Query as a Resource: Activity-Cost Guided Remote Sensing Domain-Incremental Object Detection

Xi Yang Zhengyu Yuan De Cheng* Nannan Wang
the State Key Laboratory of Integrated Services Networks
School of Telecommunications Engineering
Xidian University, Xi'an, China
yangx@xidian.edu.cn, yuanzhy@stu.xidian.edu.cn
dcheng@xidian.edu.cn, nnwang@xidian.edu.cn

Abstract

Remote sensing object detection often proceeds in a domain-incremental manner, where detectors continuously encounter new domains arising from changes in regions, resolutions, sensors, and modalities. This challenge is particularly severe when passive optical red-green-blue (RGB) imagery and active synthetic aperture radar (SAR) imagery coexist, because their distinct imaging mechanisms create large domain gaps. Existing methods mainly rely on feature alignment or regularization, but often overlook how object-level semantic representations are preserved across heterogeneous domains. For DETR-like detectors, cross-domain interference can additionally disrupt decoder queries that serve as object-centric semantic slots. Under a **query-as-resource** view, we propose **Q-cost** for remote sensing domain-incremental object detection. Q-cost contains three explicitly connected modules. The Prototype-Driven Feature Homogenization Adapter (PFHA) uses modality- and domain-aware prototypes to guide dual-level mixture-of-experts feature adaptation. The Domain-Context Prompted Encoder (DCPE) transforms the same prototypes into encoder prompts for domain-conditioned semantic aggregation. The Domain-Aware Query Activity Regulator (DQAR) estimates current and historical query activity, models cumulative activity cost, and gates query updates to balance adaptation and retention. Experiments on multiple remote sensing benchmarks show that Q-cost improves new-domain adaptation and old-domain retention under severe domain and modality shifts. Code is available at https://github.com/XI-Lab-XDU/Q-cost.

1 Introduction

Remote sensing object detection supports long-term Earth observation systems for applications such as disaster assessment, maritime surveillance, and land monitoring [4, 45, 17]. In real-world deployment, data are rarely drawn from a fixed distribution [43, 46]. Instead, detectors continuously encounter new domains arising from variations in regions, resolutions, acquisition conditions, sensors, and modalities [25, 24, 11]. This leads to a domain-incremental learning scenario, where a model must adapt to each new domain while preserving knowledge of previous domains under a fixed label space [36]. The challenge becomes more severe when the data stream involves heterogeneous sensing mechanisms. Passive optical imagery emphasizes texture and color, whereas active synthetic aperture radar (SAR) imagery captures scattering and geometric responses [27]. As illustrated in Fig. 1(a), such modality-dependent characteristics introduce larger distribution gaps and more destructive parameter drift than intra-modality shifts [31]. Nevertheless, objects retain shared semantic identities across domains, providing a basis for knowledge transfer, as shown in Fig. 1(b) [5].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

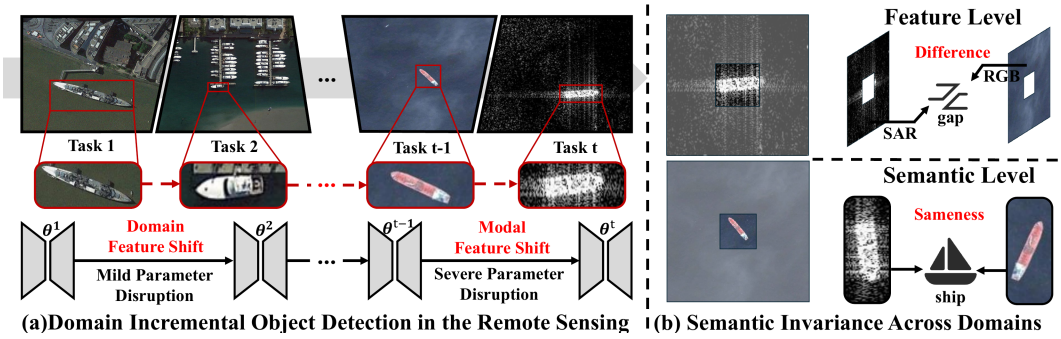


Figure 1: Motivation of domain-incremental object detection in remote sensing. (a) Introducing cross-modality domains leads to a much larger domain gap and more destructive parameter drift than within-modality shifts. (b) Despite severe feature-level discrepancies across modalities, objects remain semantically shared, providing a basis for cross-domain knowledge transfer.

Although domain-incremental learning has been widely studied, most existing settings involve a limited number of domains or relatively mild shifts [7, 32]. Detection-oriented methods mitigate forgetting through regularization, replay, adapters, or domain-specific designs, but they typically assume shared visual statistics or a common imaging modality [30, 29, 23]. This assumption becomes unreliable for heterogeneous optical and SAR streams, where adapting to a new domain can overwrite knowledge acquired from earlier feature distributions and object appearances. For DETR-like detectors, this cross-domain interference also acts on the shared decoder queries, which serve as object-centric semantic slots and are repeatedly updated across domains [1, 10]. Existing methods rarely examine how the activities of these queries change, overlap, or conflict across heterogeneous domains. Our object-grounded query analysis directly verifies this detector-internal effect. Using Hungarian matching to identify the query responsible for each ground-truth object, the Query Attention Localization Ratio (QALR) decreases from 0.68 before the RGB-to-SAR shift to 0.39 after sequential DINO fine-tuning, while Q-cost retains 0.61. These observations establish query-level disruption within DINO and motivate explicit regulation of query-level semantic responses under cross-modality adaptation.

To address this problem, we propose **Q-cost**, a framework that coordinates feature-side domain conditioning and semantic-side query regulation in a unified DETR-like detection pipeline [18, 22]. Q-cost contains three modules whose names directly correspond to their functions. First, the **Prototype-Driven Feature Homogenization Adapter (PFHA)** maintains modality- and domain-aware prototypes [28, 20]. It uses them to route modality experts, fuse domain experts, and adapt backbone features through a prototype-guided dual-level mixture of experts, thereby reducing heterogeneous feature discrepancies. Second, the **Domain-Context Prompted Encoder (DCPE)** transforms the same selected modality center and fused domain prototype into an encoder prompt, which guides global semantic aggregation under the current domain context [20]. PFHA and DCPE together form the feature-side domain-conditioning pathway. Third, the **Domain-Aware Query Activity Regulator (DQAR)** forms the semantic-side pathway. It estimates the current-domain activity and relevance-weighted historical activity of every decoder query, models the remaining adaptation budget through cumulative activity cost, and applies a bounded activity gate to regulate query updates [10, 18].

Building on object-centric query representations and semantic alignment [35, 38], this design treats decoder queries as limited semantic resources rather than uniformly updated parameters. Queries that are active for the current domain but weakly occupied by historical domains are allowed to adapt more strongly. Queries important to previous domains are protected, while shared or inactive queries are updated conservatively. PFHA and DCPE reduce the feature discrepancy presented to the transformer, and DQAR then regulates how the resulting semantic evidence changes the query set. The gate connects current visual evidence to the historical role of each query in detecting previously learned objects. The three modules therefore have distinct but connected roles: PFHA adapts the backbone features, DCPE conditions encoder aggregation, and DQAR controls decoder-query optimization. Together, they improve new-domain plasticity while reducing destructive interference with historical query semantics.

Our contributions are summarized as follows:

- We introduce a cross-modality remote sensing domain-incremental object detection setting, extending conventional scenarios to heterogeneous imaging streams with large modality gaps. To our knowledge, this is the first study of cross-modality incremental detection in remote sensing, and experiments show improved new-domain adaptation with reduced forgetting on historical domains.
- We propose Q-cost, a framework that revisits continual detection from the perspective of semantic resource allocation. PFHA uses modality- and domain-aware prototypes to guide dual-level MoE feature adaptation, while DCPE converts the same prototypes into encoder prompts for domain-conditioned semantic aggregation.
- DQAR regulates decoder queries by estimating current and historical activity and modeling cumulative activity cost. Its bounded activity gate selectively adapts or preserves queries according to their roles across domains, reducing interference under large modality shifts.

2 Related Work

Remote sensing object detection. Object detection in remote sensing imagery has evolved from geometry-aware detectors tailored to aerial scenes toward Transformer-based formulations [6, 13, 47, 4]. Early progress mainly addressed oriented objects, large geometric variations, and cluttered backgrounds in bird's-eye-view imagery, where representative methods such as RoI Transformer and ReDet explicitly model rotation and orientation-sensitive features [37]. More recently, DETR reformulates detection as a set prediction problem with learnable object queries, while its variants such as Deformable DETR and DINO further improve convergence, multi-scale aggregation, and query learning [39, 18, 22]. These advances provide a strong foundation for remote sensing detection, but they are still mostly developed under static training settings where all data are jointly available [43, 14, 46]. Such an assumption is often mismatched with long-term Earth observation pipelines, where data arrive sequentially across regions, resolutions, sensors, and even sensing modalities [21, 15].

Domain-incremental learning. Incremental object detection has progressed from generic distillation- or replay-based strategies to detector-specific designs [30, 33]. RGR uses image-level generative replay to preserve knowledge during class-incremental object detection [41]. For Transformer detectors, CL-DETR adapts knowledge distillation and exemplar replay through detector-aware distillation and replay calibration [23], while LEA introduces task-specific attention modules and an energy-based task modulator to alleviate annotation conflicts in class-incremental detection. Query-oriented designs have also appeared: DyQ-DETR incrementally expands object queries and isolates matching across phases to reduce inter-class confusion [40]. However, these methods are mainly designed for category-incremental scenarios with evolving label spaces rather than fixed-label domain shifts [30]. In the domain-incremental setting, DISC performs statistical correction for different domains, and LDB learns domain-specific biases with a domain selector in a non-exemplar manner [26, 32]. More recently, Active-DDC combines active replay with query-based domain shift control for remote sensing detection [34]. Nevertheless, existing methods either rely on replay or domain-specific branches, or use queries mainly for category expansion or cross-domain alignment [23, 33]. Shared DETR queries are seldom treated as limited semantic resources under cross-modality shifts [40]. Q-cost instead regulates them using current and historical activity and cumulative activity cost.

Outside incremental detection, PhaseLoRA conditions low-rank adaptation on control regimes in vision-language-action policies [12]. Interference-factor disentanglement also supports identity robustness in cloth-changing person re-identification [3].

