ESP-YOLO: STABILIZING INCREMENTAL OBJECT DETECTION VIA EIGENVALUE SCALED PROJECTION REGULARIZATION AND CONFIDENCE-AWARE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Incremental Object Detection (IOD) requires sequentially learning to localize and classify novel object categories while preserving the stability of historical representations to mitigate catastrophic forgetting. While single-stage frameworks like YOLO are ideal baselines for real-world IOD deployment due to their efficiency, they remain under-explored. This gap primarily stems from YOLO's end-to-end prediction mechanism, where backpropagated gradients from new tasks cause drastic perturbations to backbone parameters. Unlike two-stage models that inherently isolate such impacts via class-agnostic region proposal networks, YOLO suffers from significantly more severe intermediate feature drift and exacerbated class confusion. To overcome these architectural bottlenecks, we propose a framework featuring Eigenvalue Scaled Projection Regularization (ESPReg) and Confidence-Aware Distillation (CAD). Specifically, ESPReg utilizes the eigenvalues of feature covariance to quantify the importance of spatial directions, guiding weight updates to minimize interference with old tasks. Furthermore, CAD leverages high-confidence regions to guide the distillation process, isolating background interference to resolve severe class confusion. Extensive experiments across single-stage and multi-stage settings demonstrate that our approach surpasses state-of-the-art IOD methods by 1.4% and 7.3% on the PASCAL VOC and MS COCO benchmarks, respectively.

1 INTRODUCTION

Unlike human cognition, which can accumulate knowledge incrementally, deep neural networks typically suffer from catastrophic forgetting when exposed to non-stationary data distributions. Consequently, Incremental Learning (IL) (Wang et al., 2024) is proposed to enable models to sequentially learn new tasks without compromising the proficiency acquired from historical data. While IL has garnered widespread attention in image classification, Incremental Object Detection (IOD) holds significantly greater practical application value, standing as a critical yet challenging frontier. Unlike classification, object detection requires the model to simultaneously categorize and localize objects of interest. Consequently, IOD must preserve both the discriminative knowledge and the precise localization capabilities for historical categories, which imposes more stringent requirements on preserving model stability for previously learned tasks.

Recent studies (Farajtabar et al., 2020; Wang et al., 2021; Yang et al., 2025; Zhu et al., 2025) find that by projecting the gradients of weight to the null space of previous tasks' features, models can preserve their performance on previous trained datasets after being trained on new dataset. This strategy is referred to as Null-Space Gradient Projection (NSGP) (Wu et al., 2025). However, there has been very limited attempts to apply NSGP to IOD. Only recently, the state-of-the-art IOD method NSGP-RePRE (Wu et al., 2025) introduces NSGP into IOD setting. However, existing methods naively apply NSGP to two-stage object detection models. Given that object detection is an inherently application-driven task, it imposes stringent demands on both efficiency and effectiveness. Consequently, single-stage frameworks like YOLO (Redmon et al., 2016) serve as ideal baselines for real-world IOD. We identify two critical bottlenecks in applying NSGP to YOLO, as detailed below.

First, single-stage detectors suffer from severe intermediate feature drift. Unlike two-stage models (e.g., Faster R-CNN (Ren et al., 2016)) that inherently isolate the backbone from drastic semantic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shifts via a class-agnostic Region Proposal Network (RPN), YOLO relies on an end-to-end dense prediction mechanism. Consequently, backpropagated gradients from novel tasks cause drastic perturbations to the backbone parameters. Traditional NSGP has been empirically proven to be less effective on YOLO-based incremental detectors.

Second, the lack of architectural decoupling in YOLO leads to cross-task class confusion. In a class-incremental setting, the ground truth frequencies of old and novel classes are naturally imbalanced. Therefore, when the YOLO-based detector is incrementally trained on both old and novel classes, it has a pronounced tendency to misclassify old classes into visually similar novel classes (e.g., cow→sheep, bus→train). This leads to low performance in a task-agnostic inference scenario.

To overcome these architectural bottlenecks, we propose a novel framework, ESP-YOLO, to stabilize single-stage IOD via Eigenvalue Scaled Projection Regularization (ESPReg) and Confidence-Aware Distillation (CAD). ESPReg resolves the intermediate feature drift with an adaptive regularization loss. Our approach is grounded in the theoretical analysis of the feature drift risk, in which we identify that the risk's upper bound is contributed by the projection of parameter updates onto each principal components of historical features, scaled by corresponding feature covariance eigenvalues. Based on this key insight, ESPReg replaces rigid gradient projection in NSGP (Wu et al., 2025) with a soft, adaptive regularization in loss function. Specifically, following the training of each task, we apply Principal Component Analysis (PCA) (Pearson, 1901) to the input feature space. Subsequent parameter updates are projected onto these principal components, and the norms of these projections are weighted by their respective eigenvalues and summed to formulate a regularization term. By converting hard constraints into soft, eigenvalue-scaled energy barriers, ESPReg permits updates within the significant subspace when task-specific gradients outweigh the projection penalty. This mechanism effectively bounding weight deviation in critical directions to prevent feature drift.

Furthermore, minimizing feature drift through ESPReg alone is insufficient to fully resolve the challenges inherent in YOLO-based IOD. To address the challenge of cross-task class confusion, we introduce the CAD mechanism. Unlike standard approaches that indiscriminately distill all output channels, thereby risking the transfer of task-irrelevant background noise, CAD selectively transfers knowledge guided by the base model's confidence. By generating a confidence mask based on the maximum classification scores across classes, CAD compels the model to anchor its predictions on regions where the base model exhibits high certainty. This targeted strategy effectively isolates background interference and preserves discriminative capabilities for historical categories, ensuring the stability of YOLO in IOD. The contributions of our work are summarized as:

- To address the severe intermediate feature drift in single-stage incremental detectors, we propose ESPReg. Grounded in a theoretical analysis of the feature drift risk, ESPReg adaptively penalizes parameter updates along critical historical feature directions based on their covariance eigenvalues. This mechanism effectively preserves structural stability while maintaining the necessary plasticity for learning novel tasks.
- To mitigate the cross-task class confusion in IOD, we introduce CAD. Instead of indiscriminate global distillation, CAD anchors the knowledge transfer exclusively on high-certainty regions, shielding the IOD process from background noise and ensuring semantic discrimination for incremental classes.
- Extensive experiments on Pascal VOC demonstrate the effectiveness of ESPReg across various baseline models. Specifically, it achieves a new state-of-the-art result on the 15+5 class incremental setting of Pascal VOC and 40+40 class incremental setting of MS COCO by 1.4% and 7.1% improvement over the NSGP-RePRE respectively.

