

Article

Stable Incremental Underwater Object Detection via Adaptive Representation Routing and Topology-Preserved Replay

Shaodong Zhang ¹, Feng Tian ¹, Haiyang Yao ^{2,*} , Jinhao Shi ² and Yongsheng Yan ³ ¹ Marine Design and Research Institute of China, Shanghai 200011, China² School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710016, China³ School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

* Correspondence: yaohaiyang@sust.edu.cn

Abstract

Incremental object detection (IOD) is critical for autonomous underwater perception, where detectors deployed on long-duration underwater platforms must continuously adapt to evolving marine environments while retaining previously learned recognition and localization capabilities. However, underwater IOD is particularly challenging because visual degradation, small-object ambiguity, background dominance, rare-class dilution, and non-stationary data distributions jointly cause structure instability in incremental representations. Existing response-based distillation methods, such as elastic response distillation (ERD), mainly preserve output-level responses and often rely on rigid backbone updating and static optimization strategies, making them insufficient for maintaining hierarchical representation stability under degraded and imbalanced underwater observations. To address these limitations, we propose a structure-stable underwater IOD framework that jointly regulates representation update, replay topology, and optimization dynamics within a unified stability–plasticity formulation. Specifically, Structure-Adaptive Residual Routing (SARR) replaces binary layer freezing with adaptive residual paths, task-aware gradient routing, and semantic-sensitive update gates, enabling parameter-efficient incremental representation routing. Topology-Aware Semantic Replay (TASR) maintains class prototypes, teacher-guided relation matrices, and decision-boundary anchor samples to preserve the neighborhood topology of rare old classes in the feature manifold. Uncertainty-Aware Plasticity–Stability Feedback (UPSF) dynamically adjusts classification and localization distillation strengths according to old-class forgetting risk, new-class learning difficulty, and localization uncertainty. Extensive experiments on UTDAC2020 and DUO demonstrate that the proposed method outperforms representative distillation- and transformer-based IOD baselines in most incremental settings. In particular, our method achieves 30.7% AP on UTDAC2020 under the 2 + 2 setting and narrows the gap to full-data training, validating the effectiveness of structure-stable representation evolution for robust incremental underwater object detection.



Academic Editor: Chia-Hung Yeh

Received: 1 June 2026

Revised: 9 July 2026

Accepted: 17 July 2026

Published: 19 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: underwater object detection; vision; continual learning; topology-preserved distillation; stability–plasticity dilemma; uncertainty-aware optimization

1. Introduction

Autonomous underwater perception is a fundamental capability for long-duration marine exploration, ecological monitoring, and intelligent intervention by autonomous

underwater vehicles (AUVs) [1,2]. Unlike object detection in closed and relatively stable terrestrial scenarios, underwater object detection operates in open, dynamic, and partially observable environments, where visual evidence is continuously affected by turbidity, light attenuation, color distortion, scattering, occlusion, and background clutter [1,3]. During long-term missions, AUVs may encounter unseen marine species, novel object appearances, unfamiliar seabed structures, and changing imaging conditions caused by variations in illumination, depth, and water quality. These factors make the data distribution evolve over time. Therefore, conventional detectors trained once on predefined categories and fixed environmental conditions often degrade when deployed in real underwater missions. Since repeatedly retraining detectors from scratch is impractical due to limited historical data, expensive annotation, and constrained onboard computation and storage, underwater object detection requires an incremental learning paradigm that can continuously acquire new visual knowledge while preserving previously learned recognition and localization capabilities [4,5].

Underwater IOD faces a severe stability–plasticity dilemma. Visual degradation weakens object boundaries and suppresses discriminative textures, making newly observed categories difficult to distinguish from previously learned ones. Many marine organisms are small, deformable, partially occluded, or visually similar to surrounding substrates, which increases small-object ambiguity. Meanwhile, underwater images are often dominated by non-target regions such as rocks and sand, causing object-level supervision to be diluted, especially when replay memory is limited. In addition, underwater object distributions are typically long-tailed and category-scarce: frequent species provide relatively sufficient samples, whereas rare or newly emerging targets may appear only in few-shot or episodic forms. These factors jointly perturb the feature space during incremental updates, leading to old-class forgetting and decision-boundary collapse. Thus, underwater IOD is not merely a task of adding new categories but a problem of maintaining hierarchical representation stability and class-topology consistency under degraded, imbalanced, and non-stationary marine visual distributions.

Existing IOD methods have attempted to mitigate catastrophic forgetting through knowledge distillation, experience replay, transformer-based detection, and vision–language model-assisted adaptation [4–17]. Distillation-based methods preserve old knowledge by aligning feature maps, prediction responses, semantic relations, or optimization trajectories between teacher and student detectors, while replay-based approaches rehearse historical samples, foreground objects, or compact prototypes under memory constraints [8–13]. Transformer-based IOD further adapts object queries and detector-specific distillation to incremental detection [14], and recent vision–language model-based methods exploit cross-modal priors to refine pseudo labels or preserve global semantic consistency [15–17]. Despite their effectiveness, most existing methods are designed and evaluated on generic natural-image benchmarks, where illumination, object appearance, and background distributions are relatively more stable than those in underwater scenes [8,14]. They mainly regard forgetting as an output-level or annotation-level mismatch, while paying insufficient attention to the hierarchical instability of visual representations caused by underwater degradation. As a result, rigid backbone updating or global distillation constraints may either erode generic underwater optical features or restrict the adaptation of high-level semantics to new marine categories.

To address this problem, we propose a structure-stable underwater IOD framework that jointly regulates representation update, replay topology, and optimization dynamics within a unified stability–plasticity formulation. Specifically, Structure-Adaptive Residual Routing (SARR) introduces lightweight residual adapters and task-aware gates to decouple the update behavior of different backbone stages, preserving stable low-level

optical representations while enabling high-level semantic adaptation. Topology-Aware Semantic Replay (TASR) maintains class prototypes, teacher-guided relation matrices, and boundary-sensitive anchor samples to construct informative replay batches, thereby alleviating background dominance, rare-class dilution, and topology drift. In addition, Uncertainty-Aware Plasticity–Stability Feedback (UPSF) dynamically adjusts the optimization balance between new-category learning and old-knowledge preservation according to old-class forgetting risk, new-class learning difficulty, and localization uncertainty. Through these complementary designs, the proposed framework provides a practical and unified solution for maintaining representation stability while retaining sufficient semantic plasticity in incremental underwater object detection. The novelty of this work lies in shifting underwater IOD from output-level response preservation to representation-structure stability. Instead of simply distilling old predictions or replaying historical samples, the proposed framework jointly controls three coupled factors: layer-wise representation evolution, class-topology preservation, and uncertainty-aware optimization. This design is particularly suitable for underwater scenarios, where degraded imaging conditions, scarce categories, and background-dominated observations can jointly distort feature hierarchies, class–neighborhood relations, and decision boundaries.

In this paper, “structure stability” refers to the stability of learned representation structures during incremental learning, including hierarchical feature organization, class–neighborhood topology, and decision-boundary consistency. It does not refer to the physical structural stability of underwater vehicles or marine objects.

The main contributions of this work are summarized as follows.

- (1) We formulate incremental underwater object detection as a structure-stability problem under visually degraded, category-scarce, long-tailed, and non-stationary marine distributions. This perspective reveals that the limitations of conventional response distillation are not restricted to output-level knowledge transfer but also arise from its insufficient ability to preserve hierarchical feature structures, class–neighborhood topology, and decision-boundary consistency under underwater-induced representation drift. First, we formulate incremental underwater object detection as a representation-structure stability problem under degraded, category-scarce, long-tailed, and non-stationary marine distributions. Second, we propose a unified framework that integrates adaptive representation routing, topology-preserved semantic replay, and uncertainty-guided optimization to jointly regulate old-knowledge preservation and new-category adaptation. Third, experiments on UTDAC2020 and DUO demonstrate the effectiveness and limitations of the proposed framework under different incremental settings.
- (2) We propose a unified structure-stable incremental detection framework that integrates Structure-Adaptive Residual Routing (SARR), Topology-Aware Semantic Replay (TASR), and Uncertainty-Aware Plasticity–Stability Feedback (UPSF). Instead of relying on rigid backbone updating, random replay, or fixed distillation weights, the proposed framework jointly regulates layer-wise representation evolution, topology-preserved semantic rehearsal, and uncertainty-guided adaptive optimization, enabling the detector to retain old-category stability while acquiring new marine concepts with controlled plasticity.
- (3) We conduct extensive experiments on UTDAC2020 and DUO to evaluate the proposed framework under different incremental settings. The results show that our method achieves clear improvements in most scenarios and remains competitive in challenging cases. Class-wise comparisons, qualitative visualizations, and ablation studies further demonstrate the effectiveness of the proposed components in

improving old-knowledge retention, new-class adaptation, rare-class robustness, and practical detection reliability in complex underwater environments.