3 Preliminary and Problem Formulation

Problem Definition. Let $\mathcal{D}^{(t)} = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t}$ denote the dataset of the t-th incremental task, where $x_i^{(t)}$ is a remote sensing image and $y_i^{(t)}$ is its annotation. After learning task t, the detector is denoted by $f(\cdot;\theta^{(t)})$. At each step, the model is updated from $\theta^{(t-1)}$ to $\theta^{(t)}$ using only $\mathcal{D}^{(t)}$, which makes it adapt to the current domain while risking forgetting previous ones. Thus, domain-

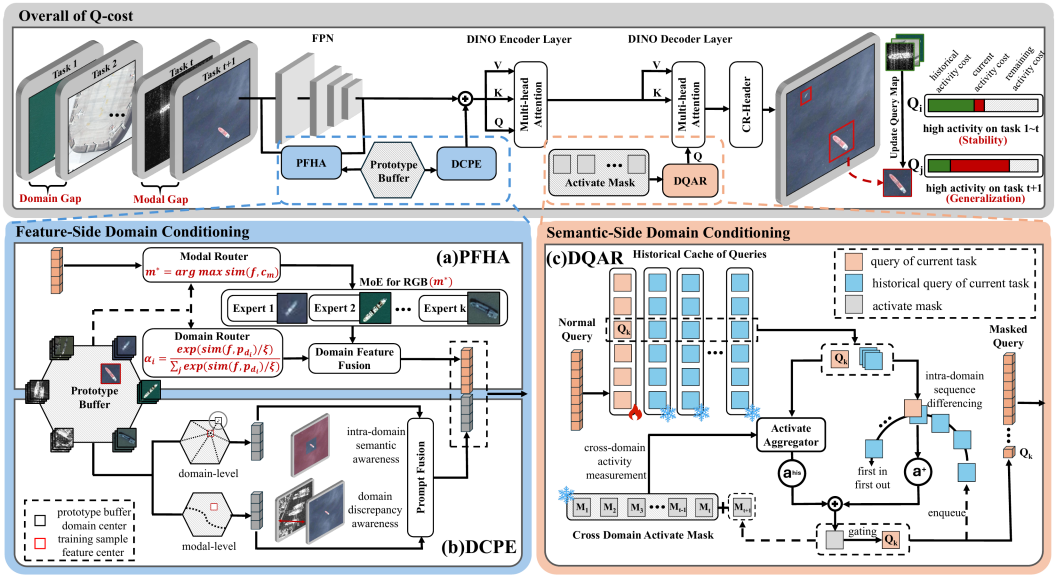


Figure 2: The overall framework of Q-cost. (a) PFHA performs prototype-guided dual-level MoE feature adaptation. (b) DCPE constructs prototype-derived prompts for domain-conditioned encoder aggregation. In the feature-side pathway, orange blocks denote PFHA-adapted feature tensors and gray blocks denote DCPE's prototype-derived prompts; the prompts are combined with adapted visual tokens before encoder processing. (c) DQAR estimates current and historical query activity, models activity cost, and gates query updates.

incremental remote sensing detection aims to maintain adaptation to new domains without sacrificing performance on historical domains under both intra- and inter-modality shifts.

DETR-like detection preliminaries. We build on DINO [39], where an input image is encoded into visual tokens $\mathbf{Z}(x)$ and decoded by a fixed set of object queries $\mathbf{Q}$ into predictions. Training uses Hungarian matching and optimizes the standard detection loss together with the denoising term:

$$\mathcal{L}_{\text{DINO}} = \mathcal{L}_{\text{det}} + \lambda_{\text{dn}}\mathcal{L}_{\text{dn}}, \tag{1}$$

where $\mathcal{L}_{\text{det}}$ combines classification, box regression, and IoU losses. In this view, each query $\mathbf{q}_k$ acts as an object-centric semantic slot, and continual learning updates the shared query set through gradients of $\mathcal{L}^{(t)}_{\text{DINO}}$, which makes query adaptation vulnerable to destructive interference across domains. This motivates explicit domain-aware query regulation.

4 Method

4.1 Overall Framework

Q-cost consists of three explicitly connected modules: PFHA implements prototype-guided dual-level MoE feature adaptation, DCPE generates prototype-derived encoder prompts, and DQAR performs query activity-cost estimation and update gating. This module map connects feature-level discrepancy reduction to query-level semantic regulation throughout the pipeline.

PFHA first reduces modality/domain-induced feature variations by projecting samples from different domains into a more homogeneous feature space. Based on the learned prototypes, DCPE further injects domain-aware guidance into the encoder, encouraging global semantic aggregation under the current domain context. Built upon these purified encoder representations, DQAR serves as the core module by explicitly modeling query activity under different domains and selectively regulating query updates. Accordingly, the overall objective at task t is

$$\mathcal{L}^{(t)}_{\text{obj}} = \mathcal{L}^{(t)}_{\text{DINO}} + \lambda_p\mathcal{L}^{(t)}_{\text{PFHA}} + \lambda_e\mathcal{L}^{(t)}_{\text{DCPE}} + \lambda_q\mathcal{L}^{(t)}_{\text{DQAR}}, \tag{2}$$

where λ_p, λ_e, and λ_q are balancing weights, and the three additional terms respectively regularize feature homogenization, prompt-guided encoder aggregation, and query preservation.

4.2 Feature-Side Domain Conditioning: PFHA and DCPE

Prototype-Driven Feature Homogenization Adapter (PFHA). To reduce feature-level domain discrepancy before query adaptation, we insert a Prototype-Driven Feature Homogenization Adapter (PFHA) in parallel with intermediate backbone stages. Given the feature at stage l, $\mathbf{F}^l \in \mathbb{R}^{N\times C}$, PFHA predicts a residual feature $\mathbf{R}^l$ and outputs $\widetilde{\mathbf{F}}^l = \mathbf{F}^l + \mathbf{R}^l$.

PFHA maintains a prototype buffer with a domain prototype $\mathbf{p}_d \in \mathbb{R}^C$ for each domain, updated by EMA from mini-batch pooled features, and derives a modality center $\mathbf{c}_m$ by averaging and normalizing the prototypes within modality m. For a sample with pooled feature $\mathbf{f}^l = \mathrm{Pool}(\mathbf{F}^l)$, PFHA first performs hard modality routing by selecting the closest modality center, and then applies a modality-specific adapter followed by a domain-fusion adapter that mixes the prototypes $\{\mathbf{p}_{d_i}\}_{i=1}^K$ under the routed modality using fusion weights $\{\alpha_i\}_{i=1}^K$. To explicitly encourage feature homogenization, we impose

$$\mathcal{L}_{\mathrm{PFHA}} = \left\|\mathrm{Norm}(\mathbf{f}^l) - \mathrm{Norm}(\mathbf{p}_d)\right\|_2^2 + \rho\left\|\mathrm{Norm}(\mathbf{f}^l) - \mathbf{c}_{m^\star}\right\|_2^2, \tag{3}$$

where ρ balances domain- and modality-level alignment. This loss is summed over inserted backbone stages and mini-batch samples.

Domain-Context Prompted Encoder (DCPE). Based on the routing result and prototype buffer in PFHA, DCPE transforms domain statistics into an encoder-side prompt. Specifically, the selected modality center and the fused domain prototypes are projected into the encoder space and combined as

$$\mathbf{e}_p = \mathbf{W}_m\mathbf{c}_{m^\star} + \mathbf{W}_d\left(\sum_{i=1}^{K}\alpha_i\mathbf{p}_{d_i}\right), \tag{4}$$

where $\mathbf{W}_m, \mathbf{W}_d \in \mathbb{R}^{C\times C}$ are learnable projections. We prepend $\mathbf{e}_p$ to the encoder token sequence $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$ to obtain the prompted input $\widetilde{\mathbf{Z}} = \{\mathbf{e}_p, \mathbf{z}_1, \ldots, \mathbf{z}_N\}$. To ensure that the prompt effectively guides encoder aggregation, we further use

$$\mathcal{L}_{\mathrm{DCPE}} = 1 - \mathrm{sim}\left(\mathrm{Pool}(\widetilde{\mathbf{Z}}), \mathbf{e}_p\right), \tag{5}$$

which aligns the prompted encoder representation with the prototype-derived domain context. In this way, DCPE bridges PFHA and DQAR by injecting lightweight domain priors without introducing additional domain labels.

4.3 Semantic-Side Query Regulation: DQAR

After feature-side conditioning, encoder tokens largely suppress low-level domain discrepancies while preserving domain-relevant semantics. We thus focus on the decoder, where the query set $\mathbf{Q} \in \mathbb{R}^{M\times C}$ carries object-level semantics and continual adaptation manifests as query updates. Instead of uniformly optimizing all queries, we regulate them with domain-conditioned activity patterns.

Query Activity Buffer. We maintain a domain-query activity buffer $\mathbf{B} \in \mathbb{R}^{T\times M}$ where T is the number of incremental domains and M is the number of decoder queries. The t-th row $\mathbf{B}_{t,:}$ records the activity profile of all queries under domain t. Given a sample x from the current domain t, let $\widetilde{\mathbf{Z}}(x) \in \mathbb{R}^{(N+1)\times C}$ denote the encoder tokens, and let

$$\mathbf{H}(x) = \Psi(\mathbf{Q}, \widetilde{\mathbf{Z}}(x)) \in \mathbb{R}^{M\times C}, \tag{6}$$

be the query response matrix, where $\Psi(\cdot,\cdot)$ denotes the decoder interaction function and the k-th row of $\mathbf{H}(x)$ corresponds to the response of the k-th query vector $\mathbf{q}_k$. DQAR operates on $\mathbf{H}(x)$ as a structured semantic response of the whole query set.