2 Related Work

Incremental Learning (IL). IL aims to sequentially learn from novel data without catastrophically forgetting prior knowledge. In visual tasks, IL methods are primarily researched under two settings: class-incremental learning (accommodating new categories) and domain-incremental learning (adapting to distribution shifts) (Wang et al., 2024). Although Incremental Image Classification (IIC) is primarily focused (Wang et al., 2024), Incremental Object Detection (IOD) has also gained popularity in recent years (Menezes et al., 2023).

Incremental Object Detection (IOD). IOD introduces unique challenges beyond image classification, such as the coupled nature of localization and classification and the spatial sensitivity of features. The field was largely formalized by (Shmelkov et al., 2017) who introduced a distillation-based paradigm.

Subsequently, both distillation-based (Shmelkov et al., 2017; Hao et al., 2019; Liu et al., 2023a; Dong et al., 2023; Zhang et al., 2025) and replay-based (Shieh et al., 2020; Yang et al., 2023a; Liu et al., 2023b) methods have seen extensive study. However, the significant limitations of dominant paradigms persist: Given the effeciency demands of deployed object detection systems, dynamic architecture-based methods generally fall short because of its constantly expanding model size. Furthermore, replay-based strategies suffer from privacy concerns and memory burdens. Therefore, we direct our focus toward regularization-based methods, as they offer a more promising and practically viable paradigm for IOD.

Null-Space Gradient Projection (NSGP). NSGP (Farajtabar et al., 2020; Saha et al., 2021; Lin et al., 2022; Zhao et al., 2023; Yang et al., 2025; Zhu et al., 2025), a family of regularization-based methods, is steadily emerging as a promising paradigm. It mitigates forgetting by projecting gradients onto the null space of the historical features, theoretically ensuring zero interference to old tasks. Despite its success in classification and two-stage detection, the application of NSGP in YOLO is hindered by two critical bottlenecks, severe intermediate feature drift and class confusion. Departing from existing paradigms, we introduce a novel framework featuring ESPReg and CAD. It adaptively penalizes weight updates in critical feature sub-spaces and selectively distills semantic knowledge, thereby applying NSGP in single-stage IOD.

3 PRELIMINARIES

In this section, we formulate the Incremental Object Detection (IOD) problem in Sec. 3.1. Next, in Sec. 3.2, we present empirical results demonstrating two major architectural bottlenecks of YOLO-based incremental detectors: feature drift and cross-task class confusion. Finally, we establish a theoretical analysis for the upper bound of feature drift, explicitly relating it to the eigenvalues of the feature covariance matrix (Sec. 3.3).

3.1 PROBLEM DEFINITION

We formulate IOD as a sequential learning process over a stream of datasets $\mathbb{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$. At stage t, the dataset $\mathcal{D}_t = \{(\mathcal{X}_t, \mathcal{Y}_t)\}$ becomes available, where $\mathcal{X}_t$ denotes the images and $\mathcal{Y}_t$ contains annotations for the task-specific class set $\mathcal{C}_t$. Following the class-incremental setting, the class sets are mutually disjoint across stages, *i.e.*, $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \forall i \neq j$. The primary objective is to evolve the detector parameters from $\mathbf{\Theta}_{t-1}$ to $\mathbf{\Theta}_t$ using only $\mathcal{D}_t$ for training, while preventing performance degradation on all accumulated classes $\mathcal{C}_{1:t} = \bigcup_{k=1}^{t} \mathcal{C}_k$.

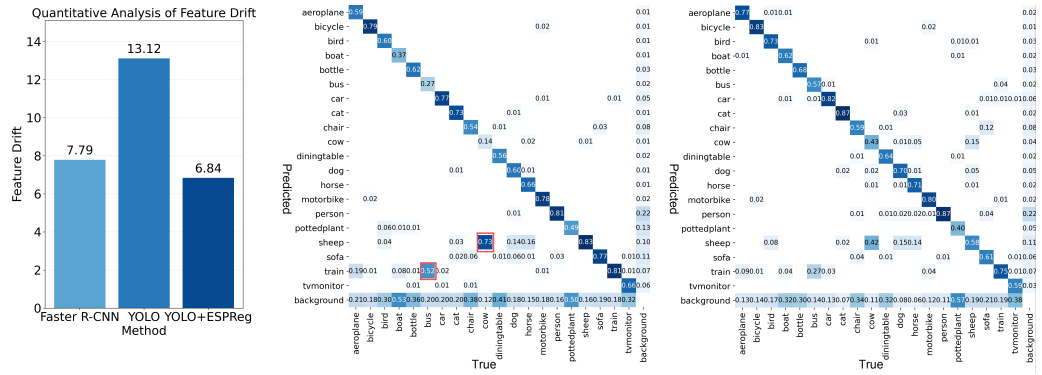
3.2 BOTTLENECKS OF YOLO INCREMENTAL DETECTOR

Applying the YOLO architecture directly to IOD reveals significant challenges in mitigating catastrophic forgetting. As shown in Fig. 1, our empirical analysis highlights two major architectural bottlenecks of YOLO-based incremental detectors: feature drift and cross-task class confusion.

We first investigate the bottleneck of feature drift. In IOD, the stability of the intermediate features for previous tasks is paramount to mitigating catastrophic forgetting. A critical observation in our study is the disparity in feature drift between YOLO-based and Faster R-CNN based incremental detectors. To quantify this phenomenon, we employ the drift at the output layer of the feature extractor (i.e., representations prior to the prediction heads) as a representative metric. As Fig. 1a indicates, Faster R-CNN exhibits a significantly lower degree of drift compared to YOLO detectors (the calculation details of feature drift are provided in Appendix A.1).

