Real-Time Source-Free Object Detection

Sairam VCR[1], Varun Gopal[1], Poornima Jain[1],
Vineeth N Balasubramanian[1,2], and Muhammad Haris Khan[3]

[1]IIT Hyderabad, India [2]Microsoft Research [3]MBZUAI
{ai20resch13001@,co22btech11015@,ai24resch11002@,
vineethnb@cse.}iith.ac.in, muhammad.haris@mbzuai.ac.ae

Abstract. Real-world detectors for autonomous driving, surveillance, and robotics must handle domain-shifts under strict latency and memory constraints, yet existing source-free object detection (SFOD) methods rely on heavyweight architectures that prioritize accuracy alone. We show this trade-off is unnecessary: building on YOLOv10, an NMS-free dual-head detector, we achieve state-of-the-art adaptation accuracy while being faster and more compact. We observe that directly applying vanilla mean-teacher self-training to dual-head detectors leads to sub-optimal adaptation performance due to two key factors. First, simple pseudo-label generation strategies, such as using a single head or directly combining high-confidence predictions from both heads, yield sub-optimal supervision under domain-shift. We propose DHF (Dual-Head Pseudo-Label Fusion) which selectively admits one-to-one (O2O) and one-to-many (O2M) head predictions, preserving precision and recovering missed objects. Second, we observe domain-shift collapses multi-scale feature discriminability. We propose the use of our MARD (Multi-scale Adaptive Representation Diversification) loss which mitigates this by enforcing detection-aware variance and covariance constraints on multi-scale feature maps. Both modules are training-time only, leaving inference unchanged. Across domain-shift benchmarks, our method, RT-SFOD yields 1.4 to 3.5% mAP gains, 1.3× higher throughput, with ∼2× fewer parameters than prior state-of-the-art SFOD methods, thus advancing the Pareto frontier of the speed-accuracy-model size trade-off. We report main results with YOLOv10, and demonstrate generalizability with additional YOLO- and DETR-based dual-head detectors. Code is available here: https://github.com/Sairam13001/RT-SFOD/.

Keywords: Source-Free Object Detection · Efficient Object Detection

1 Introduction

Source-Free Object Detection (SFOD) adapts a detector trained on labeled source data to an unlabeled target domain *without* using source samples [43, 12, 42, 51]. This setting is increasingly relevant in applications such as autonomous driving, surveillance, robotics, and medical imaging, where privacy concerns, proprietary restrictions, or storage limitations preclude access to source

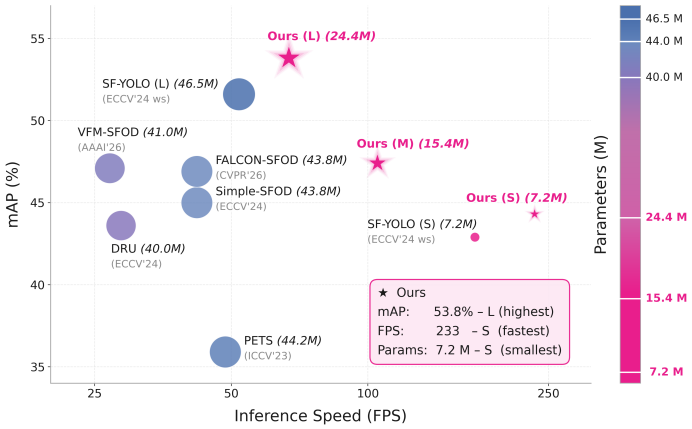
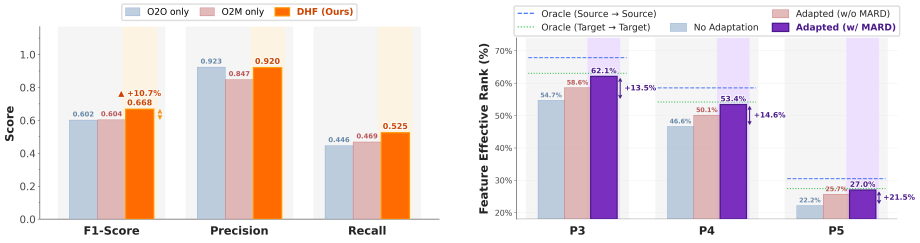