Current-Domain Activity. To estimate current-domain activity, we maintain for each domain t a queue $\mathcal{U}_t = \left\{\mathbf{H}_t^{(1)}, \ldots, \mathbf{H}_t^{(S-1)}\right\}$, where S is the queue length and each $\mathbf{H}_t^{(\tau)} \in \mathbb{R}^{M\times C}$ is a recent response matrix collected from domain t. Using a short queue instead of one-step differencing reduces response fluctuation. We first compute the queue mean

$$\bar{\mathbf{H}}_t = \frac{1}{S-1}\sum_{\tau=1}^{S-1}\mathbf{H}_t^{(\tau)}, \tag{7}$$

and define the response deviation matrix of the current sample as $\mathbf{\Delta}_t(x) = \mathbf{H}(x) - \bar{\mathbf{H}}_t \in \mathbb{R}^{M\times C}$. Let $\mathbf{\Delta}_{t,k}(x) \in \mathbb{R}^C$ denote the k-th row of $\mathbf{\Delta}_t(x)$. The current-domain activity vector is then obtained by row-wise aggregation:

$$\mathbf{a}_t^{(x)} = \mathcal{N}\left(\|\mathbf{\Delta}_{t,1}(x)\|_2, \ldots, \|\mathbf{\Delta}_{t,M}(x)\|_2\right) \in \mathbb{R}^M, \tag{8}$$

where $\mathcal{N}(\cdot)$ denotes vector normalization. The corresponding buffer row $\mathbf{B}_{t,:}$ is updated by EMA.

Historical Activity and Query Activity Cost. The historical-domain activity is computed by aggregating the activity rows of all previous domains. To avoid treating all historical domains equally, we weight them according to their similarity to the current domain condition. Recall that Section 4.2 produces a soft domain representation $\hat{\mathbf{p}}_t = \sum_{i=1}^K \alpha_i \mathbf{p}_{d_i}$, where $\{\mathbf{p}_{d_i}\}$ are the prototypes of the selected modality and $\{\alpha_i\}$ are the corresponding routing weights. We define the historical relevance weights as

$$w_{t,j} = \frac{\exp\left(\mathrm{sim}(\hat{\mathbf{p}}_t, \mathbf{p}_j)/\xi\right)}{\sum_{r=1}^{t-1}\exp\left(\mathrm{sim}(\hat{\mathbf{p}}_t, \mathbf{p}_r)/\xi\right)}, \quad j < t, \tag{9}$$

where $\mathbf{p}_j$ is the prototype of the j-th historical domain and $\xi > 0$ is a temperature parameter. Let $\mathbf{w}_t = [w_{t,1}, \ldots, w_{t,t-1}]^\top \in \mathbb{R}^{t-1}$ collect these relevance weights, and aggregate historical activity in matrix form

$$\mathbf{a}_t^{\text{his}} = \mathbf{w}_t^\top \mathbf{B}_{1:t-1}. \tag{10}$$

To further model the limited adaptability of each query, we introduce the notion of query activity cost. Instead of viewing queries as passive parameters, we treat them as schedulable resources whose capacity is gradually consumed during incremental learning. The cumulative activity cost before learning domain t is defined as $\mathbf{c}_t = \sum_{j=1}^{t-1}\mathbf{B}_{j,:} \in \mathbb{R}^M$, and the corresponding remaining activity budget is $\mathbf{r}_t = \exp\left(-\lambda \mathbf{c}_t\right) \in \mathbb{R}^M$, where $\lambda > 0$ is the activity-decay coefficient. The effective current-domain promotion signal is then

$$\mathbf{a}_t^+ = \mathbf{r}_t \odot \mathbf{a}_t^{(x)}. \tag{11}$$

where $\odot$ denotes element-wise multiplication. This formulation naturally suppresses overly aggressive adaptation on queries that have already been heavily occupied by previous domains. At the objective level, we additionally penalize changes on historically important queries by

$$\mathcal{L}_{\text{DQAR}} = \frac{1}{M}\sum_{k=1}^{M} a_{t,k}^{\text{his}} \left\|\mathbf{q}_k^{(t)} - \mathbf{q}_k^{(t-1)}\right\|_2^2, \tag{12}$$

where $a_{t,k}^{\text{his}}$ is the k-th element of the historical activity vector $\mathbf{a}_t^{\text{his}}$, and $\mathbf{q}_k^{(t-1)}$ and $\mathbf{q}_k^{(t)}$ denote the k-th query vectors before and after the current incremental update. This term explicitly protects query slots that are strongly associated with previous domains.

Activity-Aware Query Gating. The final gate combines current-domain promotion and historical-domain inhibition:

$$\mathbf{g}_t^{\text{raw}} = \alpha \mathbf{a}_t^+ - \beta \mathbf{a}_t^{\text{his}}, \tag{13}$$

where α and β control the strengths of self-domain promotion and cross-domain inhibition. This form favors queries that are active for the current domain but weakly occupied by historical domains, while suppressing queries strongly associated with previous ones. We further smooth the raw gate around 1:

$$\mathbf{g}_t = 1 + \gamma \cdot \tanh\left(\tau \mathbf{g}_t^{\text{raw}}\right), \tag{14}$$

where γ and τ control the modulation amplitude and sensitivity. The resulting gate remains bounded around 1, so DQAR refines rather than replaces the default query optimization. Accordingly, for the k-th query, its gradient is modulated as $\widetilde{\nabla}_{\mathbf{q}_k}\mathcal{L}_{\text{obj}}^{(t)} = g_{t,k}\nabla_{\mathbf{q}_k}\mathcal{L}_{\text{obj}}^{(t)}$, and the gated gradient is used to update the query set.

Overall, DQAR turns the decoder into a domain-aware resource allocation mechanism. By jointly modeling current-domain demand, historical-domain interference, and cumulative activity cost, it enables selective query specialization for new domains while preserving previously acquired knowledge.

Table 1: Statistics of the datasets used in our experiments.

Dataset	**HRRSD**	**LEVIR**	**DIOR**	**DOTA**	**MSAR**	**SSDD_MSAR**
Modality	RGB	RGB	RGB	RGB	SAR	SAR
Train size	3650	3032	4430	4216	21756	1024
Test size	1827	518	2216	2824	10880	486
Classes	ship, airplane, storage-tank					

Table 2: Comparison with State-of-the-Art Methods on Six Domain Datasets.

Dataset	HRRSD		LEVIR		DIOR		DOTA		MSAR		SSDD-MSAR		JOINT TEST	
Methods	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**	**mAP**	**AP$_{50}$**
Fine Tuning	14.6	25.3	15.9	26.3	13.9	34.1	8.0	20.9	13.2	34.3	33.2	40.5	13.4	31.0
Joint Training	75.9	95.0	61.3	88.3	56.3	87.2	42.8	79.7	45.9	82.1	29.3	48.5	50.4	83.3
CL-DETR [23]	46.3	68.2	30.2	52.1	23.8	53.6	18.1	39.4	20.5	40.8	16.0	22.5	23.8	45.2
Incre-DETR [8]	50.1	71.5	31.4	54.3	24.9	55.2	18.9	41.0	21.8	42.6	15.6	23.6	25.2	47.1
LDB [32]	51.2	73.0	30.9	55.0	26.8	56.8	19.3	41.8	22.4	43.3	13.5	24.1	25.8	48.0
RGR* [41]	55.9	81.6	35.8	61.2	28.7	61.5	21.4	46.0	24.8	47.2	15.4	26.5	28.5	52.7
LEA [33]	54.8	79.8	34.1	60.1	29.1	62.0	20.3	45.1	25.2	46.8	14.7	26.0	28.4	52.1
CoMBO* [9]	53.6	78.2	33.9	58.9	27.5	60.4	20.8	44.3	23.9	45.5	15.3	25.4	27.4	50.8
HyCal* [16]	55.6	80.7	34.9	60.9	28.9	61.9	20.6	45.6	25.1	47.0	14.5	26.3	28.5	52.4
Active-DDC [34]	60.0	87.5	34.8	61.8	30.1	63.2	**23.3**	**50.5**	25.5	47.8	16.2	26.8	29.5	53.9
DyQ-DETR [40]	57.8	83.6	35.5	62.5	26.6	63.0	21.7	47.2	25.8	48.1	15.7	27.1	29.1	53.8
Ours	**61.4**	**89.7**	**38.0**	**66.5**	**31.9**	**67.5**	22.9	49.6	**27.6**	**50.9**	**16.5**	**28.2**	**31.4**	**57.1**

* The marked continual-learning strategies are adapted to the unified DINO-R50 detector with consistent architectural enhancements for fair comparison.

5 Experiments

5.1 Experimental Setup

As shown in Table 1, we evaluate on a six-stage domain-incremental benchmark built from HRRSD [44], LEVIR [48], DIOR [19], DOTA [37], MSAR [2], and SSDD_MSAR [42]. The first four domains are RGB and the last two are SAR. All datasets share three categories, so the setting isolates domain shift rather than label expansion. Training follows the six-domain sequence without access to previous domains. We mainly compare with recent DETR-based continual detectors and also report CNN-based detectors as supplementary references.

5.2 Comparison with State-of-the-Art Methods

Table 2 compares the proposed method with recent state-of-the-art approaches on the six-stage domain-incremental remote sensing detection benchmark. Here, Joint Training denotes the joint-training upper bound on all six datasets, Fine Tuning denotes standard sequential domain-incremental training.

Overall, our method achieves the best performance among all compared approaches. Under joint test, it attains the highest overall results with mAP = **31.4%** and AP50 = **57.1%**, outperforming the strongest baseline, Active-DDC [34], by **1.9%** mAP. More importantly, it shows stronger resistance to forgetting on previously learned domains, especially under the challenging RGB-to-SAR transition. This advantage is further supported by the qualitative comparison in Fig. 3, where we compare with three strong baselines: the representative query-oriented DETR methods DyQ-DETR and Active-DDC, and DINO-adapted HyCal; our method yields fewer false positives and more stable cross-domain detections, indicating a better balance between continual adaptation and historical knowledge preservation.

5.3 Sequential Experiment

To further examine the robustness of our method to domain order, we conduct sequential experiments on four datasets from two modalities, including the RGB domains HRRSD, LEVIR, and DIOR, as

Table 3: Comparison results under different incremental sequences.