This discrepancy is largely due to the end-to-end dense prediction paradigm of YOLO, which induces overwhelming back-propagated gradients that easily perturb historical knowledge embedded in the feature extractor parameters. In contrast, Faster R-CNN demonstrates superior feature stability because its class-agnostic Region Proposal Network requires minimal reconfiguration for new categories, naturally constraining back-propagated gradients and isolating the backbone from drastic shifts. Therefore, in response to the feature drift faced by the incremental learning of YOLO, we designed ESPReg in Sec. 4.1.

Fig. 1b demonstrates the second bottleneck of YOLO-based incremental detectors: cross-task class confusion. When sequentially trained on the Pascal VOC 15+5 class-incremental setting, relying solely on ESPReg leaves the classifier biased towards the new task's distribution. Consequently, the model exhibits a pronounced tendency to misclassify instances from the initial 15 historical categories into the 5 visually similar novel ones, as highlighted by the red boxes (e.g., cow→sheep, bus→train). We hypothesize that YOLO's severe cross-task class confusion stems from its one-stage training paradigm. Under extreme incremental data imbalance, while some images in the new task happen



(a) Feature Drift (b) Without CAD (c) With CAD

Figure 1: Bottlenecks in YOLO-based IOD and the effectiveness of our solutions. (a) The comparison of feature drift between Faster R-CNN, YOLO and YOLO+ESPReg. The drift is measured by the average Euclidean distance at the final feature extractor layer over all historical samples.(b) and (c) represent the confusion matrices which report the average post-softmax confidence scores computed across all test samples of the relevant categories. The experiment follows the Pascal VOC 15+5 incremental setting, where *pottedplant, sheep, sofa, train, and tvmonitor* are introduced as novel categories.

to contain instances of historical categories, they are rare. Consequently, the BCE loss may induce constant negative gradients that systematically suppress the output channels of historical tasks. We infer that this mechanism creates a critical confidence scale mismatch, overshadowing old categories during task-agnostic inference. Regardless of whether this specific gradient-level dynamic is the sole trigger, the macroscopic consequence, the skewed confidence distribution, is an observable phenomenon that can be directly corrected. Motivated by this, we design a Confidence-Aware Distillation (CAD) mechanism (Sec. 4.2) to explicitly calibrate cross-task classification confidence from a distributional perspective. As shown in Fig. 1c, CAD effectively rectifies the task bias and mitigates this cross-task confusion.

3.3 Feature Drift Analysis

Feature drift is a key bottleneck affecting the incremental learning of YOLO. Therefore, we conducted a theoretical analysis of the feature drift. To establish a theoretical bound on the feature drift of YOLO-based incremental detectors, we analyze the influence of parameter updates on the network's intermediate features. The detailed analysis process and formula derivation can be found in the appendix A.1. Without loss of generality, consider the l-th layer of the network with input features $\boldsymbol{x}^l \in \mathbb{R}^{d_{in}}$. Let $\boldsymbol{W}_{t-1}^l \in \mathbb{R}^{d_{out} \times d_{in}}$ denote the parameters converged at stage $t-1$. The forward propagation is given by:

$$\boldsymbol{x}^{l+1} = \sigma(\boldsymbol{W}_{t-1}^l \boldsymbol{x}^l), \tag{1}$$

where $\sigma(\cdot)$ is a non-linear activation function. When learning the new task at stage t, the weights are updated by a perturbation $\Delta \boldsymbol{W}_t^l$, resulting in $\boldsymbol{W}_t^l = \boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l$. The resulting shift in the output feature is:

$$\Delta \boldsymbol{x}^{l+1} = \sigma((\boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l)\boldsymbol{x}^l) - \sigma(\boldsymbol{W}_{t-1}^l \boldsymbol{x}^l). \tag{2}$$

To quantify forgetting, we examine the magnitude of this shift. Adopting the standard assumption that the activation function $\sigma(\cdot)$ (e.g., ReLU, Leaky ReLU) satisfies Lipschitz continuity with Lipschitz constant $K = 1$, we bound the feature deviation as:

$$\|\Delta \boldsymbol{x}^{l+1}\|_2 = \|\sigma\left((\boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l)\boldsymbol{x}^l\right) - \sigma\left(\boldsymbol{W}_{t-1}^l \boldsymbol{x}^l\right)\|_2 \leq \|\Delta \boldsymbol{W}_t^l \boldsymbol{x}^l\|_2. \tag{3}$$

We define the **feature drift risk** ($\mathcal{E}_{\text{feat}}$) as the expected squared change over the historical data distribution $\mathcal{D}_{1:t-1} = \bigcup_{i=1}^{t-1} \mathcal{D}_i$. Using the linearity of expectation and the cyclic property of the trace operator, the upper bound of the stability risk is derived as:

$$\mathcal{E}_{\text{feat}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}}\left[\|\Delta \boldsymbol{x}^{l+1}\|_2^2\right] \leq \text{Tr}\left((\Delta \boldsymbol{W}_t^l)^\top (\Delta \boldsymbol{W}_t^l) \boldsymbol{\Sigma}_{1:t-1}^l\right), \tag{4}$$

where $\boldsymbol{\Sigma}_{1:t-1}^l = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}}[\boldsymbol{x}^l (\boldsymbol{x}^l)^\top]$ denotes the uncentered covariance matrix of the input features at the l-th layer from stages 1 to $t-1$. We perform spectral decomposition on this covariance matrix.

Let

$$\mathbf{\Sigma}_{1:t-1}^{l} = \sum_{k=1}^{d_{in}} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^{\top}, \tag{5}$$

where $\{\boldsymbol{u}_k, \lambda_k\}$ denote the normalized eigenvector-eigenvalue pairs. Substituting this into Eq. (4) yields a decoupled upper bound:

$$\mathcal{E}_{\text{feat}} \leq \text{Tr}\left((\Delta \boldsymbol{W}_t^l)^{\top} (\Delta \boldsymbol{W}_t^l) \sum_{k=1}^{d_{in}} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^{\top} \right) = \sum_{k=1}^{d_{in}} \lambda_k \left\| \Delta \boldsymbol{W}_t^l \boldsymbol{u}_k \right\|_2^2. \tag{6}$$

This formulation reveals that the feature drift risk of a parameter update direction is proportional to its corresponding eigenvalue in the historical feature covariance matrix. Consequently, we design a regularization term scaled by the eigenvalues of the parameters to minimize feature drift.