Fig. 1: **Accuracy-speed-size trade-off in SFOD.** We compare state-of-the-art SFOD methods on Cityscapes → Foggy Cityscapes domain-shift, in terms of mAP (%) and inference speed (FPS), with model size (# parameters) encoded by color-graded bubble/star size. Our method achieves the best trade-off across all model-scales, obtaining 53.8% mAP with the large model while remaining the fastest (233 FPS) and most compact (7.2 M params) with the small model.

data[36, 33, 29, 1]. Such real-time applications typically operate under strict latency and memory constraints[27, 35]. However, following the pioneering SFOD method [22], existing works[43, 24, 12, 52, 42] largely build on two-stage Faster R-CNN [31], with a few recent ones[19, 52, 2, 51] relying on transformer-based detectors [61, 50]. While these backbones are strong, they are ill-suited for real-time applications [46, 32]. One-stage detectors such as YOLO [39, 15, 44] are designed for efficiency, but limited attention has been given to developing SFOD methods based on them [41]. To bridge this gap, we build a new SFOD framework on YOLOv10[44], which beats the existing methods on accuracy while using a fraction of their parameter count, thus establishing a new state-of-the-art on the accuracy-speed-size frontier as shown in Fig. 1. While we adopt YOLOv10 as the primary architecture for all main experiments, our framework applies to any NMS-free dual-head detector, as we validate on other YOLO- and DETR-based detectors in Table 6.

YOLOv10 is the first non-maximum suppression (NMS)-free YOLO detector, employing a dual-head training strategy with one-to-many (O2M) and one-to-one (O2O) heads under a consistent matching scheme to enable true end-to-end, real-time detection without post-processing overhead, while maintaining high accuracy. This makes YOLOv10 appear to be an ideal backbone for real-time SFOD; however, unlike Faster R-CNN and transformer-based detectors, it does not explicitly generate region proposals or object queries that provide structured object-level representations for cross-domain alignment, thereby making source-free adaptation more challenging.

The Mean-Teacher (MT) framework [38] facilitates training without ground-truth labels through teacher-student consistency, and has therefore become a



(a) Pseudo-label quality (b) Feature discriminability

Fig. 2: **Motivation for our proposed modules** on the Cityscapes → Foggy Cityscapes benchmark. **(a)** O2O pseudo-labels are precise but miss many objects; O2M pseudo-labels with standard post-processing provide broader coverage but introduce additional noise; our proposed DHF achieves the best F1-score. **(b)** Effective rank of YOLOv10 multi-scale PAN features (P3, P4, P5) under domain-shift. The dashed blue line marks the source oracle: effective rank of the source model on Cityscapes val set feature maps. The dotted green line marks the target oracle: effective rank of the target model on Foggy Cityscapes val set feature maps. (No Adaptation) denotes the effective rank of the source model on Foggy Cityscapes val set feature maps. MARD consistently recovers the effective rank best across all scales, closing the gap toward the target oracle.

fundamental component of state-of-the-art SFOD methods [12, 52, 42, 51]. However, we observe that a vanilla MT on YOLOv10 is suboptimal for two reasons: (i) As shown in Fig. 2a, the O2O head produces high-precision (0.923) pseudo-labels but tends toward low recall (0.446), missing objects. In contrast, use of the O2M head for pseudo-labeling improves coverage (recall 0.469) but introduces additional noise (precision 0.847). Using either head alone for pseudo-labeling, or directly combining head predictions remains suboptimal (Tab. 5b). (ii) The dual-head consistent assignment mechanism benefits from highly discriminative features [44], however, we hypothesize that domain-shift erodes the discriminability. To verify this, for each scale level ℓ, we reshape the feature map $F_\ell \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$ into a spatially flattened matrix $Z_\ell \in \mathbb{R}^{(H_\ell W_\ell) \times C_\ell}$, where rows correspond to spatial locations and columns correspond to channels. We compute singular values $\{\sigma_i\}$ of Z_ℓ, normalize them as $p_i = \sigma_i / \sum_j \sigma_j$, and define effective rank as $\text{erank}_\ell = \exp(-\sum_i p_i \log p_i)$, which quantifies the effective number of channel dimensions carrying independent information. Over a dataset of N images, we calculate the dataset-level mean effective rank as $\overline{\text{erank}}_\ell = \frac{1}{N}\sum_{n=1}^{N} \text{erank}_\ell^{(n)}$, and report its normalized percentage as $\text{erank}\%_\ell = 100 \times \frac{\overline{\text{erank}}_\ell}{C_\ell}$ in Fig.2b. Domain-shift erodes the effective rank of features across scales as observed in Fig.2b (No Adaptation), indicating the collapse of features toward lower-dimensional subspaces. Standard mean-teacher self-training (Adapted w/o MARD) recovers less than half of this lost diversity, placing a ceiling on adaptation performance.