	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}	mAP	AP_{50}
SEQ I	**HRRSD**		**LEVIR**		**DIOR**		**MSAR**		**JOINT-TEST**	
Joint Training	76.3	94.8	62.7	88.0	58.8	89.3	37.6	74.5	47.4	80.2
Fine Tuning	18.4	30.0	16.3	29.4	18.0	40.1	26.9	46.8	23.7	42.3
DyQ-DETR [40]	46.8	75.2	32.6	58.9	29.7	57.8	27.9	48.2	31.9	53.5
Active-DDC [34]	50.3	78.4	34.1	61.2	31.0	60.3	28.4	50.6	32.2	55.0
HyCal [16]	51.0	77.6	34.6	60.5	31.5	59.8	27.6	50.1	31.5	54.8
Ours	**56.1**	**85.7**	**37.9**	**66.2**	**34.8**	**65.4**	**30.2**	**54.3**	**34.7**	**60.7**
SEQ II	**LEVIR**		**HRRSD**		**MSAR**		**DIOR**		**JOINT-TEST**	
Fine Tuning	13.0	27.1	19.8	32.6	22.5	40.0	37.9	60.3	23.4	40.6
DyQ-DETR [40]	30.8	55.4	55.1	82.0	26.1	49.2	42.3	67.0	32.5	56.7
Active-DDC [34]	32.1	57.3	57.6	84.7	27.8	51.6	44.0	69.2	34.2	58.6
HyCal [16]	31.6	56.1	56.8	83.5	26.5	48.9	43.1	67.8	33.7	56.5
Ours	**33.5**	**59.2**	**60.3**	**88.6**	**29.5**	**54.0**	**46.4**	**73.1**	**35.9**	**61.2**
SEQ III	**DIOR**		**MSAR**		**HRRSD**		**LEVIR**		**JOINT-TEST**	
Fine Tuning	19.6	32.0	16.0	28.1	30.2	46.8	42.0	63.4	21.5	35.4
DyQ-DETR [40]	24.8	36.5	23.9	36.9	37.5	52.8	47.3	68.2	28.9	39.4
Active-DDC [34]	26.1	38.7	25.8	39.8	39.6	56.1	49.5	71.0	30.6	42.0
HyCal [16]	25.6	37.9	25.1	38.6	38.9	54.8	48.7	69.9	30.1	41.1
Ours	**27.9**	**41.2**	**28.6**	**43.5**	**42.5**	**60.3**	**52.8**	**75.4**	**32.7**	**46.8**
SEQ IV	**MSAR**		**DIOR**		**LEVIR**		**HRRSD**		**JOINT-TEST**	
Fine Tuning	15.6	27.3	22.0	34.1	19.4	31.5	47.9	60.8	20.7	32.7
DyQ-DETR [40]	23.8	33.9	26.1	37.6	22.8	34.5	50.2	63.9	29.6	39.2
Active-DDC [34]	25.1	35.8	27.3	39.4	24.0	36.2	52.0	66.1	30.7	41.0
HyCal [16]	23.3	34.8	26.6	38.5	23.2	35.4	51.1	65.0	29.2	40.1
Ours	**26.9**	**38.7**	**28.9**	**42.1**	**25.6**	**38.7**	**54.3**	**69.8**	**31.8**	**44.4**

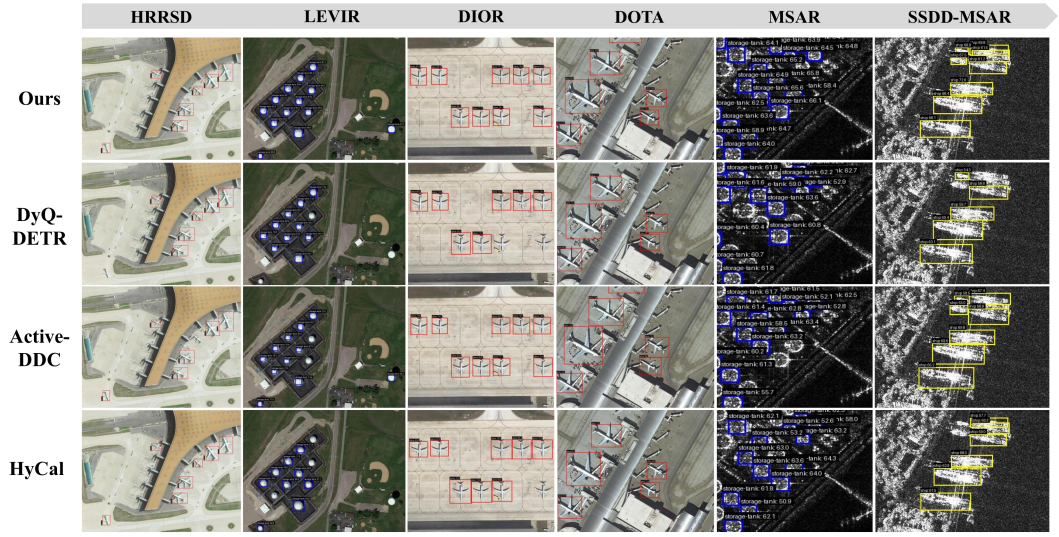


Figure 3: Qualitative comparison on six domain-incremental benchmarks. Our method yields fewer false positives and more stable cross-domain detection, especially across-modality shifts.

well as the SAR domain MSAR. We evaluate four training orders to assess sensitivity to domain sequence. We compare our approach with three strong baselines, namely DyQ-DETR, Active-DDC, and HyCal, which represent the most competitive recent incremental detection methods. As shown in Table 3, our method consistently achieves the best performance across all four orders. This indicates that the proposed design is not only effective in a fixed sequence, but also robust to different domain permutations. In particular, it maintains a better balance between adapting to newly arrived domains and preserving knowledge from earlier ones. These results further verify the stability and generality of Q-cost under sequential remote sensing domain shifts.

5.4 Ablation Study

Table 4 and Fig. 4 ablate PFHA, DCPE, and DQAR under cross-modality and intra-modality settings, respectively. In the RGB-only incremental sequence, DQAR raises joint mAP from **29.6%** to **36.4%**, showing that activity-cost estimation and query-update gating are effective even under fine-grained shifts within one modality. Combining PFHA and DCPE with DQAR further improves joint mAP to **46.0%**, confirming the benefit of the complete framework.

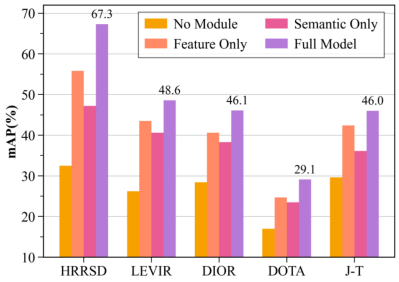


Figure 4: Intra-modality ablation of PFHA and DCPE (feature side) and DQAR (semantic side).

The cross-modality setting shows that PFHA's dual-level MoE feature adaptation and DCPE's prototype-derived encoder prompting complement DQAR. PFHA and DCPE each improve robustness to the RGB-to-SAR transition and reduce forgetting relative to the plain baseline. Combining either module with DQAR consistently outperforms DQAR alone, showing that feature-side conditioning supports query regulation under severe modality discrepancy. Such feature-side conditioning helps DQAR preserve historical query semantics despite modality-dependent appearance shifts. The full PFHA–DCPE–DQAR pipeline achieves the best old RGB, new RGB, new SAR, and overall mAP, while reducing forgetting to **3.1%**.

Table 4: Ablation under coarse-grained cross-modality incremental setting.

Modules			Modal(mAP)				
PFHA	DCPE	DQAR	Old RGB	New RGB	New SAR	Overall	Forgetting
-	-	-	37.9	19.0	23.7	21.7	18.9
✓	-	-	38.6	32.7	25.7	28.7	5.9
-	✓	-	37.1	30.3	24.9	27.2	6.8
-	-	✓	43.2	29.4	29.5	29.5	13.8
✓	✓	-	39.8	33.9	26.7	29.8	5.9
✓	-	✓	45.8	41.4	30.2	35.0	4.4
-	✓	✓	44.0	39.5	30.7	34.5	4.5
✓	✓	✓	**47.9**	**44.8**	**33.2**	**38.2**	**3.1**

5.5 Hyperparameter Analysis

Fig. 5 shows the effects of the sensitivity factor τ, modulation amplitude γ, and branch balance α/β. Intermediate values perform best, consistent with DQAR's soft query regulation. The factor τ controls the gate's sensitivity to activity differences, while γ sets the modulation strength. Small values yield insufficient specialization; large values amplify activity fluctuations and impair retention. The ratio α/β balances current-domain promotion against historical-domain inhibition: low ratios hinder adaptation, whereas high ratios favor new domains at the expense of retention. Effective regulation thus requires moderate sensitivity and modulation strength, with balanced plasticity and stability.

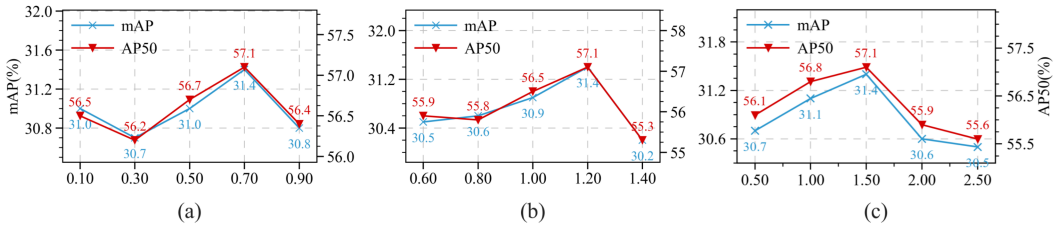


Figure 5: Hyper-parameter analysis across six domains, evaluated on the joint test set. (a) τ: effect of the sensitivity factor; (b) γ: effect of the modulation amplitude; (c) α/β: effect of the branch balance.

6 Conclusion

To the best of our knowledge, this is the first work to investigate cross-modality domain-incremental object detection for remote sensing, including both intra-modality evolution and transitions across heterogeneous sensing mechanisms such as RGB and SAR. From a query-as-resource perspective, Q-cost coordinates three connected modules: PFHA uses modality- and domain-aware prototypes for dual-level MoE feature adaptation; DCPE transforms the same prototypes into encoder prompts for domain-conditioned semantic aggregation; and DQAR estimates query activity and cumulative activity cost to gate decoder-query updates. Together, these modules encourage new-domain adaptation while preserving historically important query responses. Experiments demonstrate improved new-domain plasticity and old-domain retention under severe domain and modality shifts, while object-grounded QALR analysis verifies query preservation within DINO.