4 METHODOLOGY

4.1 EIGENVALUE SCALED PROJECTION REGULARIZATION

Motivated by the theoretical analysis in Sec. 3.3, we propose the Eigenvalue Scaled Projection Regularization method to demystifying feature drift of YOLO-based incremental detectors. Consider the incremental learning step from task t to $t+1$. We define the parameter deviation of the l-th layer as the difference between the current weights and the consolidated weights from the previous stage:

$$\Delta \boldsymbol{W}_{t+1}^{l} = \boldsymbol{W}_{t+1}^{l} - \boldsymbol{W}_{t}^{l}. \tag{7}$$

To control the feature drift caused by this deviation, we utilize the feature drift risk upper bound in Eq. (6) as a regularization objective. Directly using raw eigenvalues λ_k as regularization coefficients can lead to optimization difficulties, because the eigenvalues of significant directions change vastly (as is shown in Fig. 2b). To prevent the penalty on the most significant direction dominating the regularization, we introduce a calibrated importance score λ_k^*. Let $0 \leq r \leq d_{in}$ be the elbow point of eigenvalue sequence, where the second order difference reach the maximum value (refer (Wu et al., 2025)). The calibrated score is defined as:

$$\lambda_k^* = \min\left(\frac{\lambda_k}{\lambda_r}, 1 \right). \tag{8}$$

This formulation creates a truncated penalty weight: for dominant principal components ($\lambda_k > \lambda_r$), the penalty saturates at 1 and contributes to the regularization evenly, while for tail components ($\lambda_k < \lambda_r$), the penalty decays linearly with their importance.

Based on the feature drift risk upper bound derived in Eq. (6), we formulate the Eigenvalue Scaled Projection Regularization (ESPReg) loss as:

$$\mathcal{L}_{\text{reg}} = \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{d_{in}} \lambda_k^* \left\| \Delta \boldsymbol{W}_t^l \boldsymbol{u}_k \right\|_2^2. \tag{9}$$

Here, the term $\|\Delta \boldsymbol{W}_t^l \boldsymbol{u}_k\|_2^2$ quantifies the projection of the weight update onto the k-th principal component of historical features. By weighting this projection with λ_k^*, ESPReg effectively reshapes the optimization landscape: it imposes stringent regularization penalties along principal directions crucial for historical tasks (where λ_k^* is large), while permitting unconstrained parameter updates within the null space (where $\lambda_k^* \approx 0$).

Finally, the total objective function for the current task combines the standard detection loss with the proposed regularization:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{reg}}, \tag{10}$$

where $\mathcal{L}_{\text{det}}$ includes classification and regression losses for the new task.

4.2 CONFIDENCE-AWARE DISTILLATION

As empirical results in Fig. 1b, **ESPReg** alone cannot completely mitigate the severe cross-task class confusion.

To resolve this issue, we introduce a Confidence-Aware Distillation (CAD) method. Let $\mathcal{M}_t$ represent the frozen base model from the previous stage and $\mathcal{M}_{t+1}$ denote the current model being trained. For a given input image, the YOLO detection head outputs predictions for N anchors. Let $\boldsymbol{p}_{t,i}$ and $\boldsymbol{p}_{t+1,i}$ denote the classification probability vectors for the i-th anchor generated by the base and current models, respectively.

Instead of indiscriminately distilling all output channels across all anchors, which inevitably includes task-irrelevant background noise and suppresses the learning of novel classes, we selectively distill knowledge guided by the base model's confidence. Specifically, for the i-th anchor, we calculate the maximum classification score across all previous class channels $\mathcal{C}_t$. We define this maximum score as the confidence mask m_i:

$$m_i = \max_{c \in \mathcal{C}_t} p_{t,i,c}. \tag{11}$$

Let $\hat{c}_i$ be the specific category index where this maximum score is achieved:

$$\hat{c}_i = \arg\max_{c \in \mathcal{C}_t} p_{t,i,c}. \tag{12}$$

To effectively transfer the classification knowledge while avoiding interference with new classes, we compute the Kullback-Leibler (KL) divergence () between the base and current models exclusively on this most prominent channel $\hat{c}_i$. By employing m_i as a weighting factor, we compel the optimizer to strictly anchor its predictions on regions where the base model exhibits high confidence, while granting plasticity to ambiguous or background anchors. The proposed classification distillation loss is formulated as a weighted average:

$$\mathcal{L}_{\text{cad}} = \frac{1}{\sum_{j=1}^{N} m_j} \sum_{i=1}^{N} m_i \cdot \mathcal{D}_{\text{KL}} \left(p_{t,i,\hat{c}_i} \parallel p_{t+1,i,\hat{c}_i} \right), \tag{13}$$

where $\mathcal{D}_{\text{KL}}(p \parallel q) = p \log(\frac{p}{q}) + (1-p) \log(\frac{1-p}{1-q})$ denotes the binary KL divergence.

Consequently, the overall objective function presented in Eq. (10) is updated to incorporate this distillation strategy:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cad}}. \tag{14}$$

This dual-regularization approach ensures that feature drift is minimized via ESPReg while cross-task class confusion is demystified via Confidence-Aware Distillation.

5 Experiments

Datasets and Protocols. Our experiments are conducted on Pascal VOC (Everingham et al., a;b) and MS COCO 2017 (Lin et al., 2014) datasets, which are widely adopted benchmarks for IOD. We comprehensively evaluate the performance of our ESP-YOLO on both single-step incremental settings and multi-step incremental settings. Specifically, for Pascal VOC, we adopt two single-step splits (19+1 and 15+5) and three multi-step splits (10-5 with 3 tasks, 5-5 with 4 tasks, and 10-2 with 6 tasks); for MS COCO 2017, we carry out evaluation on two standard single-step splits (70+10 and 40+40), covering different incremental learning difficulty levels. In the A+B or A-B protocol, the detector is first trained on A base classes and then sequentially updated with B novel classes at each incremental stage. Following prior IOD protocols (Liu et al., 2023b; Wu et al., 2025), we report *Old*, *New*, *All*, and *Avg* on Pascal VOC, denoting mAP over old classes, new classes, all seen classes, and the average of Old and New, respectively.