To address these two bottlenecks, we introduce two modules respectively: **(DHF) Dual-Head Pseudo-Label Fusion.** Treating duplicate-free high-precision O2O predictions as anchors, we *selectively* supplement them with non-

redundant, high-scoring O2M predictions to recover objects that O2O misses. Fig.2a shows how our *DHF loss increases F-1 score by 10.7%*, leading to a gain in adaptation performance (Tab.5b). This fusion is applied *only* during training-time pseudo-label construction; at inference, the detector retains its original NMS-free O2O-only inference with no additional overhead. **(MARD) Multi-scale Adaptive Representation Diversification.** MARD enforces structured representational constraints on the multi-scale features, preserving the discriminative capacity (Fig.2b (Adapted w/ MARD)) that domain-shift otherwise erodes. Together, these components facilitate stable adaptation of YOLOv10, achieving similar or higher accuracy than state-of-the-art approaches, with 1.3x higher FPS and requiring significantly fewer parameters. Our main contributions are:

- We present, to the best of our knowledge, the first systematic study of NMS-free dual-head detectors for SFOD. Our proposed method, *RT-SFOD*, establishes a new state-of-the-art on the accuracy-speed-size frontier.
- We propose DHF, a training-time fusion of high-precision O2O and selectively filtered O2M predictions for superior pseudo-label quality, and MARD, a multi-scale feature diversification mechanism that counteracts feature rank degradation under domain-shift.
- Across domain-shift scenarios, our proposed modules yield state-of-the-art or competitive adaptation accuracy, while the efficient YOLOv10 backbone contributes 1.3x higher throughput with ∼2x fewer parameters than prior SFOD methods.

2 Related Works

Source-Free Object Detection: SFOD was introduced in [22], which identified noisy pseudo-labels as the main challenge and addressed it using entropy-based confidence thresholding and false-negative augmentation. LODS [21] applied bidirectional knowledge distillation on original and style-transferred target images with graph-based feature alignment to learn domain-invariant representations. [8] proposed A^2SFOD, which uses variance-based target splitting and adversarial alignment in a mean-teacher framework. IRG [43] adopted a mean-teacher framework with graph CNN-guided contrastive learning for target feature alignment. Simple-SFOD [12] adapted batch normalization statistics and leveraged strong augmentations with fixed pseudo-labels in a teacher-student setup for domain adaptation. FALCON-SFOD [42] showed that domain-shift weakens object-focus in features and leveraged foundation model priors to address this issue. Following [22], all above methods adopt Faster R-CNN as the base detector. Recent SFOD [19, 52] works utilize Detection Transformers (DETRs). In [19], a historical student loss is used to stabilize student-training and mitigate noisy pseudo-labels. FRANCK [52] is an SFOD framework that enhances DETR adaptation via category-level contrastive learning, instance reweighting, and uncertainty-weighted feature distillation. Faster R-CNN and DETR are computationally intensive [28], and SFOD methods built on them focus mainly on adaptation accuracy, overlooking real-time performance metrics such as FPS and model size.