2. Related Works

2.1. Incremental Object Detection

Incremental object detection (IOD) aims to extend a detector to newly arriving categories while retaining the recognition and localization capabilities learned from previous categories. Compared with class-incremental classification, IOD is more challenging because the model must simultaneously preserve category discrimination, bounding-box regression, foreground–background separation, and proposal-level localization knowledge [4,5]. Early continual learning studies mainly addressed catastrophic forgetting through regularization, rehearsal, or classifier rebalancing strategies [4,6]. When extended to object detection, these strategies require detector-specific adaptation because old objects may be unlabeled in new-task images and background regions may contain categories learned in previous stages. IncDet introduced elastic weight consolidation into incremental detection and demonstrated the importance of protecting important parameters during detector adaptation [7]. ERD further showed that response-level knowledge from classification and regression heads is highly valuable for IOD and proposed elastic response selection to distill informative prediction responses from the old detector [8].

Recent IOD research has moved beyond simple output matching toward more structured knowledge preservation. DMD analyzed catastrophic forgetting as the destruction of within-class consistency and between-class discriminability and introduced separate distillation constraints to preserve feature separability [9]. BPF identified the information asymmetry among past, present, and future stages in IOD, arguing that inconsistent foreground definitions across stages can mislead optimization; it therefore aligns models across stages and leverages future-aware distillation to improve stability and plasticity [10]. GDA further investigated gradient conflicts in IOD and decomposed gradients into components related to old and new knowledge, highlighting that the direction of parameter updates is critical for balancing retention and adaptation [11]. Replay-based approaches provide another important solution. ABR revealed the foreground shift problem in image replay and proposed augmented box replay to store and replay foreground objects rather than full images [12]. PCR used proxy-based contrastive replay to mitigate imbalance and stabilize online class-incremental learning [13]. These studies indicate that IOD is not merely a problem of storing old samples or matching logits; it requires preserving the structural relationship between old and new knowledge under limited supervision. Transformer-based and vision–language-based IOD methods have recently attracted increasing attention. CL-DETR showed that directly applying knowledge distillation and exemplar replay to detection transformers is ineffective and proposed detector-specific knowledge distillation and replay calibration for transformer-based IOD [14]. Vision–language models are used to validate pseudo labels for class-incremental object detection [15], while GCD advances vision–language detectors for IOD through global alignment and correspondence distillation, preserving both global semantic structures and local topological relationships [16]. Open-set vision–language detectors such as Grounding DINO further demonstrate the potential of language-guided object detection for novel categories [17].

A key challenge in incremental learning is how to adapt a model to new tasks without unnecessarily modifying parameters that encode reusable knowledge. Adapter-based learning introduces small trainable modules into a frozen backbone, enabling new tasks to be learned with limited additional parameters while preserving shared representations [18]. LoRA freezes pretrained weights and injects trainable low-rank matrices into network layers, substantially reducing the number of updated parameters during adaptation [19].

Prompt-based continual learning methods learn task-adaptive prompts to retrieve useful knowledge without requiring rehearsal buffers or explicit task identity at inference time [20]. In vision models, VPT introduces trainable visual prompts into the input space of pretrained vision transformers, while AdaptFormer inserts lightweight adaptation modules into transformer blocks for scalable visual recognition [21,22]. These works demonstrate that full fine-tuning is not always necessary and that controlled parameter updates can preserve general representations while enabling task-specific adaptation.

Although parameter-efficient methods have been widely studied in classification, natural language processing, and vision transformer adaptation, their direct use in underwater IOD remains insufficient. Underwater incremental detection requires not only reducing trainable parameters but also deciding where and how new knowledge should be injected into the detection backbone. Low-level layers encode generic optical cues such as edges, textures, color attenuation patterns, and local contrast, which are easily corrupted by underwater degradation and should remain relatively stable. High-level layers encode category semantics and decision boundaries, which require stronger plasticity when new marine species appear. A rigid all-or-nothing strategy, such as freezing early layers and updating later layers, cannot fully handle this heterogeneous stability–plasticity demand. Dynamic routing mechanisms provide a conceptual basis for adaptive information flow by assigning representations to different transformation paths according to input- or task-dependent states [23]. Inspired by adapter learning, parameter-efficient tuning, and dynamic routing, SARR transforms fixed layer freezing into continuous incremental representation routing. Each backbone stage contains a stable anchor path and a lightweight residual adaptation path, and task-aware gates regulate the residual update strength according to feature drift, gradient conflict, and semantic uncertainty. This design enables parameter-efficient continual tuning while explicitly maintaining the hierarchical structure of underwater visual representations.

2.2. Underwater Object Detection

Underwater object detection (UOD) is essential for marine monitoring, underwater robotics, ecological observation, and autonomous intervention. However, underwater images differ substantially from natural images due to wavelength-dependent light attenuation, scattering, turbidity, non-uniform illumination, low contrast, background clutter, and small target size [1]. Generic detection frameworks usually rely on convolutional backbones such as ResNet [24] and feature pyramid necks such as FPN [25] to extract multi-scale visual representations. However, these components were mainly developed for natural-image detection, and their performance may degrade in underwater scenes because discriminative textures and object boundaries are often weakened by turbidity, scattering, and low contrast. SWIPENET addressed noisy underwater scenes by designing a sample-weighted hyper network and a curriculum multi-class AdaBoost training paradigm, emphasizing the importance of robust learning from noisy and small-object underwater data. The DUO benchmark consolidated multiple underwater datasets and provided a more standardized evaluation protocol for underwater object detection, highlighting practical issues such as annotation inconsistency, small targets, and embedded deployment requirements. Recent surveys also indicate that UOD remains challenging because underwater visual degradation directly affects feature extraction, localization accuracy, and detector robustness.

Despite these advances, most underwater detection methods assume a closed-set and offline training paradigm. They aim to improve accuracy on fixed datasets but do not consider the long-term deployment scenario in which an AUV continuously encounters new species, novel appearances, and changing imaging environments. In such an

incremental setting, underwater degradation not only reduces instantaneous detection accuracy but also accelerates representation drift: old class features may be distorted by new degraded observations, rare categories may lose their neighborhood relations in the feature manifold, and foreground decision boundaries may be overwhelmed by dominant backgrounds. This issue is related to relational and topology-preserving knowledge transfer. Relational knowledge distillation transfers mutual relations among examples rather than only matching individual outputs [26], while centroid- and proxy-based rehearsal strategies show that preserving class prototypes or feature-space relations can mitigate semantic drift and sample imbalance in continual learning [13,27]. However, these ideas have not been sufficiently integrated into underwater IOD.

Therefore, underwater IOD requires a domain-specific formulation that treats continual detection as a structure-stability problem under visually degraded, long-tailed, and background-dominated distributions. Existing generic IOD methods mainly focus on output distillation, replay calibration, or query-level transfer, while existing underwater detectors mainly focus on robust feature extraction in static training settings. In contrast, our work bridges these two research lines by explicitly regulating representation routing, Topology-Aware Semantic Replay, and Uncertainty-Aware Plasticity–Stability Feedback. This provides a unified framework for preserving hierarchical underwater visual representations while enabling controlled adaptation to new marine categories.

3. Method

3.1. Motivation and Overview

The goal of underwater IOD is to learn new marine categories from an incremental stream while preserving old recognition and localization knowledge. Existing response-based methods are effective in transferring prediction-level reasoning but do not sufficiently regulate the structural evolution of the detector representation. We therefore propose a unified framework consisting of three coupled mechanisms. SARR controls how hierarchical features are updated through adaptive residual routing. TASR maintains the topology of old and new categories by replaying prototype-sensitive boundary anchors and imposing topology-preserved distillation. UPSF adjusts the stability–plasticity trade-off according to old-class uncertainty, new-class learning difficulty, and localization uncertainty. The overall pipeline is shown in Figure 1.

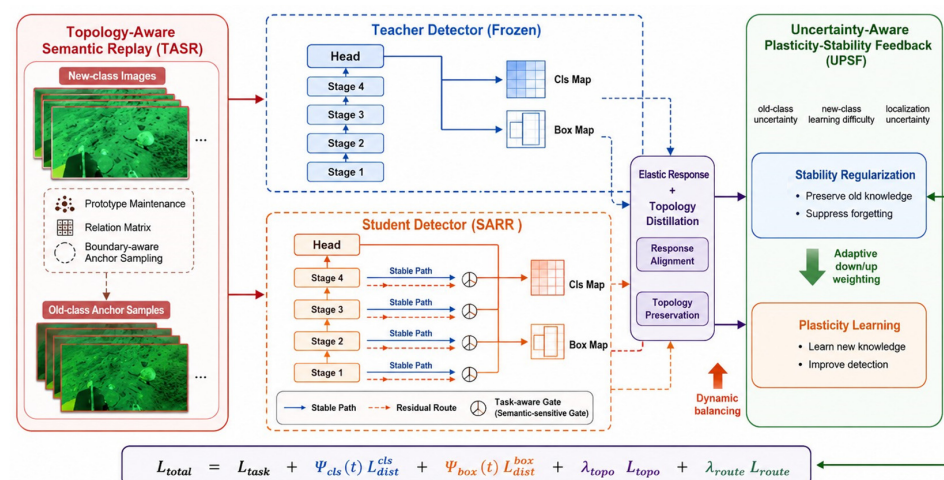


Figure 1. Overview of the proposed structure-stable underwater IOD framework. The framework jointly regulates representation routing, topology-aware replay, and uncertainty-aware feedback.