7 Acknowledgments and Disclosures

This work was supported in part by the National Natural Science Foundation of China under Grants 62372348 and 62576262, in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-ZDCYL-02-10, in part by the Natural Science Basic Research Program of Shaanxi under Grants 2026JC-YXQN-202 and 2026SYS-SYSZD-105, in part by the Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Key Research and Development Program of Shaanxi under Grant 2026JC-ZDGG-42, in part by the Establishment Fund of the State Key Laboratory of Wakefulness/Sleep and Cognition (Chongqing Institute for Brain and Intelligence), in part by the Fundamental Research Funds for the Central Universities under Grants QTZX25083 and QTZX26138.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229. Springer, 2020.

[2] Jie Chen, Zhixiang Huang, and Runfan Xia. Large-scale multi-class sar image target detection dataset-1.0, 2022. Journal of Radars Open Data Platform.

[3] De Cheng, Yubo Li, Chaowei Fang, Shizhou Zhang, Nannan Wang, and Xinbo Gao. Isolating interference factors for robust cloth-changing person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6089–6106, 2026.

[4] Gong Cheng and Junwei Han. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:11–28, 2016.

[5] Song Cui, Ailong Ma, Liangpei Zhang, Miaozhong Xu, and Yanfei Zhong. MAP-Net: SAR and optical image matching via image-based convolutional network with attention mechanism and spatial pyramid aggregated pooling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[6] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2849–2858, 2019.

[7] Li Ding, Xiang Song, Yuhang He, Changxin Wang, Songlin Dong, Xing Wei, and Yihong Gong. Domain incremental object detection based on feature space topology preserving strategy. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):424–437, 2024.

[8] 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, pages 543–551, 2023.

[9] Kai Fang, Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Combo: Conflict mitigation via branched optimization for class incremental segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25667–25676, 2025.

[10] Yulu Gao, Yifan Sun, Xudong Ding, Chuyang Zhao, and Si Liu. EASE-DETR: Easing the competition among object queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17282–17291, 2024.

[11] Shengxi Gui, Shuang Song, Rongjun Qin, and Yang Tang. Remote sensing object detection in the deep learning era—a review. *Remote Sensing*, 16(2):327, 2024.

[12] Yufei Guo, Yinan Wu, Haoran Duan, Guiguang Ding, and Jungong Han. PhaseLoRA: Control-regime-conditioned low-rank adaptation for continuous-action vision-language-action policies. *arXiv preprint arXiv:2608.15285*, 2026.

[13] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2786–2795, 2021.

[14] Thorsten Hoeser, Felix Bachofer, and Claudia Kuenzer. Object detection and image segmentation with deep learning on earth observation data: A review-part ii: Applications. *Remote Sensing*, 12(18):3053, 2020.

[15] Thorsten Hoeser and Claudia Kuenzer. Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. *Remote Sensing*, 12(10):1667, 2020.

[16] Eunju Lee, MiHyeon Kim, JuneHyoung Kwon, Yoonji Lee, JiHyun Kim, Soojin Jang, and YoungBin Kim. Hycal: A training-free prototype calibration method for cross-discipline few-shot class-incremental learning, 2026.

[17] Jiaxu Leng, Yongming Ye, Mengjingcheng Mo, Chenqiang Gao, Ji Gan, Bin Xiao, and M. A. Meyers. Recent advances for aerial object detection: A survey. *ACM Computing Surveys*, 56(12):1–36, 2024.

[18] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13619–13627, 2022.

[19] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

[20] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5291–5300, 2022.

[21] Zheng Li, Yongcheng Wang, Ning Zhang, Yuxi Zhang, Zhikang Zhao, Dongdong Xu, Guangli Ben, and Yunxiao Gao. Deep learning-based object detection techniques for remote sensing images: A survey. *Remote Sensing*, 14(10):2385, 2022.

[22] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. In *International Conference on Learning Representations*, 2022.

[23] 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 (CVPR)*, pages 23799–23808, 2023.

[24] Cheng Luo, Yueting Zhang, Jiayi Guo, Yuxin Hu, Guangyao Zhou, Hongjian You, and Xia Ning. SAR-CDSS: A semi-supervised cross-domain object detection from optical to SAR domain. *Remote Sensing*, 16(6):940, 2024.

[25] You Ma, Lin Chai, Lizuo Jin, and Jun Yan. Hierarchical alignment network for domain adaptive object detection in aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:39–52, 2024.

[26] M. Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3000–3010, 2022.

[27] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1(1):6–43, 2013.

[28] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2239–2247, 2019.

[29] Can Peng, Kun Zhao, and Brian C. Lovell. Faster ILOD: Incremental learning for object detectors based on faster RCNN. *Pattern Recognition Letters*, 140:109–115, 2020.

[30] 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 (ICCV)*, pages 3400–3409, 2017.

[31] Oscar Sommervold, Michele Gazzea, and Reza Arghandeh. A survey on sar and optical satellite image registration. *Remote Sensing*, 15(3):850, 2023.

[32] Xiang Song, Yuhang He, Songlin Dong, and Yihong Gong. Non-exemplar domain incremental object detection via learning domain bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15056–15065, 2024.

[33] Xiang Song, Yuhang He, Jingyuan Li, Qiang Wang, and Yihong Gong. Learning endogenous attention for incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30354–30364, 2025.

[34] Jiachen Sun, De Cheng, Xi Yang, and Nannan Wang. Dual domain control via active learning for remote sensing domain incremental object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3809–3818, 2025.

[35] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14454–14463, 2021.

[36] Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[37] Gui-Song Xia, Xiang Bai, Jian Ding, Zihao Zhu, Serge Belongie, Jiebo Luo, M. D. Datcu, Marcello Pelillo, and Jun Li. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3974–3983, 2018.

[38] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 949–958, 2022.

[39] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[41] Shizhou Zhang, Xueqiang Lv, Yinghui Xing, Qirui Wu, Di Xu, and Yanning Zhang. Revisiting generative replay for class incremental object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20340–20349, June 2025.

[42] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, and Hao Su. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18):3690, 2021.

[43] Xiangrong Zhang, Tianyang Zhang, Guanchun Wang, Peng Zhu, Xu Tang, Xiuping Jia, and Licheng Jiao. Remote sensing object detection meets deep learning: A metareview of challenges and advances. *IEEE Geoscience and Remote Sensing Magazine*, 11(4):8–44, 2023.

[44] Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5535–5548, 2019.

[45] Chenjie Zhao, Ryan Wen Liu, Jingxiang Qu, and Ruobin Gao. Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons. *Engineering Applications of Artificial Intelligence*, 128:107513, 2024.

[46] Xiaoxiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

[48] Zhengxia Zou and Zhenwei Shi. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Transactions on Image Processing*, 27(3):1100–1111, 2018.

Appendix

A Algorithm Details

Algorithm 1: Training and inference procedures of Q-cost.

Input: current-task dataset $\mathcal{D}^{(t)}$, previous detector $f(\cdot;\theta^{(t-1)})$, query set $\mathbf{Q}$, prototype buffer $\{\mathbf{p}_d\}$, domain-query activity buffer $\mathbf{B}$, queues $\{\mathcal{U}_j\}_{j=1}^{t}$, queue length S, and hyperparameters in Eqs. (2), (3), (9), and (11)–(14)

Output: updated detector $f(\cdot;\theta^{(t)})$, updated prototype buffer $\{\mathbf{p}_d\}$, updated activity buffer $\mathbf{B}$, updated queue $\mathcal{U}_t$, and test prediction $\hat{y}$

`/* Part I: Overall training algorithm of Q-cost at incremental task` t. `*/`

foreach *mini-batch sampled from* $\mathcal{D}^{(t)}$ **do**

  `/* Feature-side conditioning: update prototypes, perform modality/domain routing, and build the prompt-guided encoder input. */`

  Extract backbone features $\mathbf{F}^l$ and pooled features $\mathbf{f}^l = \mathrm{Pool}(\mathbf{F}^l)$ for each inserted stage l

  Update the current domain prototype $\mathbf{p}_d$ by EMA from the mini-batch pooled features, and derive modality centers $\mathbf{c}_m$ from the prototype buffer

  Route the sample to $m^\star = \arg\max_m \mathrm{sim}(\mathbf{f}^l, \mathbf{c}_m)$; apply the modality-specific adapter and the domain-fusion adapter to obtain fusion weights $\{\alpha_i\}_{i=1}^{K}$, fused prototype $\hat{\mathbf{p}}_t = \sum_{i=1}^{K} \alpha_i \mathbf{p}_{d_i}$, and residual $\mathbf{R}^l$; set $\widetilde{\mathbf{F}}^l = \mathbf{F}^l + \mathbf{R}^l$

  Compute $\mathcal{L}_{\mathrm{PFHA}}^{(t)}$ by Eq. (3)

  Construct the encoder prompt $\mathbf{e}_p = \mathbf{W}_m \mathbf{c}_{m^\star} + \mathbf{W}_d \hat{\mathbf{p}}_t$ by Eq. (4)

  Prepend $\mathbf{e}_p$ to the encoder tokens $\mathbf{Z}$ to obtain $\widetilde{\mathbf{Z}}$, and compute $\mathcal{L}_{\mathrm{DCPE}}^{(t)}$ by Eq. (5)

  Decode $\widetilde{\mathbf{Z}}$ to obtain $\mathbf{H}(x) = \Psi(\mathbf{Q}, \widetilde{\mathbf{Z}}(x))$ by Eq. (6), and compute $\mathcal{L}_{\mathrm{DINO}}^{(t)}$ by Eq. (1)

  `/* Query-side regulation: estimate current/historical activities, compute activity-aware gates, and update queries selectively. */`

  Compute $\bar{\mathbf{H}}_t$ by Eq. (7), $\mathbf{\Delta}_t(x) = \mathbf{H}(x) - \bar{\mathbf{H}}_t$, and $\mathbf{a}_t^{(x)}$ by Eq. (8); update $\mathbf{B}_{t,:}$ by EMA

  if $t > 1$ **then**

    Compute the historical relevance weights $\mathbf{w}_t$ by Eq. (9) and the historical activity $\mathbf{a}_t^{\mathrm{his}}$ by Eq. (10)

    Compute $\mathbf{c}_t = \sum_{j=1}^{t-1} \mathbf{B}_{j,:}$, $\mathbf{r}_t = \exp(-\lambda \mathbf{c}_t)$, and $\mathbf{a}_t^{+}$ by Eq. (11)

