3d object classification. *IEEE Transactions on Image Processing*, 2021. 2

[41] Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[42] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[43] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[44] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[45] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 1, 2

[46] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[47] Can Peng, Kun Zhao, and Brian C Lovell. Faster ilod: Incremental learning for object detectors based on faster rcnn. *Pattern recognition letters*, 2020. 2, 3, 4, 6, 7

[48] Can Peng, Kun Zhao, Sam Maksoud, Tianren Wang, and Brian C Lovell. Diode: Dilatable incremental object detection. *Pattern Recognition*, 2023. 3

[49] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 7

[50] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015. 1, 2, 3, 6

[52] Jeng-Lun Shieh, Qazi Mazhar ul Haq, Muhamad Amirul Haq, Said Karam, Peter Chondro, De-Qin Gao, and Shanq-Jang Ruan. Continual learning strategy in one-stage object detection framework based on experience replay for autonomous driving vehicle. *Sensors*, 2020. 3

[53] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[54] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4, 6, 7

[55] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[56] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[57] Eli Verwimp, Kuo Yang, Sarah Parisot, Hong Lanqing, Steven McDonagh, Eduardo Pérez-Pellitero, Matthias De Lange, and Tinne Tuytelaars. Re-examining distillation for continual object detection. *arXiv preprint arXiv:2204.01407*, 2022. 5

[58] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[59] Dongbao Yang, Yu Zhou, Aoting Zhang, Xurui Sun, Dayan Wu, Weiping Wang, and Qixiang Ye. Multi-view correlation distillation for incremental object detection. *Pattern Recognition*, 2022. 3, 6, 7

[60] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[61] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[62] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[63] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[64] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 4

[65] Wang Zhou, Shiyu Chang, Norma Sosa, Hendrik Hamann, and David Cox. Lifelong object detection. *arXiv preprint arXiv:2009.01129*, 2020. 3, 6, 7