3.2. Overall Optimization Objective

Given the incremental batch constructed by TASR, the student detector is optimized under task supervision, response distillation, topology-preserved distillation, and route regularization. The total objective is formulated as

$$L_{total} = L_{task} + \Psi_{cls}(t)L_{dist}^{cls} + \Psi_{box}(t)L_{dist}^{box} + \lambda_{topo}L_{topo} + \lambda_{route}L_{route} \quad (1)$$

where L_{task} optimizes detection on current incremental samples, L_{dist}^{cls} and L_{dist}^{box} preserve classification and localization responses of old categories, L_{topo} constrains the teacher–student class topology, and L_{route} regularizes the adaptive routing behavior. $\Psi_{cls}(t)$ and $\Psi_{box}(t)$ are uncertainty-aware feedback weights generated by UPSF.

3.3. Structure-Adaptive Residual Routing

Conventional incremental detectors usually adopt an all-or-nothing update strategy for the backbone, either freezing most feature extraction layers to preserve old knowledge or fine-tuning all parameters to adapt to new categories. Such a rigid strategy is suboptimal for underwater IOD, where different layers exhibit distinct stability–plasticity requirements. Inspired by adapter learning, parameter-efficient tuning, and dynamic routing [18–23], SARR equips each backbone stage with a stable anchor path and a lightweight residual adaptation path, allowing new knowledge to be injected through gated residual routes rather than unrestricted backbone updates, as depicted in Figure 2.

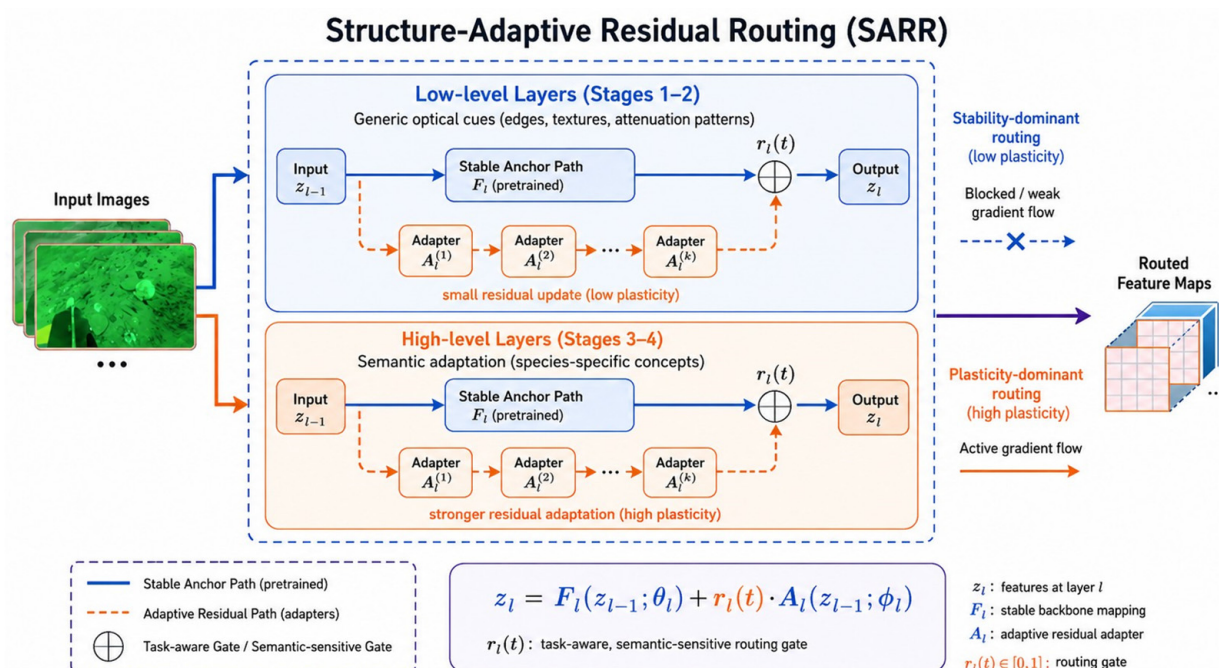


Figure 2. Illustration of Structure-Adaptive Residual Routing (SARR). Each backbone stage contains a stable anchor path, a lightweight residual adapter, and a dynamic routing gate.

To address this issue, SARR upgrades binary layer freezing into continuous incremental representation routing. Each backbone stage contains a stable anchor path and an adaptive residual path. The anchor path preserves previously learned underwater visual representations, while the residual path injects task-specific incremental knowledge through lightweight adapters. For the l -th layer, the routed representation is defined as

$$z_l = F_l(z_{l-1}; \theta_l) + r_l(t) A_l(z_{l-1}; \phi_l) \quad (2)$$

where F_l denotes the original backbone transformation, A_l denotes a lightweight residual adapter, and $r_l(t)$ in $[0,1]$ is a dynamic routing gate. The original backbone parameters θ_l are frozen or weakly updated, while only the residual adapter parameters ϕ_l and routing gates are optimized during incremental learning. The gate is estimated from old-representation drift, gradient conflict, new-category uncertainty, and layer-wise stability importance:

$$r_l(t) = \sigma(\beta_0 + \beta_u U_l(t) - \beta_d D_l(t) - \beta_c C_l(t) - \beta_w \Omega_l) \quad (3)$$

Here, $\sigma(\cdot)$ denotes the sigmoid function, which maps the routing score into $[0,1]$. The coefficients β_0 , β_u , β_d , β_c , and β_w are scalar hyperparameters selected according to validation performance rather than independently learned network parameters. The sign of each term is determined by its role in the stability–plasticity trade-off. A larger semantic uncertainty $U_l(t)$ indicates that the student detector still has difficulty learning new-category semantics and therefore increases the residual routing strength. In contrast, larger teacher–student feature drift $D_l(t)$, gradient conflict $C_l(t)$, or layer-wise stability importance Ω_l indicates a higher risk of old-knowledge erosion and therefore suppresses excessive residual adaptation. In this way, the routing gate adaptively controls the amount of incremental knowledge injected into each backbone stage.

$$C_l(t) = \max(0, -\cos(\nabla_{\phi_l} \mathcal{L}_{task}, \nabla_{\phi_l} \mathcal{L}_{dist})), \quad (4)$$

where the gradients are computed with respect to the residual adapter parameters ϕ_l , because the residual adapters are the main trainable components in SARR.

The route regularization term penalizes unnecessary updates in important layers and smooths gate variation across iterations:

$$L_{route} = \sum_l r_l(t) \Omega_l + \lambda_s \sum_l |r_l(t) - r_l(t-1)| \quad (5)$$

It is worth noting that SARR is designed as a parameter-efficient adaptation mechanism rather than a full backbone retraining strategy. During incremental learning, the teacher detector is frozen and only provides response supervision, feature references, and uncertainty estimates. For the student detector, the original backbone transformation is used as the stable anchor path, while lightweight residual adapters and routing gates are introduced as the main trainable components for incremental adaptation. In practice, the backbone anchor path can be kept fixed or updated with a much smaller learning rate than the residual adapters, depending on the deployment requirement. No additional detection head is introduced by SARR; the original classification and regression heads of the baseline detector are retained. Therefore, SARR only adds a small number of trainable parameters through residual adaptation paths and task-aware gates, making it more suitable for resource-constrained underwater platforms than full backbone fine-tuning.

3.4. Topology-Aware Semantic Replay

Underwater incremental replay is not only class-imbalanced but also topology-biased: old rare classes tend to lose their neighborhood relations and decision-boundary anchors in the feature manifold. This motivation is consistent with relational distillation and prototype-based rehearsal, which suggest that preserving feature-space relations can improve knowledge transfer beyond independent output matching [13,26,27]. In underwater scenes, rare marine categories are often small, low-contrast, partially occluded, and visually similar to the surrounding seabed or other species, as depicted in Figure 3.

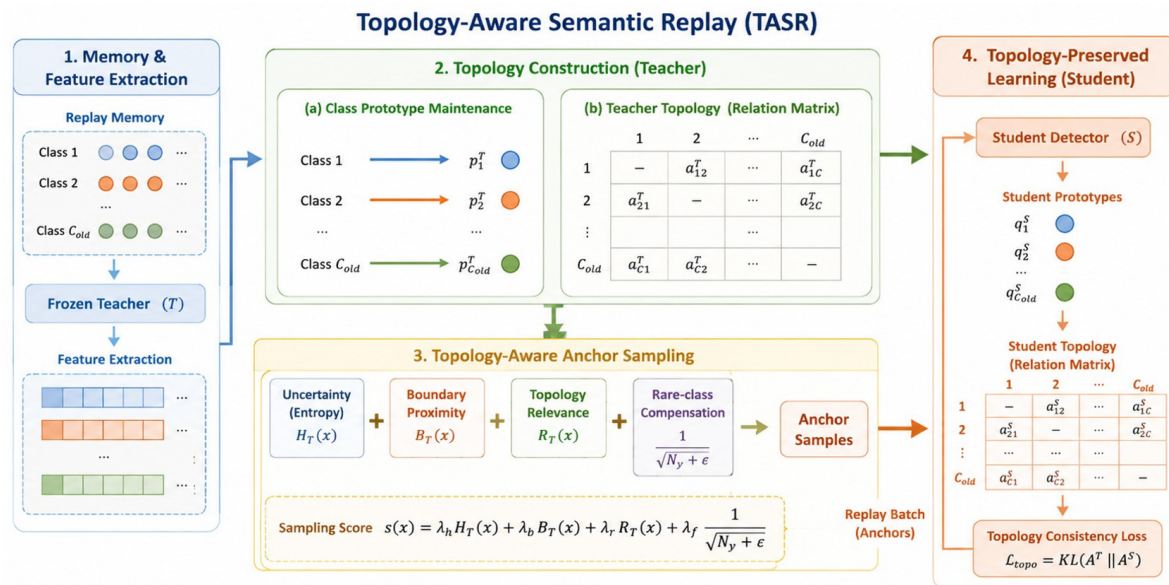


Figure 3. Illustration of Topology-Aware Semantic Replay (TASR). TASR maintains class prototypes, constructs teacher-guided relation matrices, selects topology-aware anchor samples, and imposes topology-preserved distillation to maintain class–neighborhood consistency.