    Compute $\mathcal{L}_{\mathrm{DQAR}}^{(t)}$ by Eq. (12)

  else

    Set $\mathbf{a}_t^{\mathrm{his}} = \mathbf{0}$, $\mathbf{r}_t = \mathbf{1}$, $\mathbf{a}_t^{+} = \mathbf{a}_t^{(x)}$, and $\mathcal{L}_{\mathrm{DQAR}}^{(t)} = 0$

  Compute $\mathbf{g}_t^{\mathrm{raw}}$ and $\mathbf{g}_t$ by Eqs. (13) and (14)

  Form $\mathcal{L}_{\mathrm{obj}}^{(t)}$ by Eq. (2), backpropagate it, and modulate each query gradient as $\widetilde{\nabla}_{\mathbf{q}_k} \mathcal{L}_{\mathrm{obj}}^{(t)} = g_{t,k} \nabla_{\mathbf{q}_k} \mathcal{L}_{\mathrm{obj}}^{(t)}$

  Update $\theta^{(t-1)} \to \theta^{(t)}$ using the gated gradients for $\mathbf{Q}$ and standard gradients for the remaining parameters

  Push $\mathbf{H}(x)$ into $\mathcal{U}_t$ and discard the oldest entry if $|\mathcal{U}_t| > S - 1$

`/* Part II: Inference algorithm of Q-cost after learning task` t. `*/`

`/* Test-time conditioning: infer domain context from prototypes and construct the same prompt-enhanced encoder representation. */`

Extract backbone features $\mathbf{F}^l$ and pooled feature $\mathbf{f}^l = \mathrm{Pool}(\mathbf{F}^l)$ for each inserted stage l

Derive modality centers $\mathbf{c}_m$ from the prototype buffer, route x to $m^\star = \arg\max_m \mathrm{sim}(\mathbf{f}^l, \mathbf{c}_m)$, and apply the modality-specific adapter and the domain-fusion adapter to obtain fusion weights $\{\alpha_i\}_{i=1}^{K}$, fused prototype $\hat{\mathbf{p}} = \sum_{i=1}^{K} \alpha_i \mathbf{p}_{d_i}$, and PFHA residual $\mathbf{R}^l$

Set $\widetilde{\mathbf{F}}^l = \mathbf{F}^l + \mathbf{R}^l$, and construct the encoder prompt $\mathbf{e}_p = \mathbf{W}_m \mathbf{c}_{m^\star} + \mathbf{W}_d \hat{\mathbf{p}}$ by Eq. (4)

Prepend $\mathbf{e}_p$ to the encoder token sequence to obtain $\widetilde{\mathbf{Z}}$

Feed $\widetilde{\mathbf{Z}}$ and $\mathbf{Q}$ into the decoder to obtain $\mathbf{H}(x) = \Psi(\mathbf{Q}, \widetilde{\mathbf{Z}}(x))$ as in Eq. (6)

Predict classification scores and bounding boxes from $\mathbf{H}(x)$, and apply the standard DINO post-processing to obtain $\hat{y}$

return $f(\cdot;\theta^{(t)})$, $\{\mathbf{p}_d\}$, $\mathbf{B}$, $\mathcal{U}_t$, *and* $\hat{y}$

B Justification of Q-cost

Consider the incremental update from task $t-1$ to task t for a DETR-like detector with decoder responses $\mathbf{H}(x) = \Psi(\mathbf{Q}, \widetilde{\mathbf{Z}}(x))$ in Eq. (6). Let $\mathbf{h}_k(x) \in \mathbb{R}^C$ denote the k-th row of $\mathbf{H}(x)$, and let $\delta\mathbf{q}_k = \mathbf{q}_k^{(t)} - \mathbf{q}_k^{(t-1)}$. Under a local second-order approximation of the slot loss and a query-wise isotropic curvature approximation, the locally optimal query update admits the form

$$\delta\mathbf{q}_k^\star = -\eta\, g_{t,k}^\star\, \nabla_{\mathbf{q}_k} \mathcal{L}_{\mathrm{obj}}^{(t)}, \qquad g_{t,k}^\star = 1 + \nu\big(\alpha a_{t,k}^+ - \beta a_{t,k}^{\mathrm{his}}\big), \tag{15}$$

where $a_{t,k}^{(x)}$ is proportional to the response deviation norm $\|\mathbf{\Delta}_{t,k}(x)\|_2$, $a_{t,k}^+ = r_{t,k} a_{t,k}^{(x)}$, and $a_{t,k}^{\mathrm{his}}$ is the relevance-weighted historical activity. Consequently, the practical gate defined by Eqs. (13) and (14) is a bounded smooth realization of the locally optimal query-wise step-size modulation.

C Derivation Details

For DETR-like decoders, each query is an object-centric semantic slot. Writing one cross-attention update explicitly,

$$\mathbf{h}_k(x) = \sum_{n=1}^{N} \alpha_{k,n}(x)\, \mathbf{V}\mathbf{z}_n(x), \qquad \alpha_{k,n}(x) = \frac{\exp\Big((\mathbf{W}_q\mathbf{q}_k)^\top(\mathbf{W}_k\mathbf{z}_n(x))/\sqrt{C}\Big)}{\sum_{r=1}^{N}\exp\Big((\mathbf{W}_q\mathbf{q}_k)^\top(\mathbf{W}_k\mathbf{z}_r(x))/\sqrt{C}\Big)}. \tag{16}$$

Hence $\mathbf{q}_k$ controls how the k-th slot aggregates encoder evidence. Any domain shift therefore acts on the decoder primarily through the change of slot responses $\mathbf{h}_k(x)$.

After Hungarian matching, the DINO loss is slot-wise decomposable:

$$\mathcal{L}_{\mathrm{DINO}}(x) = \sum_{k=1}^{M} \ell_k(\mathbf{h}_k(x); \sigma^*(x), y) + \mathcal{L}_{\mathrm{dn}}(x). \tag{17}$$

Thus,

$$\nabla_{\mathbf{q}_k}\mathcal{L}_{\mathrm{DINO}}(x) = \mathbf{J}_k(x)^\top \mathbf{g}_k(x), \qquad \mathbf{J}_k(x) = \frac{\partial \mathbf{h}_k(x)}{\partial \mathbf{q}_k}, \quad \mathbf{g}_k(x) = \nabla_{\mathbf{h}_k}\ell_k. \tag{18}$$

So the optimization pressure on $\mathbf{q}_k$ is completely mediated by its slot response $\mathbf{h}_k(x)$.

Let $\bar{\mathbf{h}}_{t,k}$ be the k-th row of the queue mean $\bar{\mathbf{H}}_t$ in Eq. (7), and define

$$\mathbf{\Delta}_{t,k}(x) = \mathbf{h}_k(x) - \bar{\mathbf{h}}_{t,k}. \tag{19}$$

A second-order expansion of ℓ_k around $\bar{\mathbf{h}}_{t,k}$ gives

$$\ell_k(\mathbf{h}_k(x)) = \ell_k(\bar{\mathbf{h}}_{t,k}) + \mathbf{u}_{t,k}^\top \mathbf{\Delta}_{t,k}(x) + \frac{1}{2}\mathbf{\Delta}_{t,k}(x)^\top \mathbf{M}_{t,k}\mathbf{\Delta}_{t,k}(x) + o\big(\|\mathbf{\Delta}_{t,k}(x)\|_2^2\big)\,, \tag{20}$$

where $\mathbf{u}_{t,k} = \nabla\ell_k(\bar{\mathbf{h}}_{t,k})$ and $\mathbf{M}_{t,k} = \nabla^2\ell_k(\bar{\mathbf{h}}_{t,k})$. Since $\bar{\mathbf{h}}_{t,k}$ is the local response center estimated from recent samples of domain t, we have $\mathbb{E}_{x\sim\mathcal{D}^{(t)}}[\mathbf{\Delta}_{t,k}(x)] \approx \mathbf{0}$, and therefore

$$\mathbb{E}\big[\ell_k(\mathbf{h}_k(x)) - \ell_k(\bar{\mathbf{h}}_{t,k})\big] \approx \frac{1}{2}\mathrm{tr}(\mathbf{M}_{t,k}\mathbf{\Sigma}_{t,k})\,, \qquad \mathbf{\Sigma}_{t,k} = \mathbb{E}\big[\mathbf{\Delta}_{t,k}(x)\mathbf{\Delta}_{t,k}(x)^\top\big]\,. \tag{21}$$

Under the isotropic approximation $\mathbf{M}_{t,k} \approx \lambda_{t,k}\mathbf{I}$,

$$\mathbb{E}\big[\ell_k(\mathbf{h}_k(x)) - \ell_k(\bar{\mathbf{h}}_{t,k})\big] \approx \frac{\lambda_{t,k}}{2}\,\mathbb{E}\|\mathbf{\Delta}_{t,k}(x)\|_2^2. \tag{22}$$

Hence the activity definition in Eq. (8), namely $a_{t,k}^{(x)} \propto \|\mathbf{\Delta}_{t,k}(x)\|_2$, is a principled proxy for the current-domain contribution of query k.

Next consider forgetting on a previous domain $j < t$. A second-order expansion in parameter space yields

$$\Delta\mathcal{L}_j := \mathbb{E}_{x\sim\mathcal{D}^{(j)}}\Big[\mathcal{L}(x;\mathbf{Q}^{(t)}) - \mathcal{L}(x;\mathbf{Q}^{(t-1)})\Big] \approx \frac{1}{2}\sum_{k=1}^{M}\delta\mathbf{q}_k^\top \mathbf{F}_{j,k}\delta\mathbf{q}_k, \tag{23}$$

where $\mathbf{F}_{j,k}$ is the query-wise curvature induced by domain j. By Eq. (18), $\mathbf{F}_{j,k}$ is determined by the response Jacobian and slot-loss curvature, hence scales with the historical response variance of query k. Under the same isotropic approximation $\mathbf{F}_{j,k} \approx \kappa_{j,k}\mathbf{I}$ with $\kappa_{j,k} \propto a_{j,k}$, we obtain

$$\Delta\mathcal{L}_j \propto \sum_{k=1}^{M} a_{j,k} \|\delta\mathbf{q}_k\|_2^2. \tag{24}$$

Aggregating all previous domains with the relevance weights in Eq. (9) gives

$$\Delta\mathcal{L}_{\text{his}} \propto \sum_{k=1}^{M} a_{t,k}^{\text{his}} \|\delta\mathbf{q}_k\|_2^2, \qquad \mathbf{a}_t^{\text{his}} = \mathbf{w}_t^\top \mathbf{B}_{1:t-1}, \tag{25}$$

which is exactly the rationale behind Eq. (12).