Unsupervised Domain Adaptive Object Detection: UDAOD addresses domain-shift using both labeled source and unlabeled target data. Early methods focused on Faster R-CNN [4, 7, 10, 49], while later works adopted DETR-based models: MTM [48] proposed two-stage DETR adaptation with source pre-training on style-transferred images and masked feature alignment; MTTrans [55] applied a ResNet50-based DETR in a mean-teacher framework; DA-DETR [57] and SFA [47] introduced transformer detectors with dedicated alignment modules; and BiADT [25] leveraged box-coordinate queries as positional priors using a DAB-DETR backbone. UDOAD requires access to labeled source data, which limits its applicability in real-world applications where source data privacy and unavailability are major concerns, shifting the focus more towards SFOD.

Efficiency-centric Object Detection: Research on fast/lightweight detectors includes SSD-MobileNet [26, 13] and EfficientDet [37]. NanoDet [30] adopts depthwise convolutions with an anchor-free design, while the YOLO line [39, 45, 44, 18] has been progressively optimized for speed. Real-time transformer detectors include Real-Time-DETR [59], RF-DETR [32], LightWeight-DETR [6], and Quantized RT-DETR [14], but they generally trail YOLO in throughput. For instance, [44] reports RT-DETR-R18 at ∼46.5 AP on COCO with ∼4.6 ms/frame (∼217 FPS), whereas YOLOv10-S achieves similar AP (46.3) in 2.49 ms/frame (∼401 FPS), about *1.8× faster*. The same study [44] also shows Faster R-CNN and DETR variants have higher compute/latency, while YOLOv10's streamlined heads and lightweight blocks surpass YOLOv8 at higher accuracy with ∼2× fewer parameters, making it especially efficient for real-time systems.

SF-YOLO [41] is the first work to apply real-time detectors to SFOD. It adapts a single-head, NMS-based YOLO detector (YOLOv5 [39]) using mean-teacher self-training with student stabilization and learned augmentations. In contrast, our work focuses on NMS-free dual-head detectors, which bring new challenges beyond those considered in SF-YOLO, such as pseudo-label precision-recall trade-off across heads. We further provide the evidence of a previously unreported phenomenon: feature discriminability collapse under domain-shift; and provide a scale-aware regularization loss to mitigate this effect and preserve feature diversity. To our knowledge, no prior work systematically studies efficiency-driven SFOD for NMS-free detectors, leaving newer efficient architectures such as YOLOv10 [44] underexplored. We fill this gap with an SFOD framework for NMS-free dual-head detectors, instantiated primarily on YOLOv10, where our proposed modules drive adaptation accuracy gains while the throughput and latency advantages stem from the efficient YOLO backbone. We also demonstrate generality on other YOLO- and DETR-based dual-head detectors.

3 Methodology

3.1 Preliminaries

Problem statement. Let the unlabeled target dataset be $\mathcal{D}_t = \{x_i\}_{i=1}^{N}$. Given a detector pre-trained on a labeled source domain, our objective is to adapt