For each category c , TASR maintains a class prototype p_c^T in the teacher feature space. Given region-level teacher features f_i^T from memory samples M_c , the prototype is computed and updated as

$$p_c^T = (1/|M_c|) \sum_{x_i \in M_c} f_i^T / \|f_i^T\|_2, \quad (6)$$

$$p_c^T \leftarrow \rho p_c^T + (1 - \rho) \bar{f}_c^T$$

A teacher-guided class relation matrix is then constructed from prototype similarity:

$$A_{ij}^T = \exp(-\|p_i^T - p_j^T\|_2 / \tau) / \sum_{k=1}^C \exp(-\|p_i^T - p_k^T\|_2 / \tau) \quad (7)$$

The relation matrix encodes neighborhood structure among categories. Classes with high similarity are more likely to share confusing visual patterns, and their decision boundaries should be preferentially preserved. TASR selects anchor samples using a topology-aware score:

$$s(x) = \lambda_h H_T(x) + \lambda_b B_T(x) + \lambda_r R_T(x) + \lambda_f / \sqrt{N_y + \epsilon} \quad (8)$$

where $H_T(x)$ is teacher entropy, $B_T(x)$ measures boundary proximity, $R_T(x)$ measures relevance to highly confused neighboring classes, N_y denotes the number of memory samples belonging to class y , and the last term compensates rare classes. Boundary proximity is defined as

$$B_T(x) = \exp(-|d(f^T, p_y^T) - \min_{j \neq y} d(f^T, p_j^T)|) \quad (9)$$

During student training, student prototypes and a student relation matrix A^S are computed in the same manner. The topology-preserved distillation loss is formulated as

$$L_{topo} = (1/C) \sum_{i=1}^C KL(A_i^T || A_i^S) \quad (10)$$

This loss encourages the student to preserve the teacher-guided class–neighborhood structure and prevents rare old categories from being absorbed by visually similar new or frequent classes.

TASR mainly operates during the training stage and does not change the inference architecture of the detector. Specifically, class prototypes are maintained only from replay features during incremental training, and the teacher-guided topology relation matrix is computed from frozen teacher representations. The topology-aware anchor score is used to select informative replay samples for constructing balanced and boundary-sensitive training batches. The topology-preserved loss is also applied only during optimization to constrain the student feature topology to remain consistent with the teacher topology. Once training is completed, class prototypes, topology relation matrices, anchor-score computation, and topology-preserved loss are discarded. Therefore, TASR improves replay quality and representation stability without introducing additional inference-time computation.

3.5. Uncertainty-Aware Plasticity–Stability Feedback

Incremental underwater detection requires a dynamic balance between stability and plasticity. Static distillation weights or loss-magnitude-only adjustment cannot accurately reflect the real optimization state, because old-category forgetting, new-category underfitting, and localization instability may occur simultaneously. UPSF therefore regulates the optimization according to explicit uncertainty signals.

The first signal measures old-class forgetting risk by combining normalized prediction entropy and teacher–student distillation discrepancy:

$$U_{old}(t) = \lambda_h H(p_{old}^S) / \log |C_{old}| + \lambda_k KL(p_{old}^T || p_{old}^S) \quad (11)$$

The second signal describes the learning difficulty of new categories using the normalized detection loss and the classification margin:

$$D_{new}(t) = \lambda_d L_{det}^{new}(t) / (L_{det}^{new}(0) + \epsilon) + \lambda_m [1 - (p_y^S - \max_{j \neq y} p_j^S)] \quad (12)$$

The third signal estimates localization uncertainty. For old-class replay samples, it is measured by the predicted box variance or teacher–student localization inconsistency; for new classes, it is estimated by the normalized box regression loss:

$$\begin{aligned} U_{loc}^{old}(t) &= \lambda_v \text{var}(b_{old}^S) + \lambda_i [1 - \text{IoU}(b_{old}^S, b_{old}^T)], \\ D_{loc}^{new}(t) &= L_{box}^{new}(t) / (L_{box}^{new}(0) + \epsilon) \end{aligned} \quad (13)$$

All uncertainty signals are smoothed using an exponential moving average, $Q_{bar}(t) = \gamma Q_{bar}(t-1) + (1-\gamma)Q(t)$, to avoid high-frequency oscillation. The classification and localization distillation weights are then generated as

$$\Psi_{cls}(t) = \text{clip}[\Psi_{cls}^0 \exp(\alpha_o U_{old}(t) - \beta_n D_{new}(t)), \Psi_{min}, \Psi_{max}] \quad (14)$$

$$\Psi_{box}(t) = \text{clip}[\Psi_{box}^0 \exp(\alpha_l U_{loc}^{old}(t) - \beta_l D_{loc}^{new}(t)), \Psi_{min}, \Psi_{max}] \quad (15)$$

When old-class uncertainty or localization inconsistency increases, UPSF strengthens the corresponding distillation constraint to prevent forgetting. Conversely, when new-category classification or localization difficulty remains high, UPSF relaxes the stability constraint and allocates more capacity to new-category adaptation. The clipping operation guarantees bounded feedback weights and prevents either excessive regularization or uncontrolled plasticity.

UPSF is implemented as a lightweight training-time feedback mechanism rather than an additional prediction model. The uncertainty signals used in UPSF, including old-class forgetting risk, new-class learning difficulty, and localization uncertainty, are derived from existing classification probabilities, detection losses, teacher–student response discrepancies, and bounding-box predictions. No extra uncertainty estimation network or auxiliary detection branch is introduced. To ensure stable optimization, exponential moving average is used to smooth the uncertainty signals, and clipping is applied to keep the adaptive distillation weights within a bounded range. UPSF only adjusts the weights of classification and localization distillation losses during training; it does not modify the detector architecture or increase inference complexity. Since UPSF only adjusts the weights of classification and localization distillation losses during training, it does not introduce additional prediction branches or inference-time computation.

3.6. Theoretical Discussion

This section provides theoretical insights into why the proposed framework can alleviate representation instability in underwater incremental object detection. Rather than treating catastrophic forgetting merely as an output-level mismatch, we analyze underwater IOD from the perspective of structure-stable representation evolution. In an incremental step, the student detector is optimized with new-category samples and limited replay exemplars, while the teacher detector preserves the representation structure learned from previous stages. Due to underwater visual degradation, background dominance, and long-tailed category distributions, the student representation may deviate from the teacher representation at both the layer-wise feature level and the class-topology level. We denote this structural drift as

$$\Delta S_t = \sum_l \left\| \mathbf{z}_{l,t}^S - \mathbf{z}_{l,t-1}^T \right\|_2 + \left\| \mathbf{A}_t^S - \mathbf{A}_{t-1}^T \right\|_F \quad (16)$$

where $\mathbf{z}_{l,t}^S$ denotes the student feature at layer l , $\mathbf{z}_{l,t-1}^T$ denotes the teacher feature inherited from the previous stage, and $\mathbf{A}^S$ and $\mathbf{A}^T$ are the student and teacher class-relation matrices, respectively. The first term measures layer-wise feature drift, while the second term measures topology drift in the category manifold. The objective of the proposed framework is to reduce ΔS_t while still allowing sufficient plasticity for new-category learning.

For Structure-Adaptive Residual Routing (SARR), each layer is decomposed into a stable anchor path and an adaptive residual path:

$$\mathbf{z}_l = F_l(\mathbf{z}_{l-1}; \theta_l) + r_l(t) A_l(\mathbf{z}_{l-1}; \phi_l) \quad (17)$$

where $F_l(\cdot)$ is the stable backbone transformation, $A_l(\cdot)$ is the residual adapter, and $r_l(t) \in [0, 1]$ is the task-aware and semantic-sensitive routing gate. The residual adaptation component is bounded by

$$\|\Delta \mathbf{z}_l^{res}\| \leq r_l(t) \|A_l(\mathbf{z}_{l-1}; \phi_l)\|_2. \quad (18)$$

This inequality indicates that the adaptation strength of each layer is explicitly controlled by the routing gate. When old-feature drift or gradient conflict is large, $r_l(t)$ is reduced to suppress destructive updates. When new-category uncertainty is high, $r_l(t)$ is increased to provide additional plasticity. Therefore, SARR transforms rigid layer freezing into continuous representation routing, enabling structure-preserving yet adaptive feature evolution.