For the current task, Q-cost further discounts the instantaneous activity by a remaining plasticity budget. If the residual plasticity of query k decays proportionally to its cumulative occupation $c_{t,k}$, then

$$\frac{dr_{t,k}}{dc_{t,k}} = -\lambda r_{t,k}, \qquad r_{t,k}(0) = 1, \tag{26}$$

whose solution is

$$r_{t,k} = \exp(-\lambda c_{t,k}), \qquad a_{t,k}^{+} = r_{t,k} a_{t,k}^{(x)}, \tag{27}$$

which matches Eq. (11).

Now write the query update as

$$\delta\mathbf{q}_k = -\eta s_k \nabla_{\mathbf{q}_k} \mathcal{L}_{\text{obj}}^{(t)}, \tag{28}$$

where s_k is a query-wise step-size factor. Consider the local utility

$$\mathcal{U}_t(\{s_k\}) = \sum_{k=1}^{M} \left[(\alpha a_{t,k}^{+} - \beta a_{t,k}^{\text{his}}) s_k - \frac{(s_k - 1)^2}{2\nu} \right] \left\| \nabla_{\mathbf{q}_k} \mathcal{L}_{\text{obj}}^{(t)} \right\|_2^2, \tag{29}$$

where the first term rewards plastic adaptation to the current domain and penalizes interference with historical domains, while the second term keeps the update close to the vanilla gradient step. Maximizing Eq. (29) with respect to each s_k gives

$$s_k^\star = 1 + \nu(\alpha a_{t,k}^{+} - \beta a_{t,k}^{\text{his}}). \tag{30}$$

Substituting Eq. (30) into Eq. (28) yields Eq. (15). Replacing the affine factor by the bounded smooth map

$$g_{t,k} = 1 + \gamma \tanh\big(\tau(\alpha a_{t,k}^{+} - \beta a_{t,k}^{\text{his}})\big) \tag{31}$$

gives the practical gate in Eqs. (13) and (14).

Therefore, we have shown three facts: (i) response deviation norm is a valid estimator of current-domain query importance; (ii) historical query activity determines the local forgetting curvature; (iii) the Q-cost gate is a smooth approximation to the locally optimal query-wise update.

Hence, the proposed analysis elevates query activity from an empirical heuristic to a principled optimization quantity, showing that Q-cost realizes continual adaptation as activity-aware, query-wise plasticity control in DETR-like/DINO detectors.

D Experimental Details

As shown in Table 5, the base detector follows the official DINO configuration with batch size 2, while the remaining general settings are kept at their default values. The task-specific parameters of Q-cost are divided into prototype/prompt-related and query-regulation-related components. Among them, τ, γ, and α/β are selected according to the hyperparameter study, whereas the auxiliary parameters are fixed to stable default values. Unless otherwise stated, all experiments use the same configuration across datasets and training orders. All experiments are conducted on two NVIDIA RTX 4090 GPUs. Following the default DINO training protocol, the model is trained for 12 epochs on each incremental domain. Unless otherwise specified, the remaining implementation details, including data preprocessing, optimization strategy, and training schedule, follow the official DINO defaults, while the task-specific settings of Q-cost are listed in Table 5.

Table 5: Default hyper-parameter settings used in our experiments.

Type	Hyper-parameter	Default value
DINO defaults	Batch size	2
	Optimizer	AdamW
	Learning rate	1×10^{-4}
	Backbone learning rate	1×10^{-5}
	Weight decay	1×10^{-4}
	Epochs	12
	Gradient clip norm	0.1
	Backbone	ResNet-50
	Encoder/decoder layers	6 / 6
	Hidden dimension	256
	Number of object queries	900
	DN settings	100 / 0.5 / 0.4
Loss coefficients	PFHA loss weight λ_p	0.3
	DCPE loss weight λ_e	0.1
	DQAR loss weight λ_q	0.7
Prototype & prompt	Prototype momentum μ	0.99
	Similarity temperature ξ	0.07
	Response queue length S	5
	Activity decay λ	0.50
DQAR	Branch balance ratio α/β	1.50
	Sensitivity factor τ	0.70
	Modulation amplitude γ	1.20

E Hyperparameter Analysis

Effect of the sensitivity factor τ: The parameter τ controls how sensitively the final gate responds to the estimated query activity. A smaller τ makes the gating curve flatter, so even noticeable activity differences only lead to mild update modulation. As shown in Fig. 6, when τ is relatively small, the model tends to preserve earlier domains better, since the influence of activity signals on query updates remains conservative. However, such conservative regulation also weakens adaptation to newly arrived domains, leading to inferior generalization on the latest task. As τ increases, the gate becomes more responsive to activity variation, which improves domain-specific adaptation. Nevertheless, an overly large τ makes the decoder excessively sensitive to activity fluctuation, resulting in unstable query specialization and degraded overall performance. The best trade-off is achieved at τ=**0.70**, where the model balances retention and adaptation most effectively.

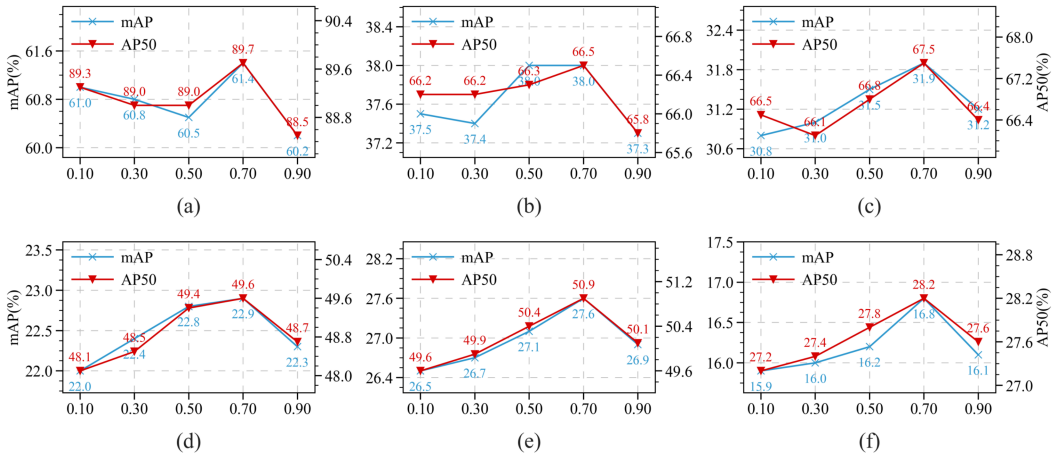


Figure 6: The Analysis of the Hyperparameter τ Includes the mAP and AP50 of HRRSD, LEVIR, DIOR, DOTA, MSAR and SSDD-MSAR.

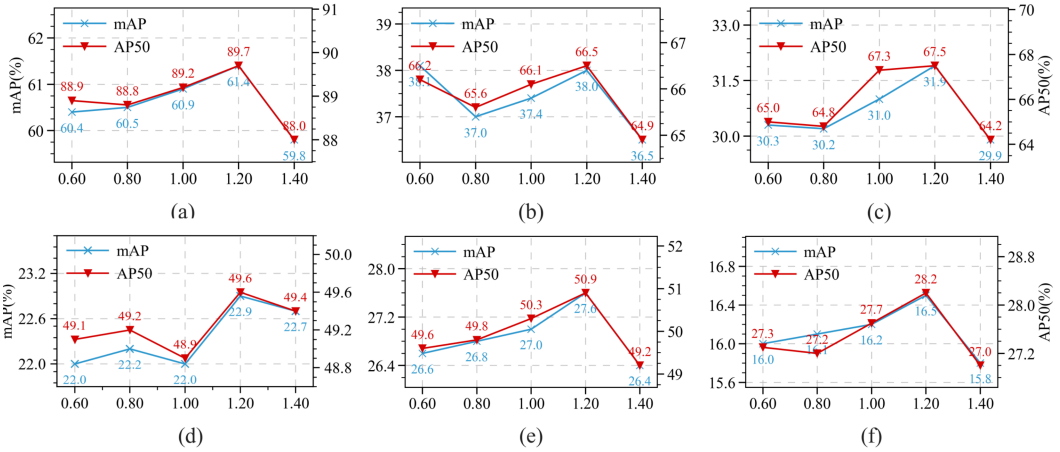


Figure 7: The Analysis of the Hyperparameter γ Includes the mAP and AP50 of HRRSD, LEVIR, DIOR, DOTA, MSAR and SSDD-MSAR.

Effect of the modulation amplitude γ: We then study the modulation amplitude γ, which determines how strongly the gate affects the actual query update magnitude. In essence, γ plays a role similar to an activity-aware step-size controller. As shown in Fig. 7, when γ is small, the final gate remains close to 1, and DQAR only slightly perturbs the natural query update. This mild modulation is beneficial to historical-domain preservation, which explains why the earliest domains remain relatively stable in the low-γ regime. However, the corresponding update strength is insufficient for effective specialization to the current domain, and the newest domain therefore shows weaker generalization. Increasing γ allows the gate to exert stronger control over domain-specialized query learning, which improves adaptation to recent domains. On the other hand, overly large γ amplifies the gating effect too aggressively, making the query update process unstable and harming the balance between plasticity and stability. The optimal performance is obtained at γ=**1.20**.