Implementation Details. We adopt YOLOv8 (Varghese & Sambath, 2024) as the base detector, upon which we apply ESPReg and CAD to build our ESP-YOLO. The model is trained for 100 epochs per incremental stage with a batch size of 16, using the SGD optimizer with a learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4}. All experiments are performed on a single NVIDIA 4090 GPU. For fair comparison, we report published results for existing IOD methods when available and re-implement several of them on our YOLOv8 baseline, including two standard reference bounds following prior protocols (Liu et al., 2023b; Mo et al., 2024; Wu et al., 2025): *Joint*, trained on the union of all stage-wise data with complete annotations, and *Fine-tuning*, sequentially trained on incremental data without replay, distillation, or regularization.

Comparison Results. Tabs. 1, 2, 3 summarize the comparison results between our proposed ESP-YOLO and comparable IOD models on different benchmarks. On the Pascal VOC single-step incremental benchmark (Tab. 1), our ESP-YOLO achieves significant performance gains and

Table 1: **Comparison with state-of-the-art IOD methods on Pascal VOC test 2007.** We highlight the **best** and second-best results. '-' indicates results not available. Note that Joint and Fine-tuning serve as the theoretical upper and lower bounds for the results, and are therefore excluded from the ranking.

Baseline	Method	19+1				15+5			
		Old	New	All	Avg	Old	New	All	Avg
YOLOv8	Joint	78.1	82.8	78.0	80.5	79.8	79.1	78.0	79.5
	Fine-tuning	0.0	82.4	3.9	41.2	0.0	78.4	11.5	39.2
Faster R-CNN	FasterILOD (Ren et al., 2016)	68.9	61.1	68.5	65.0	71.6	56.9	67.9	64.3
	ILOD-Meta (Joseph et al., 2021b)	70.9	57.6	70.2	64.2	71.7	55.9	67.8	63.8
	ORE (Joseph et al., 2021a)	69.4	60.1	68.9	64.7	71.8	58.7	68.5	65.2
	PPAS (Zhou et al., 2020)	70.5	53.0	69.2	61.8	-	-	-	-
	MVC (Yang et al., 2022)	70.2	60.6	69.7	65.4	69.4	57.9	66.5	63.7
	PROB (Zohar et al., 2023)	73.9	48.5	72.6	61.5	73.5	60.8	70.1	67.0
	PseudoRM (Yang et al., 2023b)	72.9	67.3	72.6	70.1	73.4	60.9	70.3	66.9
	MMA (Cermelli et al., 2020)	71.1	63.4	70.7	67.2	73.0	60.5	69.9	66.7
	BPF (Mo et al., 2024)	74.5	65.3	74.1	69.9	75.9	63.0	72.7	69.5
	OSR (Yang et al., 2023a)	71.8	66.5	71.5	69.2	72.2	61.2	69.5	66.7
	ABR (Liu et al., 2023b)	71.0	69.7	70.9	70.4	73.0	65.1	71.0	69.1
	NSGP-RePRE (Wu et al., 2025)	76.3	69.0	**76.0**	72.7	77.5	61.8	73.6	69.7
YOLOv8	EWC (Kirkpatrick et al., 2017)	-	-	-	-	63.2	47.1	59.2	55.2
	NSGP (Wu et al., 2025)	-	-	-	-	72.8	48.2	66.6	60.5
	NSGP-RePRE (Wu et al., 2025)	-	-	-	-	72.4	49.1	66.6	60.8
	BPF (Mo et al., 2024)	-	-	-	-	54.6	**80.1**	53.2	67.4
	ESP-YOLO (Ours)	**76.9**	**72.0**	75.5	**74.5**	**80.4**	77.1	**75.0**	**78.8**

Table 2: **Comparison with state-of-the-art IOD methods on the MS COCO 2017 validation set.** All metrics are computed over all seen classes after the single-step incremental stage. We highlight the **best** and second-best results among incremental methods. '-' indicates unavailable results. Joint training serves as the reference upper bound and is excluded from the ranking.

Method	Baseline	40+40 Task			70+10 Task		
		mAP	AP75	AP50	mAP	AP75	AP50
YOLOv8	Joint	53.9	58.8	70.4	53.9	58.8	70.4
Faster R-CNN	ILOD-Meta (Joseph et al., 2021b)	23.8	24.4	40.5	-	-	-
	FasterILOD (Ren et al., 2016)	20.6	-	40.1	21.3	-	39.9
	MMA (Cermelli et al., 2020)	33.0	34.6	56.6	30.2	31.5	52.1
	PseudoRM (Yang et al., 2023b)	25.3	-	44.4	-	-	-
	ABR (Liu et al., 2023b)	34.5	35.2	57.8	31.1	32.7	52.9
	BPF (Mo et al., 2024)	34.4	37.3	54.3	36.2	38.9	56.8
	NSGP (Wu et al., 2025)	35.4	38.6	55.3	36.5	39.8	56.0
YOLOv8	EWC (Kirkpatrick et al., 2017)	36.4	39.8	51.2	-	-	-
	NSGP (Wu et al., 2025)	32.9	35.4	47.5	-	-	-
	BPF (Mo et al., 2024)	16.7	18.2	23.0	-	-	-
	ESP-YOLO (Ours)	**42.5**	**46.4**	**58.0**	**44.8**	**48.7**	**60.6**

comprehensive superiority over existing SOTA methods across most evaluation metrics and settings. Specifically, in the 19+1 setting, our method reaches 76.9% Old class mAP, 72.0% New class mAP and 74.5% Avg mAP, all ranking first among all compared methods; in the more challenging 15+5 setting, ESP-YOLO further expands its advantages, achieving 80.4% Old class mAP, 75.0% All class mAP and 78.8% Avg mAP, all of which are the best results in the comparison, while its 77.1% New class mAP ranks a close second.