For Topology-Aware Semantic Replay (TASR), the key idea is that replay should preserve not only class frequency but also class–neighborhood topology and decision-

boundary anchors. For each category, TASR maintains a teacher-guided prototype $\mathbf{p}^T$. The teacher topology is then constructed as a class-relation matrix:

$$\mathbf{A}_{ij}^T = \frac{\exp(\text{sim}(\mathbf{p}_i^T, \mathbf{p}_j^T)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(\mathbf{p}_i^T, \mathbf{p}_k^T)/\tau)}. \quad (19)$$

This matrix encodes the neighborhood structure among old and new categories in the teacher feature space. During replay, TASR prioritizes anchor samples with high uncertainty, high boundary proximity, strong topology relevance, and rare-class importance. These samples are more informative for preserving decision boundaries than randomly selected exemplars. The topology-preserved distillation loss is defined as

$$\mathcal{L}_{topo} = \frac{1}{C} \sum_{i=1}^C KL(\mathbf{A}_i^T \| \mathbf{A}_i^S) \quad (20)$$

which constrains the student relation matrix $\mathbf{A}^S$ to remain consistent with the teacher topology $\mathbf{A}^T$. By minimizing this loss, TASR reduces topology drift and prevents rare old classes from being absorbed by visually similar new or frequent categories.

For Uncertainty-Aware Plasticity–Stability Feedback (UPSF), the stability–plasticity balance is formulated as a closed-loop feedback process. Instead of using fixed distillation weights, UPSF adjusts the classification and localization distillation strengths according to old-class uncertainty, new-class learning difficulty, and localization uncertainty. The dynamic weights are computed as

$$\begin{aligned} \Psi_{cls}(t) &= \text{clip}[\Psi_{cls}^0 \exp(\alpha U_{old}(t) - \beta_n D_{new}(t)), \Psi_{min}, \Psi_{max}], \\ \Psi_{box}(t) &= \text{clip}[\Psi_{box}^0 \exp(\alpha_l U_{loc}^{old}(t) - \beta_l D_{loc}^{new}(t)), \Psi_{min}, \Psi_{max}]. \end{aligned} \quad (21)$$

where $U_{old}(t)$ is old-class forgetting risk, $D_{new}(t)$ is new-class learning difficulty, $U_{loc}^{old}(t)$ is old-class localization uncertainty, and $D_{loc}^{new}(t)$ is new-class localization difficulty. When old-class uncertainty or localization inconsistency increases, the corresponding distillation weight is enlarged to enhance stability. Conversely, when new-class learning difficulty remains high, the weight is reduced to relax old-knowledge constraints and improve plasticity. The clipping operation ensures bounded feedback weights, while exponential moving average smoothing suppresses high-frequency oscillations caused by mini-batch noise. Thus, UPSF provides a stable and interpretable optimization mechanism for dynamically regulating the trade-off between old-knowledge preservation and new-category adaptation.

Overall, SARR, TASR, and UPSF reduce structure instability from three complementary perspectives. SARR bounds layer-wise feature perturbation through adaptive residual routing, TASR preserves class–neighborhood topology through prototype-based replay and topology distillation, and UPSF dynamically controls optimization pressure through uncertainty-aware feedback. Their combination forms a unified structure-stable incremental learning mechanism for underwater object detection under degraded, imbalanced, and non-stationary visual distributions.

4. Experiments

4.1. Dataset and Implementation Details

We evaluate the proposed method on two challenging underwater object detection datasets, UTDAC2020 [28] and DUO [2], both of which contain marine organisms captured under complex underwater imaging conditions, including low contrast, turbidity, occlusion, background clutter, and scale variations. The baseline detector is GFLV1 [29], equipped with ResNet-50 [24] as the backbone and FPN [25] as the feature pyramid neck.

Following the standard COCO evaluation protocol, we report AP, AP50, AP75, APS, APm, and API to comprehensively evaluate detection accuracy across different IoU thresholds and object scales. To assess the effectiveness of the proposed framework under different incremental learning difficulties, we consider two class-incremental settings, namely 2 classes + 2 classes and 3 classes + 1 class, corresponding to relatively balanced and more extreme incremental scenarios, respectively. The 2 + 2 setting evaluates balanced incremental adaptation, whereas the 3 + 1 setting is used to examine scarce-category adaptation, where only one new class is introduced after learning multiple base classes. This design is consistent with practical underwater deployment, where newly encountered marine categories often appear sparsely and irregularly.

All experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU and Python 3.13.13. To ensure a fair comparison, all baseline methods and the proposed method are trained and evaluated under the same hardware environment, dataset splits, detector architecture, training schedule, and evaluation protocol. Unless otherwise specified, each incremental step is trained for 12 epochs using the same optimization settings as the baseline detector, as shown in Table 1. For replay-based training, the memory buffer contains 20 images/class $\times$ the number of learned classes in total. For example, under the “2 classes + 2 classes” setting, the buffer contains 40 images after the base stage and 80 images after all four classes have been learned. To reduce class imbalance, 20 images are retained for each old class after each incremental step. The memory buffer is updated using topology-aware anchor selection. All memory-based methods use the same replay memory budget for fair comparison. TASR does not increase the memory capacity; instead, it computes topology-aware anchor scores within the same buffer and selects more informative replay samples for constructing balanced and boundary-sensitive training batches.

Table 1. Replay memory setting.

Item	Setting
Total memory size	20 images/class $\times$ number of learned classes
Per-class memory	20 images/class
Memory update frequency	After each incremental stage
Memory selection strategy	Class-balanced topology-aware exemplar selection
Same memory budget for memory-based baselines	Yes
Extra memory introduced by TASR	No

The exact class orders used in this study are shown in Table 2. The 2 + 2 setting evaluates relatively balanced incremental adaptation, whereas the 3 + 1 setting evaluates scarce-category adaptation, where only one new marine class is introduced after learning multiple base classes.

Table 2. Class orders in the incremental settings.

Dataset	Setting	Base Classes	Incremental Classes
UTDAC2020	2 + 2	Echinus Starfish	Holothurian Scallop
UTDAC2020	3 + 1	Echinus Starfish Holothurian	Scallop
DUO	2 + 2	Holothurian Echinus	Scallop Starfish
DUO	3 + 1	Holothurian Echinus Scallop	Starfish

Although the current experiments evaluate two representative incremental protocols, the sensitivity to different class orders has not been fully explored. Since class order can affect the stability–plasticity balance in incremental detection, future work will evaluate the proposed framework under more class-order permutations and longer incremental sequences.

4.2. Comparison with State-of-the-Art Methods

Tables 3–7 present the quantitative results on the UTDAC2020 dataset. Under the 2 + 2 incremental setting, the proposed method achieves 30.7% AP, outperforming catastrophic forgetting by 18.9 percentage points, ERD by 2.7 percentage points, and the newly added replay-based ABR baseline by 5.6 percentage points. ABR improves the catastrophic forgetting baseline from 11.8% AP to 25.1% AP, confirming the usefulness of replay for incremental object detection. However, our method further improves AP to 30.7%, suggesting that topology-aware replay and semantic preservation are more suitable for underwater incremental detection than conventional augmented box replay. In terms of AP₅₀, our method achieves 59.7%, improving ABR by over 7.0 percentage points. Consistent gains are also observed on AP₇₅, AP_S, AP_m, and AP_l, indicating that the proposed framework improves detection robustness across different localization thresholds and object scales.

Table 3. Quantitative comparison of incremental learning performance on the UTDAC2020 dataset under the “2 classes + 2 classes” scenario. (“Δ” represents an improvement over catastrophic forgetting. “∇” represents the gap towards the upper bound.)

Scenarios	Method	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
Full data	Upper Bound	34.0	63.5	31.2	13.3	30.5	40.8
2 classes + 2 classes	Catastrophic Forgetting	11.8	20.1	12.5	3.7	7.6	13.4
	LwF [6]	11.2(Δ−0.6/∇22.8)	19.6	11.8	3.3	6.8	13.1
	ERD [8]	28.0(Δ16.2/∇6.0)	53.8	25.6	10.0	26.1	33.1
	ABR [12]	25.1(Δ13.3/∇8.9)	52.7	22.8	9.8	22.9	29.4
	CL-DETR [14]	23.3(Δ11.5/∇10.7)	52.1	15.6	9.6	20.5	27.0
	Ours	30.7(Δ18.9/∇3.3)	59.7	27.2	11.4	26.8	36.1

Table 4. Detailed class-wise AP₅₀ performance comparison on the UTDAC2020 dataset under the “2 classes + 2 classes” scenario.

Scenarios	Method	Echinus	Starfish	Holothurian	Scallop
Full data	Upper Bound	85.6	67.3	43.5	58.1
2 classes + 2 classes	Catastrophic Forgetting	19.8	38.2	/	/
	LwF [6]	22.8	20.6	16.3	18.5
	ERD [8]	79.0	42.7	44.4	49.2
	CL-DETR [14]	80.9	45.3	39.5	43.1
	Ours	82.2	62.8	40.2	53.4

More importantly, the performance gap to the upper bound is reduced to only 3.3 percentage points, indicating that the proposed structure-stable framework can effectively mitigate incremental degradation in visually complex underwater scenes. Under the more challenging 3 + 1 setting, where only one new category is introduced after learning three base classes, our method obtains 31.0% AP and achieves clear improvements in most incremental settings while remaining competitive in the most challenging

DUO 3 + 1 setting. This demonstrates that the proposed framework remains robust under highly imbalanced incremental conditions, where the model must acquire a scarce new category while retaining previously learned knowledge.