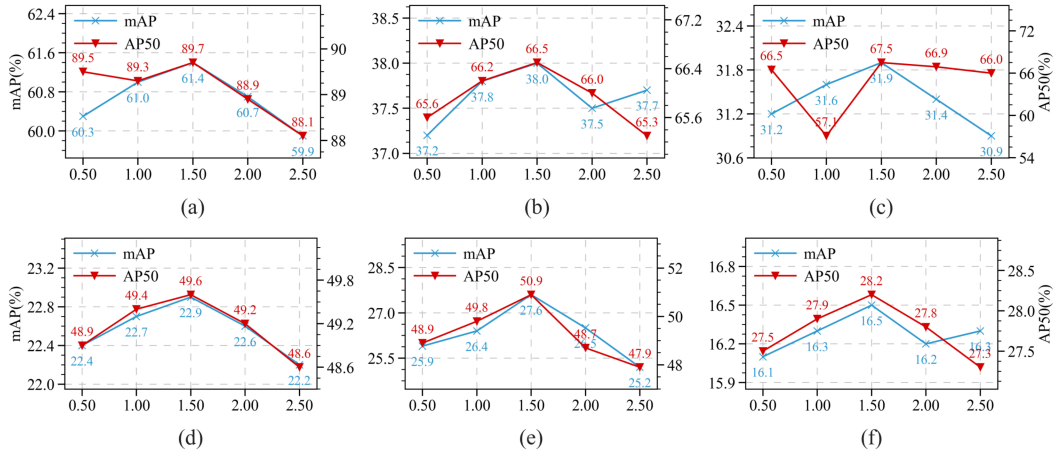


Figure 8: The Analysis of the Hyperparameter α/β Includes the mAP and AP50 of HRRSD, LEVIR, DIOR, DOTA, MSAR and SSDD-MSAR.

Effect of the branch balance α/β: The ratio α/β determines the balance between current-domain promotion and historical-domain inhibition in the query activity score. Small ratios strengthen historical suppression, improving retention but limiting adaptation to the current domain. Larger ratios encourage query specialization to newly arrived domains, with weaker preservation of earlier knowledge. As shown in Fig. 8, the best result occurs at α/β=**1.50**. This balanced setting supports DQAR's joint modeling of current and historical query activity to coordinate plasticity and stability.

F Supplementary Experiments and Analyses

F.1 Baseline Architectures

The six-domain comparison includes methods adapted to DINO-R50 and baselines evaluated with other detectors. These architectural assignments help distinguish the continual-learning strategy from the underlying detector when reading the main comparison. The marked strategies share the unified DINO-R50 configuration, while other baselines retain their reported detector families.

F.2 Computational Cost

The cumulative overhead of PFHA, DCPE, and DQAR is profiled on an NVIDIA RTX 4090, with training measured at the final incremental stage and inference measured at batch size 1 with 800×800 inputs over 500 synchronized runs after 100 warm-up runs. The full framework increases inference latency by 9.2% and training time per epoch by 38.9%, with 1.5% more parameters. These results indicate that its added cost is concentrated more in training than inference.

The measured latency rises from 96.4 ms for DINO-FT to 105.3 ms for the complete model, while peak training memory increases from 21.3 to 23.0 GB. The rows report cumulative configurations, so each percentage compares the listed configuration with DINO-FT rather than with the preceding row. This distinction allows parameter, computation, memory, and timing overheads to be read on a common baseline.

Table 6: Model cost profile. Peak memory and time per epoch refer to training; latency is end-to-end inference time. Parentheses give cumulative overheads relative to DINO-FT. Each row adds the indicated modules to DINO-FT.

Model	Params (M)	GFLOPs	Peak memory (GB)	Train time (min/epoch)	Latency (ms)
DINO-FT	46.6	201.8	21.3	18	96.4
+ PFHA	47.1 (+1.1%)	208.8 (+3.5%)	22.8 (+7.0%)	22 (+22.2%)	102.2 (+6.0%)
+ PFHA + DCPE	47.3 (+1.5%)	209.2 (+3.7%)	23.0 (+8.0%)	24 (+33.3%)	104.0 (+7.9%)
+ PFHA + DCPE + DQAR (full)	47.3 (+1.5%)	209.4 (+3.8%)	23.0 (+8.0%)	25 (+38.9%)	105.3 (+9.2%)

F.3 Memory Scalability

Memory usage consists of a fixed current-domain working queue and per-domain activity summaries and historical prototypes whose storage grows linearly with the number of domains. The working queue stays at 3.52 MiB regardless of sequence length. At 100 domains, the growing-state subtotal is still below 1 MiB, separating fixed working storage from compact historical state.

With query count and feature dimensions fixed, extending the domain sequence adds activity summaries and historical prototypes without enlarging the working queue. The growing-state subtotal increases from 45.09 KiB at six domains to 751.56 KiB at 100 domains. This subtotal excludes the fixed queue, so the components should be considered separately when interpreting the total auxiliary memory required by a longer sequence.

Table 7: Memory scalability in FP32. The growing-state subtotal includes activity summaries and historical prototypes, excluding the fixed current-domain queue.

Number of domains	Current-domain queue	Activity summaries	Historical prototypes	Growing-state subtotal
6	3.52 MiB	21.09 KiB	24.00 KiB	45.09 KiB
10	3.52 MiB	35.16 KiB	40.00 KiB	75.16 KiB
20	3.52 MiB	70.31 KiB	80.00 KiB	150.31 KiB
50	3.52 MiB	175.78 KiB	200.00 KiB	375.78 KiB
100	3.52 MiB	351.56 KiB	400.00 KiB	751.56 KiB

F.4 Results over Multiple Random Seeds

Three independent runs of the complete six-stage sequence use identical dataset splits, domain order, and training schedules to assess the consistency of Q-cost across random initializations. Q-cost achieves the highest mean mAP and AP_{50} across these runs, with standard deviations of 0.40 and 0.72 percentage points, respectively. The mean gains over Active-DDC are 2.04 percentage points in mAP and 4.60 points in AP_{50}, indicating that the advantage holds across both individual runs and their averages. This supports consistency across the reported initializations.

Table 8: Results over three independent runs on the six-domain sequence. The first three result columns report mAP/AP_{50} (%); the final two report the mean and sample standard deviation across runs.

Method	Seed 1	Seed 2	Seed 3	mAP (%) mean $\pm$ std.	AP_{50} (%) mean $\pm$ std.
DINO-FT	13.4/31.0	12.8/31.2	13.9/32.5	13.37 ± 0.55	31.57 ± 0.81
Active-DDC	29.5/53.9	29.2/52.0	29.6/52.7	29.43 ± 0.21	52.87 ± 0.96
Q-cost	31.4/57.1	31.9/58.3	31.1/57.0	$\mathbf{31.47 \pm 0.40}$	$\mathbf{57.47 \pm 0.72}$

G Limitations

Despite its strong empirical performance, Q-cost still has several limitations. First, its design assumes identifiable task boundaries and sensor/modality information when new training domains are incorporated; ambiguous or continuously drifting streams may weaken prototype estimation and query-activity tracking. An unidentified modality encountered only at inference is assigned to an existing modality center, with no guarantee of reliable performance. Second, prototype buffers, activity queues, prompt construction, and query gating increase training complexity and introduce additional hyperparameters, although the same defaults are used across the reported datasets and training orders. Computational costs and memory scaling are reported in Appendices F.2 and F.3. Third, historical activity estimation depends on prototype quality and may become less accurate for highly imbalanced, noisy, or weakly separable domains. Finally, evaluation covers RGB and SAR remote sensing streams with DINO-R50; generalization to longer sequences, other modalities, boundary-free online learning, and broader detector families remains to be validated.

H Broader Impacts and Responsible Use

Continual adaptation across sensing domains can support disaster assessment, environmental monitoring, infrastructure inspection, and maritime search and rescue. Preserving detection performance as acquisition conditions and sensors change may improve the reliability of long-term Earth-observation systems.

Remote sensing detection also has dual-use risks, including unauthorized surveillance, privacy-invasive monitoring, and military reconnaissance. False detections or missed objects under domain shifts can lead to harmful downstream decisions, even when the system is used for its intended purpose. The reported RGB/SAR benchmark results do not establish reliability for unrestricted or safety-critical deployment.

Responsible deployment should respect applicable privacy and data-governance requirements and the licenses of source datasets and models. We recommend human oversight, uncertainty assessment, validation on the intended deployment domain, and auditing of consequential decisions; access controls should be considered where the application warrants them. These are deployment recommendations rather than safeguards experimentally validated in this work. The code release should document the method's assumptions, intended uses, and limitations, and should direct users to the original dataset providers rather than redistribute data without the necessary permissions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the target setting, the proposed Q-cost framework, and its main contribution of balancing new-domain adaptation and old-domain retention under cross-modality domain shifts. These claims are consistent with the method and the experimental results reported in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix G discusses task-boundary and modality assumptions, prototype quality, computational costs, and the limited scope of the RGB/SAR evaluations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendices B and C state and derive the numbered analytical relationships under local second-order and query-wise isotropic-curvature approximations. This is a local justification of the activity-based update rule, rather than a global optimality guarantee.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix A describes the training and inference procedures, Appendix D lists the default settings and hardware, and Appendix F gives the computational profiling, memory-scaling, and multiple-seed evaluation protocols. The public implementation is linked in the abstract.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the code at `https://github.com/XI-Lab-XDU/Q-cost`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The benchmark and data splits are described in the experimental setup, default training settings are listed in Appendix D, and the supplementary comparison and measurement protocols are given in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix F.4 reports all three independent runs and the mean and sample standard deviation of mAP and AP_{50} for Q-cost, DINO-FT, and Active-DDC on the complete six-domain sequence. The variability is across random seeds with fixed data splits, domain order, and training schedules; no formal hypothesis test is claimed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computer resources in Appendices D, F.2, and F.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses existing public benchmarks without new human-subject data collection. Potential societal harms and responsible-use considerations are discussed in Appendix H.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix H discusses beneficial Earth-observation applications, surveillance and privacy risks, military misuse, and harmful downstream decisions caused by detection errors, together with recommended mitigation measures.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not introduce a new high-risk generative model or a newly collected dataset. This does not imply an absence of risk: dual-use concerns and recommended deployment safeguards are discussed in Appendix H.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The original datasets, detectors, and compared methods are credited through citations, but version-specific licenses and terms of use for all existing assets are not explicitly enumerated in the manuscript.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: The implementation is released at the repository linked in the abstract, but a complete accompanying asset-documentation package is not provided in the manuscript. The experiments use existing benchmark datasets rather than a newly collected human-subject dataset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM usage is strictly limited to linguistic polishing and formatting, with no impact on the core methodology or scientific originality.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.