On MS COCO 2017 single-step incremental benchmark (Tab. 2), ESP-YOLO also exhibits prominent performance advantages, outperforming all compared methods on all evaluation metrics for both 40+40 and 70+10 settings. In the 40+40 task, ESP-YOLO achieves 42.5% mAP, 46.4% AP75 and 58.0% AP50, surpassing the previous best method by a large margin; in the 70+10 task, our method further reaches 44.8% mAP, 48.7% AP75 and 60.6% AP50, consistently ranking first in all metrics. These results fully validate the strong generalization ability and effectiveness of our ESP-YOLO on large-scale datasets with more complex categories.

For the multi-step incremental learning settings on Pascal VOC (Tab. 3), which are closer to real-world continual learning scenarios, ESP-YOLO also outperforms existing methods by ranking first on all metrics in the 10-5, 5-5 and 10-2 settings.

6 Ablations

Our ablation experiments are conducted on the Pascal VOC 15+5 single-stage class-incremental detection setting to evaluate the contribution of each proposed component.

Table 3: **Comparison with state-of-the-art IOD methods on Pascal VOC test 2007 in multi-step settings.** For each protocol, the three columns report mAP on old/base classes, newly introduced classes, and all seen classes after completing all incremental stages, respectively. We highlight the **best** and second-best results among incremental methods. '-' indicates results not available. Note that Joint and Fine-tuning serve as the theoretical upper and lower bounds for the results, and are therefore excluded from the ranking.

Baseline	Method	10-5 (3 tasks)			5-5 (4 tasks)			10-2 (6 tasks)		
		1-10	11-20	1-20	1-5	6-20	1-20	1-10	11-20	1-20
YOLOv8	Joint	80.5	79.7	78.0	78.7	79.0	78.0	80.5	79.7	78.0
	Fine-tuning	5.3	30.6	18.0	0.5	18.3	13.8	3.8	13.6	8.7
Faster R-CNN	ABR (Liu et al., 2023b)	68.7	67.1	67.9	64.7	56.4	58.4	67.0	58.1	62.6
	FasterILOD (Ren et al., 2016)	68.3	57.9	63.1	55.7	16.0	25.9	64.2	48.6	56.4
	MMA (Cermelli et al., 2020)	66.7	61.8	64.2	62.3	31.2	38.9	65.0	53.1	59.1
	BPF (Mo et al., 2024)	69.1	68.2	68.7	60.6	63.1	62.5	68.7	56.3	62.5
	NSGP-RePRE (Wu et al., 2025)	72.4	67.6	70.0	64.6	66.1	65.7	70.1	58.8	64.4
YOLOv8	**ESP-YOLO**	**74.2**	**69.4**	**70.8**	**70.1**	**67.3**	**66.2**	**74.9**	**70.2**	**70.9**

Table 4: **Ablation study of different components on Pascal VOC test 2007 (15+5 setting).** All experiments are built upon the YOLOv8 (Base). ESPReg: Eigenvalue Scaled Projection Regularization, CAD: Confidence-Aware Distillation.

Base	ESPReg	CAD	15+5			
			Old	New	All	Avg
✓			51.0	78.3	49.0	64.7
✓	✓		72.6	76.8	66.6	74.7
✓		✓	65.3	76.9	67.4	71.1
✓	✓	✓	**80.4**	**77.1**	**75.0**	**78.8**

6.1 Effectiveness of Components

We analyze the impact of ESPReg and CAD on the incremental learning performance in Tab. 4. Here, *Base* denotes the YOLOv8 under sequential fine-tuning with pseudo labeling (Liu et al., 2023b; Shmelkov et al., 2017), which is a common baseline setting in IOD. Several key observations can be drawn: (1) **ESPReg significantly mitigates forgetting**: Compared to the Base in Tab.4, adding ESPReg improves the mAP on old classes (Task-1) from 51.0% to 72.6%, demonstrating its superior capability in preserving historical knowledge. (2) **CAD enhances task-agnostic performance**: While ESPReg focuses on stability, CAD bridges the performance gap between task-aware and task-agnostic scenarios by eliminating cross-task class confusion, which is further demonstrated in Sec. 6.3. (3) **Synergistic Effect**: The combination of both components achieves the optimal incremental result (78.8% Avg mAP).

6.2 Effectiveness of ESPReg

To quantitatively assess the stability of the representation, we first measure the **feature drift** of the feature extractor after incremental training. Specifically, we randomly select 100 samples from the historical tasks and extract their features using both the initial base model and the incrementally updated model. We then compute the average Euclidean distance between these two sets of features. As shown in Tab. 5, the model regularized by ESPReg exhibits significantly lower feature drift compared to the vanilla finetuning model, the model regularized by EWC (Kirkpatrick et al., 2017) and NSGP (Wu et al., 2025). And as is shown in Fig. 1a, the feature drift of YOLO+ESPReg is significantly lower than Faster R-CNN. These results confirm that ESPReg effectively stabilizes the intermediate feature of IOD model.

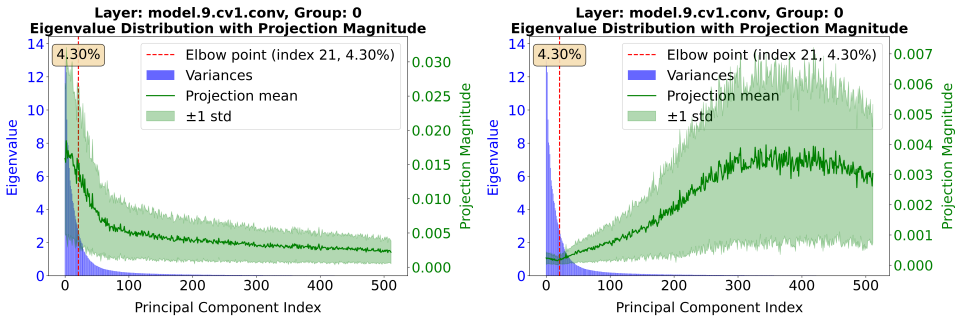
We then visualize the parameter changes $\Delta \boldsymbol{W}$ in the first convolutional block of layer 9 (last layer of feature extractor). Specifically, we plot the magnitude of these weight updates projected onto the eigenvectors of the historical feature covariance, overlaid with the eigenvalue distribution. As illustrated in Fig. 2a, without regularization, the optimization naturally gravitates towards the **significant subspace** (the left region with high eigenvalues), causing severe interference with historical knowledge. In contrast, Fig. 2b demonstrates that ESPReg effectively **suppresses updates in these high-variance directions**. Instead, it steers the optimization towards the **null space** (the right region beyond the red dotted line), thereby preserving stability while utilizing the low-risk subspace for plasticity. The red dotted line represents the elbow point defined by (Wu et al., 2025).