Table 5. Quantitative comparison of incremental learning performance on the UTDAC2020 dataset under the “3 classes + 1 class” scenario. (“ Δ ” represents an improvement over catastrophic forgetting. “ ∇ ” represents the gap towards the upper bound.)

Scenarios	Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Full data	Upper Bound	34.0	63.5	31.2	13.3	30.5	40.8
3 classes + 1 class	Catastrophic Forgetting	8.4	13.8	8.7	1.7	5.3	10.3
	LwF [6]	13.8(Δ 5.4/ ∇ 20.2)	26.4	13.4	4.6	9.2	13.4
	ERD [8]	28.7(Δ 20.3/ ∇ 5.3)	55.8	25.7	9.7	24.7	35.1
	CL-DETR [14]	25.2(Δ 16.8/ ∇ 8.8)	55.5	18.0	9.8	23.4	28.3
	Ours	31.0(Δ 22.6/ ∇ 3.0)	60.2	27.5	11.4	27.1	36.5

Table 6. Detailed class-wise AP_{50} performance comparison on the UTDAC2020 dataset under the “3 classes + 1 class” scenario.

Scenarios	Method	Echinus	Starfish	Holothurian	Scallop
Full data	Upper Bound	85.6	67.3	43.5	58.1
3 classes + 1 class	Catastrophic Forgetting	26.8	11.0	4.8	/
	LwF [6]	35.8	29.7	16.9	23.5
	ERD [8]	81.0	56.0	44.8	41.5
	CL-DETR [14]	81.3	55.7	44.6	40.9
	Ours	82.3	63.0	40.9	54.6

Table 7. Quantitative comparison of incremental learning performance on the DUO dataset under the “2 classes + 2 classes” scenario. (“ Δ ” represents an improvement over catastrophic forgetting. “ ∇ ” represents the gap towards the upper bound.)

Scenarios	Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Full data	Upper Bound	53.3	72.9	60.3	38.0	53.2	54.9
2 classes + 2 classes	Catastrophic Forgetting	32.9	40.7	38.9	31.4	31.9	35.7
	LwF [6]	28.2(Δ −4.7/ ∇ 25.1)	43.9	31.4	18.5	28.1	28.8
	ERD [8]	44.9(Δ 12.0/ ∇ 8.4)	63.8	50.5	29.0	42.8	50.1
	CL-DETR [14]	35.2(Δ 2.3/ ∇ 18.1)	58.3	38.7	25.7	35.5	37.2
	Ours	47.3(Δ 14.4/ ∇ 6.0)	66.1	53.6	28.5	45.4	51.9

The class-wise AP_{50} results further reveal the stability–plasticity advantage of the proposed method. In the UTDAC2020 2 + 2 setting, the AP_{50} of starfish reaches 62.8%, substantially higher than ERD (42.7%), while the performance on the old class echinus remains stable. This suggests that the proposed representation routing and topology-aware replay strategies improve the learning of visually ambiguous categories without causing severe old-class degradation. In the 3 + 1 setting, the newly introduced scallop class achieves 54.6% AP_{50} , compared with 41.5% for ERD. This improvement confirms that the proposed method can enhance new-category plasticity while maintaining competitive

stability on old categories, which is essential for long-term underwater perception where rare marine organisms may appear incrementally.

Tables 8–10 report the results on the DUO dataset, further validating the generalization ability of the proposed framework across different underwater data distributions. In the 2 + 2 setting, our method achieves 47.3% AP, exceeding ERD by 2.4 percentage points and CL-DETR by 12.1 percentage points. The improvement is particularly significant for the tail class scallop, where our method reaches 41.5% AP50, 8.0 percentage points higher than ERD. This result indicates that Topology-Aware Semantic Replay is effective in preserving decision-boundary anchors and improving rare-class recognition under imbalanced underwater distributions. It should be noted that the proposed method does not consistently improve APS in all settings. For example, in the DUO 2 + 2 setting, APS is slightly lower than ERD, although the proposed method improves AP, AP50, AP75, APS, APM, and APL. This indicates that the proposed framework mainly improves representation stability and semantic adaptation, whereas extremely small objects under severe degradation remain difficult due to weak boundaries and insufficient visual evidence.

Table 8. Detailed class-wise AP_{50} performance comparison on the DUO dataset under the “2 classes + 2 classes” scenario.

Scenarios	Method	Holothurian	Echinus	Scallop	Starfish
Full data	Upper Bound	71.0	84.9	48.7	87.1
2 classes + 2 classes	Catastrophic Forgetting	22.4	73.5	/	/
	LwF [6]	43.4	71.3	18.5	41.9
	ERD [8]	69.8	83.4	33.5	68.5
	CL-DETR [14]	62.5	80.1	29.3	61.2
	Ours	70.4	83.6	41.5	69.1

Table 9. Quantitative comparison of incremental learning performance on the DUO dataset under the “3 classes + 1 class” scenario. (“ Δ ” represents an improvement over catastrophic forgetting. “ ∇ ” represents the gap towards the upper bound.)

Scenarios	Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Full data	Upper Bound	53.3	72.9	60.3	38.0	53.2	54.9
3 classes + 1 class	Catastrophic Forgetting	28.5	39.5	32.2	23.3	28.3	28.6
	LwF [6]	31.6(Δ 3.1/ ∇ 21.7)	50.8	35.2	21.0	30.7	34.3
	ERD [8]	47.2(Δ 18.7/ ∇ 6.1)	66.8	53.4	31.8	45.3	51.4
	CL-DETR [14]	41.8(Δ 13.3/ ∇ 11.5)	61.2	47.7	25.2	40.2	46.2
	Ours	46.7(Δ 18.2/ ∇ 6.6)	67.2	52.7	32.4	45.7	50.2

Table 10. Detailed class-wise AP_{50} performance comparison on the DUO dataset under the “3 classes + 1 class” scenario.

Scenarios	Method	Holothurian	Echinus	Scallop	Starfish
Full data	Upper Bound	71.0	84.9	48.7	87.1
3 classes + 1 class	Catastrophic Forgetting	61.5	32.8	17.1	/
	LwF [6]	41.1	78.2	15.4	68.4
	ERD [8]	66.2	84.5	45.4	71.2
	CL-DETR [14]	69.1	83.6	27.7	64.3
	Ours	69.0	83.8	40.6	75.3

In the DUO 3 + 1 setting, our method obtains comparable overall AP to ERD while achieving a slightly higher AP₅₀ and stronger plasticity for the incrementally introduced starfish class. Specifically, the AP₅₀ of starfish reaches 75.3%, surpassing ERD by 4.1 percentage points. Although the overall AP is marginally lower than ERD, the higher AP₅₀ and improved new-class performance indicate that the proposed method is particularly effective in enhancing detection recall and semantic adaptation under visually degraded and imbalanced underwater scenarios. The proposed method achieves clear improvements in most settings and remains competitive in the DUO 3 + 1 setting, where it provides higher AP₅₀ and stronger new-class plasticity. The proposed framework is not a dataset-specific optimization but a general structure-stable incremental detection mechanism capable of balancing old-knowledge preservation and new-category acquisition in complex underwater environments. In the DUO 3 + 1 setting, the proposed method achieves higher AP₅₀ and stronger performance on the incrementally introduced starfish class, but its overall AP is slightly lower than ERD. This result suggests that the proposed framework improves detection recall and new-class semantic adaptation in this challenging setting, while high-IoU localization precision remains difficult under severe underwater degradation and class imbalance. Therefore, we do not claim that the proposed method uniformly outperforms all baselines in every metric. Instead, the results show that it achieves clear improvements in most incremental settings and remains competitive in the most challenging scenario.

The proposed framework is designed to improve incremental adaptability without substantially increasing inference complexity. SARR introduces lightweight residual adapters and routing gates, which add only a small number of trainable parameters compared with full backbone fine-tuning. TASR mainly operates during training by maintaining class prototypes, constructing teacher-guided relation matrices, and selecting topology-aware anchor samples from the replay buffer; therefore, it does not introduce additional inference-time computation. UPSF dynamically adjusts loss weights using uncertainty signals derived from existing classification and localization outputs, without requiring an additional prediction branch. Consequently, the main computational overhead of the proposed framework occurs during training, while the deployed detector remains close to the baseline detector in inference structure. This property is important for long-term underwater platforms with constrained onboard computation and storage resources.

4.3. Qualitative Visualization

Figure 4 presents qualitative comparisons on the UTDAC2020 dataset, and Figure 5 presents qualitative comparisons on the DUO. The baseline ERD tends to produce missed detections in challenging underwater scenes characterized by turbidity, low contrast, background clutter, and partial occlusion. This limitation becomes more evident when small marine organisms are visually blended with seabed textures or obscured by complex environmental interference. In contrast, the proposed method detects more difficult targets with improved robustness, particularly for small, camouflaged, and low-contrast objects. These visual comparisons demonstrate that the proposed framework improves detection robustness for low-contrast, camouflaged, and visually confusing objects in cluttered underwater scenes. However, extremely small targets remain challenging when their boundaries are severely weakened by turbidity, occlusion, or background interference.

4.4. Ablation Study

To verify the contributions of SARR, TASR, and UPSF, we perform ablation studies on UTDAC2020 [28] under the 2 + 2 setting. As shown in Tables 11 and 12, each component contributes to performance improvement. SARR provides the structural basis for preserving stable underwater features while allowing for semantic adaptation. TASR improves sample

representativeness by emphasizing rare and hard boundary anchors. UPSF further resolves the optimization conflict between old-knowledge preservation and new-category learning by dynamically regulating distillation strength. The full framework achieves the best AP of 30.7% and the highest AP_{50} of 59.7%.