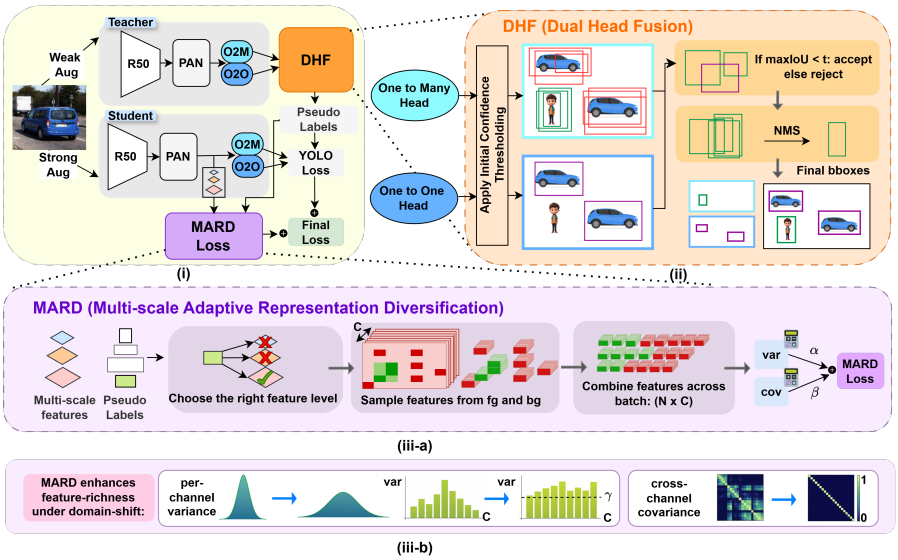


Fig. 3: **Overview of RT-SFOD.** (i) Mean-Teacher framework with DHF and MARD. (ii) **DHF:** High-precision O2O predictions serve as anchors; non-redundant O2M boxes with low overlap w.r.t. O2O anchors are selectively added to form the final pseudo-labels. (iii) **MARD:** (iii-a) Foreground and background feature vectors are sampled from multi-scale features using pseudo boxes and regularized via variance and covariance constraints, (iii-b) increasing per-channel variance for better discriminability and reducing off-diagonal covariance for greater channel diversity.

it to $\mathcal{D}_t$ *without* accessing any source samples, while maintaining a favorable accuracy-efficiency trade-off suitable for real-time use.

Mean-Teacher self-training. Most state-of-the-art SFOD methods follow the Mean-Teacher (MT) teacher-student paradigm. A teacher detector f_{θ_T} produces pseudo-labels on the target domain, which are then used to supervise a student detector f_{θ_S}. For a target image x, we construct a weakly augmented view $x^w = a_w(x)$ and a strongly augmented view $x^s = a_s(x)$. The teacher predicts pseudo-labels on x^w, and the student is trained to match them on x^s. After each step, the teacher parameters are updated via an exponential moving average (EMA) of the student with momentum μ: $\theta_T \leftarrow \mu\,\theta_T + (1-\mu)\,\theta_S$.

YOLOv10 preliminaries. We adopt YOLOv10 as the primary base detector due to its strong accuracy-efficiency trade-off and end-to-end NMS-free inference, which are desirable for real-time SFOD. YOLOv10 employs two prediction heads trained under a consistent matching scheme: a *one-to-one* (O2O) head that assigns each object to a single prediction, and a *one-to-many* (O2M) head that allows multiple matched predictions per object to strengthen supervision. YOLOv10 operates on a multi-scale feature pyramid (PAN); we denote the stu-

Algorithm 1 Proposed Real-time SFOD.

Require: AdaBN-initialized checkpoint θ^0, unlabeled target data $\mathcal{D}_t$ and hyperparameters

Initialize student $\theta_S \leftarrow \theta^0$, teacher $\theta_T \leftarrow \theta^0$
for each epoch **do**
  for each minibatch $x \sim \mathcal{D}_t$ **do**
    Form $x^w = a_w(x)$ and $x^s = a_s(x)$
    Teacher predicts $\mathcal{P}^o$ and $\mathcal{P}^m$ on x^w
    Fuse with DHF to obtain pseudo-labels $\hat{Y}^w$ (Eq. 4)
    Map $\hat{Y}^w \rightarrow \hat{Y}^s$ via the strong-view geometry
    Student predicts Y on x^s; compute detection loss $\mathcal{L}_{\text{det}}(\hat{Y}^s, Y)$
    Compute MARD loss $\mathcal{L}_{\text{mard}}$ on multi-scale PAN features (Eq. 8)
    Update θ_S using $\mathcal{L}_{\text{det}} + \lambda\mathcal{L}_{\text{mard}}$
  end for
  Update teacher: $\theta_T \leftarrow \mu\theta_T + (1-\mu)\theta_S$ (epoch-level EMA)
end for

dent features at pyramid level l as F_ℓ^S, for $\ell \in \{P3, P4, P5\}$. During inference, only the O2O head is used to produce duplicate-free detections.

Why vanilla MT is suboptimal for YOLOv10. Despite being a strong SFOD baseline, vanilla MT underperforms on YOLOv10 for two reasons. **(i)** Its dual heads yield complementary but mismatched pseudo-labels: O2O is duplicate-free and precise yet misses many objects under domain-shift, whereas use of the O2M head for pseudo-labeling improves coverage but introduces additional noise; using either alone, or directly combining them, results in suboptimal supervision. **(ii)** Domain-shift degrades YOLOv10's multi-scale feature discriminability[44], reducing effective rank across pyramid levels (Fig. 2b), and MT self-training only partially recovers this diversity. We therefore propose DHF for higher-quality pseudo-labels and MARD to preserve discriminative features.