Table 5: **Quantitative analysis of feature drift.** We report the average Euclidean distance of backbone features between the base model and the model after Task-2 training (15+5 setting).

Method	Fine-tuning	EWC	NSGP	**ESPReg**
Feature Drift (↓)	13.12	11.17	10.25	**6.84**

6.3 Effectiveness of CAD

We further investigate the cross-task confusion using confusion matrices (visualized in Fig. 1). As shown in Fig. 1b, when only ESPReg is employed to mitigate forgetting, the model's classifier remains susceptible to class imbalance within the training set. This leads to Task-1 categories being frequently misclassified as visually similar categories in Task-2, causing a noticeable drop in task-agnostic detection accuracy. As shown in Fig. 1c, the introduction of the CAD method significantly alleviates this cross-task confusion. The lower-left region of the confusion matrix becomes markedly sparser. For instance, in the challenging "cow-sheep" confusion group, the misclassification rate decreases from 73% to 42%. Consequently, the task-agnostic detection accuracy with CAD approaches the median of task-aware detection accuracies across individual tasks.



(a) Without Regularization Loss (b) With Regularization Loss

Figure 2: **Effectiveness of ESPReg.** We visualize the projection magnitude of weight updates (green curve) against the historical feature eigenvalues (blue bars). The elbow point is also represented (red dotted line). **(a)** Without ESPReg, the weight updates dominate the significant subspace, leading to drastic intermediate feature drift. **(b)** With ESPReg, weight updates are explicitly penalized in the significant subspace (left) and concentrated in the null space (right), indicating minimal intermediate feature drift.

7 Conclusion and Discussion

In this paper, we propose Eigenvalue Scaled Projection Regularization (ESPReg) to resolve the drastic feature drift and cross-task class confusion in YOLO-based Incremental Object Detector. Grounded in a rigorous theoretical analysis of the feature drift risk, ESPReg provides an adaptive, eigenvalue-weighted regularization to dynamically balance the retention of historical knowledge with the learning of new concepts. Additionally, our Confidence-Aware Distillation demystifies cross-task class confusion, ensuring performance in task-agnostic inference scenario. Extensive experiments on Pascal VOC and MS COCO confirm that ESPReg consistently outperforms state-of-the-art methods.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing and correcting grammars.

ETHICS STATEMENT

This study uses open source datasets (Pascal VOC and MS COCO), which do not raise any ethics concern.

REFERENCES

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9230–9239. IEEE, 2020.

Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Incremental-detr: Incremental few-shot object detection via self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 543–551, 2023.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, a.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, b.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

Yu Hao, Yanwei Fu, Yu-Gang Jiang, and Qi Tian. An end-to-end architecture for class-incremental object detection with knowledge distillation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2019.

K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1391–1400, 2021a.

KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9209–9216, 2021b.

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*, 114 (13):3521–3526, 2017.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. In *The Tenth International Conference on Learning Representations*, 2022.

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*, pp. 740–755. Springer, 2014.

Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, and Christian Rupprecht. Continual detection transformer for incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23799–23808, 2023a.

Yuyang Liu, Yang Cong, Dipam Goswami, Xialei Liu, and Joost Van De Weijer. Augmented box replay: Overcoming foreground shift for incremental object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11367–11377, 2023b.

Angelo G Menezes, Gustavo de Moura, Cézanne Alves, and André CPLF de Carvalho. Continual object detection: a review of definitions, strategies, and challenges. *Neural networks*, 161:476–493, 2023.

Q. Mo, Y. Gao, S. Fu, J. Yan, A. Wu, and W.-S. Zheng. Bridge past and future: Overcoming information asymmetry in incremental object detection. In *European Conference on Computer Vision (ECCV)*, pp. 463–480. Springer, 2024.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

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*, 20(23):6777, 2020.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pp. 3400–3409, 2017.

Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *International conference on advances in data engineering and intelligent computing systems (ADICS)*, 2024.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Qirui Wu, Shizhou Zhang, De Cheng, Yinghui Xing, Di Xu, PENG WANG, and Yanning Zhang. Demystifying catastrophic forgetting in two-stage incremental object detector. In *Forty-second International Conference on Machine Learning*, 2025.

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*, 131: 108863, 2022.

Dongbao Yang, Yu Zhou, Xiaopeng Hong, Aoting Zhang, and Weiping Wang. One-shot replay: Boosting incremental object detection via retrospecting one object. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3127–3135, 2023a.

Dongbao Yang, Yu Zhou, Xiaopeng Hong, Aoting Zhang, Xin Wei, Linchengxi Zeng, Zhi Qiao, and Weipinng Wang. Pseudo object replay and mining for incremental object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 153–162, 2023b.

Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Revisiting flatness-aware optimization in continual learning with orthogonal gradient projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jichuan Zhang, Wei Li, Shuang Cheng, Yali Li, and Shengjin Wang. Dynamic object queries for transformer-based incremental object detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Zhen Zhao, Zhizhong Zhang, Xin Tan, Jun Liu, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Rethinking gradient projection continual learning: Stability/plasticity feature space decoupling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3718–3727, 2023.

W. Zhou, S. Chang, N. Sosa, H. Hamann, and D. Cox. Lifelong object detection. *arXiv preprint arXiv:2009.01129*, 2020.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25613–25622, 2025.

O. Zohar, K.-C. Wang, and S. Yeung. Prob: Probabilistic objectness for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11444–11453, 2023.

A APPENDIX

A.1 TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

Disparity of Feature Drift between YOLO and Faster R-CNN. Given the architectural heterogeneity between YOLO and Faster R-CNN frameworks, a direct layer-wise comparison of their internal feature drift is intractable. Consequently, we employ the feature drift at the output layer of the feature extractor—specifically, the layer immediately preceding the classification and regression heads—as a representative metric for our analysis.