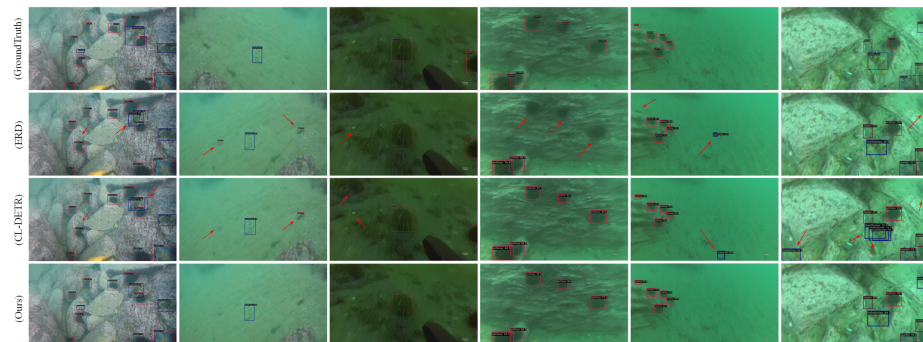


Figure 4. Qualitative comparison on UTDAC2020. From top to bottom: ground truth, ERD, CL-DETR, and ours. Red arrows indicate missed detections by baseline methods.

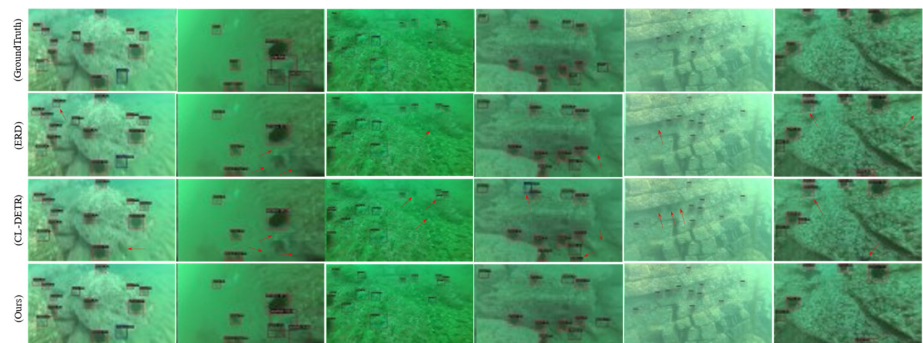


Figure 5. Qualitative comparison on DUO. The proposed method improves detection of small, occluded, and low-contrast targets compared with ERD and CL-DETR.

Table 11. Ablation study of the proposed components on the UTDAC2020 dataset under the “2 classes + 2 classes” scenario.

Method	SARR	TASR	UPSF	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Upper Bound				34.0	63.5	31.2	13.3	30.5	40.8
Catastrophic Forgetting				11.8	20.1	12.5	3.7	7.6	13.4
EXP1	✓			29.0	56.0	26.5	10.0	24.5	35.3
EXP2		✓		26.0	51.8	22.7	8.5	20.8	31.7
EXP3			✓	26.2	51.9	23.1	10.4	23.1	30.9
EXP4	✓	✓		27.3	54.9	23.2	8.8	23.3	32.2
EXP5	✓		✓	29.2	56.8	26.4	10.8	24.9	35.2
EXP6	✓	✓	✓	30.7	59.7	27.2	11.4	26.8	36.1

To verify the individual and combined contributions of SARR, TASR, and UPSF, we conduct ablation studies on the UTDAC2020 dataset under the “2 classes + 2 classes” incremental setting. The experimental settings in Tables 11 and 12 are defined as follows: EXP1 uses only SARR to evaluate the effect of selective semantic adaptation with low-level feature preservation. EXP2 uses only TASR to evaluate the contribution of topology-aware replay sampling. EXP3 uses only UPSF to examine whether uncertainty-guided semantic preservation can alleviate optimization conflicts during incremental learning. EXP4 combines SARR and TASR to test whether structural adaptation and replay construction are complementary. EXP5 combines SARR and UPSF to analyze the interaction between

selective parameter updating and uncertainty-guided loss regulation. EXP6 integrates all three components and represents the complete proposed framework.

Table 12. Class-wise AP_{50} analysis for the ablation study components on the UTDAC2020 dataset under the “2 classes + 2 classes” scenario.

Method	SARR	TASR	UPSF	Echinus	Starfish	Holothurian	Scallop
Upper Bound				85.6	67.3	43.5	58.1
Catastrophic Forgetting				19.8	38.2	/	/
EXP1	✓			81.2	57.8	36.3	48.8
EXP2		✓		79.6	51.8	35.9	39.9
EXP3			✓	81.8	63.3	22.5	40.0
EXP4	✓	✓		80.7	61.3	37.9	39.7
EXP5	✓		✓	81.1	58.5	38.0	49.5
EXP6	✓	✓	✓	82.2	62.8	40.2	53.4

As shown in Table 11, each component contributes to incremental detection performance. When used alone, SARR achieves an AP of 29.0%, outperforming TASR and UPSF, which obtain 26.0% and 26.2%, respectively. This indicates that preserving low-level underwater visual representations while updating high-level semantic layers provides an important structural basis for mitigating catastrophic forgetting. TASR and UPSF also improve the catastrophic forgetting baseline, but their individual gains are limited without the stable feature-adaptation structure provided by SARR.

The combined variants further demonstrate the complementarity of the proposed modules. EXP4, which combines SARR and TASR, achieves 27.3% AP, while EXP5, which combines SARR and UPSF, reaches 29.2% AP. This suggests that uncertainty-guided semantic preservation is particularly effective when coupled with selective parameter updating. The complete model, EXP6, achieves the best overall performance, with 30.7% AP, 59.7% AP₅₀, 27.2% AP₇₅, 11.4% APS, 26.8% APM, and 36.1% APL. Compared with EXP1, the full model further improves AP from 29.0% to 30.7%, demonstrating that TASR and UPSF provide complementary gains beyond the SARR-based adaptation framework.

Table 12 further reports the class-wise AP₅₀ results. The complete model in EXP6 achieves the most balanced class-wise performance, with 82.2% AP₅₀ on echinus, 62.8% on starfish, 40.2% on holothurian, and 53.4% on scallop. These results indicate that SARR helps retain stable old-class representations, TASR improves the representativeness of replayed samples, and UPSF adaptively balances old-knowledge preservation and new-category learning.

The experimental results confirm three important properties. First, the proposed framework improves new-class plasticity, as indicated by the large gains on starfish in UTDAC2020 2 + 2 and scallop in UTDAC2020 3 + 1. Second, topology-aware replay is particularly beneficial for rare or visually confusing classes, as shown by the improvement in scallop on DUO 2 + 2. Third, the method remains parameter-efficient because SARR introduces lightweight residual adapters and routing gates rather than updating or duplicating the entire detector. Although TASR requires prototype and anchor-score computation during training, this overhead does not affect inference and is acceptable for offline incremental updates on resource-constrained underwater platforms.

4.5. Computational Complexity and Practical Significance

The proposed framework is designed to improve incremental adaptability while keeping the deployed detector close to the baseline architecture. SARR introduces lightweight

residual adapters and routing gates to avoid unrestricted full-backbone fine-tuning. TASR introduces additional computation only during training, mainly for prototype maintenance, teacher-guided topology construction, and topology-aware anchor sampling from the replay buffer. UPSF also operates only during training by dynamically adjusting loss weights using uncertainty signals derived from existing detector outputs. Consequently, the main computational overhead occurs in the incremental training phase, whereas the inference process remains almost unchanged compared with the baseline detector.

As shown in Table 13, ERD and our method have the same total parameters, 64.53 M, because both frameworks maintain a student detector and a frozen teacher detector during incremental training. However, our method reduces the number of trainable parameters from 32.04 M in ERD to 23.7 M, and the trainable ratio from 49.65% to 36.73%. This reduction is achieved by freezing low-level backbone stages and updating only the high-level semantic components and task-related modules. Meanwhile, the FLOPs remain 206 G for GFLV1, ERD, and our method, indicating that the proposed framework does not introduce additional inference computation.

Table 13. Computational complexity comparison.

Method	Total Params	Trainable Params	Trainable Ratio	FLOPs
GFLV1	32.263 M	32.038 M	99.3%	206 G
ERD	64.53 M	32.04 M	49.65%	206 G
Ours	64.53 M	23.7 M	36.73%	206 G

These results show that the proposed method improves incremental learning mainly through training-stage adaptation rather than by increasing inference complexity. Although SARR, TASR, and UPSF introduce additional operations during incremental training, the deployed detector keeps the same inference FLOPs as the baseline detector. Therefore, the additional computational cost mainly appears in offline incremental updates, while the inference process remains suitable for deployment-oriented underwater object detection. Nevertheless, real-time onboard AUV deployment, energy-aware model optimization, and online adaptation on embedded hardware remain topics for future works.