Formally, the feature drift $\mathcal{E}_{feat}^{L_f}$ is calculated as follows:

$$\mathcal{E}_{feat}^{L_f} = \frac{1}{HW} \sum_{i=1}^{W} \sum_{j=1}^{H} \left\| \boldsymbol{F}_{t+1}^{L_f}[i,j] - \boldsymbol{F}_{t}^{L_f}[i,j] \right\|_2, \tag{15}$$

where L_f denotes the index of the final feature extraction layer before the prediction heads. $\boldsymbol{F}_t^{L_f}$ and $\boldsymbol{F}_{t+1}^{L_f}$ represent the output feature maps at layer L_f yielded by the detector after training on task t and task $t+1$, respectively. Furthermore, $\|\cdot\|_2$ denotes the L_2-norm, while H and W correspond to the spatial height and width of the feature map.

Detailed Analysis of Feature Drift. To establish a theoretical bound on the feature drift of YOLO-based incremental detectors, we analyze the influence of parameter updates on the network's intermediate features. Without loss of generality, consider the l-th layer of the network (e.g., fully connected or convolutional) with input features $\boldsymbol{x}^l \in \mathbb{R}^{d_{in}}$. Let $\boldsymbol{W}_{t-1}^l \in \mathbb{R}^{d_{out} \times d_{in}}$ denote the parameters converged at stage $t-1$. The forward propagation is given by:

$$\boldsymbol{x}^{l+1} = \sigma(\boldsymbol{W}_{t-1}^l \boldsymbol{x}^l), \tag{16}$$

where $\sigma(\cdot)$ is a non-linear activation function. When learning the new task at stage t, the weights are updated by a perturbation $\Delta \boldsymbol{W}_t^l$, resulting in $\boldsymbol{W}_t^l = \boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l$. The resulting shift in the output feature is:

$$\Delta \boldsymbol{x}^{l+1} = \sigma((\boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l)\boldsymbol{x}^l) - \sigma(\boldsymbol{W}_{t-1}^l \boldsymbol{x}^l). \tag{17}$$

To quantify forgetting, we examine the magnitude of this shift. Adopting the standard assumption that the activation function $\sigma(\cdot)$ (e.g., ReLU, Leaky ReLU) satisfies Lipschitz continuity with Lipschitz constant $K = 1$, we bound the feature deviation as:

$$\begin{aligned} \|\Delta \boldsymbol{x}^{l+1}\|_2 &= \|\sigma\left((\boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l)\boldsymbol{x}^l\right) - \sigma\left(\boldsymbol{W}_{t-1}^l \boldsymbol{x}^l\right)\|_2 \\ &\leq \|(\boldsymbol{W}_{t-1}^l + \Delta \boldsymbol{W}_t^l)\boldsymbol{x}^l - \boldsymbol{W}_{t-1}^l \boldsymbol{x}^l\|_2 \\ &= \|\Delta \boldsymbol{W}_t^l \boldsymbol{x}^l\|_2. \end{aligned} \tag{18}$$

We define the **feature drift risk** ($\mathcal{E}_{\text{feat}}$) as the expected squared change over the historical data distribution $\mathcal{D}_{1:t-1} = \bigcup_{i=1}^{t-1} \mathcal{D}_i$. Using the linearity of expectation and the cyclic property of the trace operator, the upper bound of the stability risk is derived as:

$$\begin{aligned} \mathcal{E}_{\text{feat}} &= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}} \left[\|\Delta \boldsymbol{x}^{l+1}\|_2^2 \right] \\ &\leq \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}} \left[\|\Delta \boldsymbol{W}_t^l \boldsymbol{x}^l\|_2^2 \right] \\ &= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}} \left[(\boldsymbol{x}^l)^\top (\Delta \boldsymbol{W}_t^l)^\top (\Delta \boldsymbol{W}_t^l) \boldsymbol{x}^l \right] \\ &= \text{Tr}\left((\Delta \boldsymbol{W}_t^l)^\top (\Delta \boldsymbol{W}_t^l) \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}} [\boldsymbol{x}^l (\boldsymbol{x}^l)^\top] \right) \\ &= \text{Tr}\left((\Delta \boldsymbol{W}_t^l)^\top (\Delta \boldsymbol{W}_t^l) \boldsymbol{\Sigma}_{1:t-1}^l \right), \end{aligned} \tag{19}$$

where $\boldsymbol{\Sigma}_{1:t-1}^l = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{1:t-1}} [\boldsymbol{x}^l (\boldsymbol{x}^l)^\top]$ denotes the uncentered covariance matrix of the input features at the l-th layer from stages 1 to $t-1$. We perform spectral decomposition on this covariance matrix. Let

$$\boldsymbol{\Sigma}_{1:t-1}^l = \sum_{k=1}^{d_{in}} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top, \tag{20}$$

where $\{\boldsymbol{u}_k, \lambda_k\}$ denote the normalized eigenvector-eigenvalue pairs. Substituting this into Eq. (19) yields a decoupled upper bound:

$$\begin{aligned}\mathcal{E}_{\text{feat}} &\leq \operatorname{Tr}\left((\Delta \boldsymbol{W}_t^l)^\top (\Delta \boldsymbol{W}_t^l) \sum_{k=1}^{d_{in}} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top\right) \\ &= \sum_{k=1}^{d_{in}} \lambda_k \cdot \operatorname{Tr}\left((\Delta \boldsymbol{W}_t^l)^\top (\Delta \boldsymbol{W}_t^l) \boldsymbol{u}_k \boldsymbol{u}_k^\top\right) \\ &= \sum_{k=1}^{d_{in}} \lambda_k \left\|\Delta \boldsymbol{W}_t^l \boldsymbol{u}_k\right\|_2^2 .\end{aligned} \tag{21}$$

This formulation reveals that the feature drift risk of a parameter update direction is proportional to its corresponding eigenvalue in the historical feature covariance matrix. Consequently, we design a regularization term scaled by the corresponding eigenvalues of the parameters to minimize feature drift.