Hardware-level image stabilization techniques, such as gimbal stabilization, optical image stabilization, in-body image stabilization, and IMU-assisted motion compensation, can reduce motion blur, viewpoint jitter, and frame-level instability at the sensing stage. These techniques improve the quality of visual input but do not directly solve the learning-stage problems of new-category adaptation, representation drift, rare-class dilution, and old-class forgetting. The proposed framework is therefore complementary to image stabilization: stabilized sensing can provide more reliable observations, while the proposed structure-stable incremental learning framework enables the detector to adapt to newly encountered marine categories and changing underwater distributions over time.

4.6. Multiple-Run Statistics Under the Main Incremental Settings

To reduce the influence of random initialization and memory sampling, we repeat the main experiments 24 times with different random seeds and report mean $\pm$ standard deviation where applicable, as shown in Table 14.

Due to the high computational cost of exhaustive repeated incremental training, we report multiple-run statistics for the main comparison between ERD and the proposed method. Extending repeated evaluation to all baselines will be considered in future work.

Table 14. Multiple-run statistics under the main incremental settings.

Dataset	Setting	Method	AP Mean $\pm$ Std	AP50 Mean $\pm$ Std
UTDAC2020	2 + 2	ERD	27.9 $\pm$ 0.21	53.7 $\pm$ 0.62
UTDAC2020	2 + 2	Ours	30.6 $\pm$ 0.34	59.5 $\pm$ 0.83
UTDAC2020	3 + 1	ERD	28.7 $\pm$ 0.27	55.7 $\pm$ 0.54
UTDAC2020	3 + 1	Ours	30.9 $\pm$ 0.38	60.0 $\pm$ 0.91
DUO	2 + 2	ERD	44.8 $\pm$ 0.24	63.6 $\pm$ 0.57
DUO	2 + 2	Ours	47.1 $\pm$ 0.36	66.0 $\pm$ 0.78
DUO	3 + 1	ERD	47.0 $\pm$ 0.26	66.5 $\pm$ 0.46
DUO	3 + 1	Ours	46.8 $\pm$ 0.39	67.0 $\pm$ 0.86

5. Conclusions

This paper investigated incremental underwater object detection from the perspective of representation-structure stability. Unlike conventional response-based distillation methods that mainly preserve output-level predictions, the proposed framework jointly regulates layer-wise representation evolution, class-topology preservation, and uncertainty-guided optimization. Through SARR, TASR, and UPSF, the detector can better balance old-knowledge retention and new-category adaptation under degraded, imbalanced, and non-stationary underwater distributions. Experiments on UTDAC2020 and DUO demonstrate that the proposed method achieves clear improvements in most incremental settings and remains competitive in challenging cases. The results also show that high-IoU localization and extremely small-object detection under severe degradation remain difficult.

To address this challenge, we proposed a unified structure-stable incremental detection framework composed of Structure-Adaptive Residual Routing (SARR), Topology-Aware Semantic Replay (TASR), and Uncertainty-Aware Plasticity–Stability Feedback (UPSF). SARR transforms rigid layer freezing into adaptive residual representation routing, enabling controlled feature evolution across different backbone stages. TASR preserves class prototypes, neighborhood topology, and decision-boundary anchors during replay, thereby mitigating rare-class dilution and topology drift. UPSF further regulates the optimization process using explicit uncertainty signals, including old-class forgetting risk, new-class learning difficulty, and localization uncertainty. Through these complementary designs, the proposed framework provides an integrated solution for maintaining representation stability while retaining sufficient semantic plasticity in underwater incremental learning.

Extensive experiments demonstrate clear improvements over representative distillation- and transformer-based IOD baselines in most settings. With the same replay memory budget, the proposed TASR also shows better replay effectiveness than class-balanced replay, indicating the benefit of topology-aware semantic rehearsal. The quantitative results, class-wise analysis, ablation studies, and qualitative visualizations collectively verify the effectiveness of the proposed framework in complex underwater scenarios involving low contrast, occlusion, background clutter, small objects, and imbalanced category distributions. These results suggest that structure-stable representation evolution is a promising direction for long-term autonomous underwater perception. From a practical perspective, the proposed framework provides a feasible solution for long-duration underwater platforms that must update their perception capability without repeatedly retraining detectors from scratch or storing all historical data. Since the additional topology replay and uncertainty feedback modules mainly operate during training, the deployed detector can maintain a structure close to the baseline model, which is beneficial for resource-constrained AUV applications.

Despite its effectiveness, several limitations remain. Existing underwater detection datasets contain a limited number of object categories, which restricts the evaluation of

long-horizon and multi-stage incremental learning. Although the topology-preserved formulation improves representation stability, its generalization ability still needs further validation on larger-scale, open-world underwater datasets with more diverse categories and environmental conditions. We will explore feature-level generative replay and more efficient prototype updating strategies to support open-world underwater perception.

Future work will focus on four directions. First, we will evaluate the proposed framework under more class-order permutations and longer incremental sequences. Second, we will explore more efficient replay and prototype updating strategies to reduce training-time overhead. Third, we will extend the framework to larger-scale and open-world underwater datasets with more diverse marine categories. Finally, we will investigate real-time deployment on embedded underwater platforms and integration with stabilized sensing systems.

Author Contributions: Conceptualization, S.Z., H.Y. and F.T.; software, S.Z. and J.S.; methodology, S.Z., J.S., H.Y. and F.T.; validation, F.T. and Y.Y.; project administration, F.T. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China grant No. 62301302 and 62271404.

Data Availability Statement: Data and program will be made available on request.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Chen, L.; Huang, Y.; Dong, J.; Xu, Q.; Kwong, S.; Lu, H.; Lu, H.; Li, C. Underwater optical object detection in the era of artificial intelligence: Current, challenge, and future. *ACM Comput. Surv.* **2025**, *58*, 1–34. [\[CrossRef\]](#)
2. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A dataset and benchmark of underwater object detection for robot picking. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*; IEEE: New York, NY, USA, 2021; pp. 1–6.
3. Chen, L.; Zhou, F.; Wang, S.; Dong, J.; Li, N.; Ma, H.; Wang, X.; Zhou, H. SWIPENET: Object detection in noisy underwater scenes. *Pattern Recognit.* **2022**, *132*, 108926. [\[CrossRef\]](#)
4. Zhou, D.-W.; Wang, Q.-W.; Qi, Z.-H.; Ye, H.-J.; Zhan, D.-C.; Liu, Z. Class-incremental learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9851–9873. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Zhang, J.; Liu, L.; Silven, O.; Pietikainen, M.; Hu, D. Few-shot class-incremental learning for classification and object detection: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2924–2945. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Li, Z.; Hoiem, D. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2935–2947. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Liu, L.; Kuang, Z.; Chen, Y.; Xue, J.-H.; Yang, W.; Zhang, W. IncDet: In defense of elastic weight consolidation for incremental object detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2306–2319. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Feng, T.; Wang, M.; Yuan, H. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022; pp. 9427–9436.
9. Kang, M.; Zhang, J.; Zhang, J.; Wang, X.; Chen, Y.; Ma, Z.; Huang, X. Alleviating catastrophic forgetting of incremental object detection via within-class and between-class knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 1–6 October 2023; pp. 18894–18904.
10. Mo, Q.; Gao, Y.; Fu, S.; Yan, J.; Wu, A.; Zheng, W.-S. Bridge past and future: Overcoming information asymmetry in incremental object detection. In *Computer Vision-ECCV 2024*; Springer: Cham, Switzerland, 2025; pp. 463–480.
11. Luo, W.; Zhang, S.; Cheng, D.; Xing, Y.; Liang, G.; Wang, P.; Zhang, Y. Gradient decomposition and alignment for incremental object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, HI, USA, 19–25 October 2025; pp. 4486–4495.
12. Liu, Y.; Cong, Y.; Goswami, D.; Liu, X.; van de Weijer, J. Augmented box replay: Overcoming foreground shift for incremental object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 1–6 October 2023; pp. 11367–11377.

13. Lin, H.; Zhang, B.; Feng, S.; Li, X.; Ye, Y. PCR: Proxy-based contrastive replay for online class-incremental continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 24246–24255.
14. Liu, Y.; Schiele, B.; Vedaldi, A.; Rupprecht, C. Continual detection transformer for incremental object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 23799–23808.
15. Kim, J.; Ku, Y.; Kim, J.; Cha, J.; Baek, S. VLM-PL: Advanced pseudo labeling approach for class incremental object detection via vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 16–22 June 2024; pp. 4170–4181.
16. Wang, X.; Wang, Z.; Lin, Z. GCD: Advancing vision-language models for incremental object detection via global alignment and correspondence distillation. *AAAI Conf. Artif. Intell.* **2025**, *39*, 8015–8023. [\[CrossRef\]](#)
17. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision-ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 38–55.
18. Hounsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
19. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022; Poster.
20. Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; Pfister, T. Learning to prompt for continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 139–149.
21. Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.-N. Visual prompt tuning. In *Computer Vision-ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 709–727.
22. Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2022**, *35*, 16664–16678. [\[CrossRef\]](#)
23. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2017**, *30*, 3859–3869.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
26. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3967–3976.
27. Liu, D.; Lyu, F.; Li, L.; Xia, Z.; Hu, F. Centroid distance distillation for effective rehearsal in continual learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2023; pp. 1–5.
28. *SRMIST, UTDAC2020 Dataset*; Roboflow Universe: Des Moines, IA, USA, 2023.
29. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2020**, *33*, 21002–21012.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.