3.2 Proposed Real-Time SFOD

Following [12], we warm-start the teacher using Adaptive Batch Normalization (AdaBN): we update BN statistics on target images to obtain a stronger target-initialized checkpoint. Teacher and student are initialized with this checkpoint before the adaptation begins. After applying AdaBN, we found that updating the teacher model once per epoch yields better performance than updating it at every training step. Algorithm 1 summarizes the adaptation loop. Below, we explain the two proposed modules of our framework.

DHF: Dual-Head Pseudo-Label Fusion Let $\mathcal{P}^o$ and $\mathcal{P}^m$ denote the teacher predictions from the O2O and O2M heads on the weak view x^w, respectively. Each prediction is a tuple $p = (b, s, c)$ where $b \in \mathbb{R}^4$ denotes box coordinates (x_1, y_1, x_2, y_2), s is the confidence, and c is the class.

We first keep high-confidence O2O predictions as anchors:

$$\hat{Y}^o = \{p \in \mathcal{P}^o \mid s(p) \geq \tau_o\}. \tag{1}$$

We then consider O2M candidates $\mathcal{C}^m = \{p \in \mathcal{P}^m \mid s(p) \geq \tau_m\}$, and add only those that have low overlap w.r.t. the O2O anchors:

$$\mathcal{E} = \left\{ p \in \mathcal{C}^m \;\middle|\; \max_{q \in \hat{Y}^o} \mathrm{IoU}(b_p, b_q) \leq \tau_{\mathrm{no}} \right\}. \tag{2}$$

This rule uses the O2O head as a high-precision prior and lets O2M contribute only non-redundant boxes, improving pseudo-label coverage while mitigating noisy supervision during self-training. Because O2M predictions can contain duplicates among themselves, we apply class-wise NMS *only* on the O2M extras:

$$\mathcal{E} \leftarrow \mathrm{NMS}(\mathcal{E}, \tau_{\mathrm{dup}}). \tag{3}$$

Finally, the fused pseudo-label set is

$$\hat{Y}^w = \hat{Y}^o \cup \mathcal{E}. \tag{4}$$

MARD: Multi-scale Adaptive Representation Diversification
Let $\{F_\ell^S\}_{\ell \in \{3,4,5\}}$ be the three feature maps input to the YOLOv10 detection head (PAN outputs), where $F_\ell \in \mathbb{R}^{B \times C_\ell \times H_\ell \times W_\ell}$. We compute MARD on these features during the student forward pass on x^s. Crucially, MARD gradients propagate through the PAN into the backbone, actively shaping feature learning during adaptation. Unlike representation-learning methods that regularize *global* embeddings[3], MARD constructs its feature vector set using *detection structure*:

- **Pseudo-box guided foreground feature vectors.** From the pseudo-labels $\hat{Y}^w$, we sample K spatial locations *inside each pseudo box* and extract the corresponding vectors from the detection feature map F_ℓ. This ties MARD to *object evidence* used by the detector.
- **Complementary background feature vectors.** We also sample M locations from *background* regions (valid, non-padded pixels outside pseudo boxes), encouraging object-background separability in the same feature space.
- **Noise control under SFOD.** We restrict sampling to the top-K_b pseudo boxes with confidence above τ_b. This improves robustness to pseudo-label noise, and the adaptive weighting in Eq. (9) further downweights MARD when pseudo-label quality is low.

YOLO-style detectors distribute predictions across feature pyramid levels (PAN) according to object size. To keep MARD detection consistent, we assign each pseudo box b to a feature level based on its size $s(b) = \sqrt{w(b)h(b)}$ in pixels and stride-derived thresholds:

$$g(b) = \begin{cases} 3 & s(b) \leq \eta \cdot \mathrm{stride}_3, \\ 4 & \eta \cdot \mathrm{stride}_3 < s(b) \leq \eta \cdot \mathrm{stride}_4, \\ 5 & s(b) > \eta \cdot \mathrm{stride}_4, \end{cases} \tag{5}$$

where η is a constant and stride_ℓ is the effective stride of level ℓ. We then sample foreground/background feature vectors *per level* using only boxes assigned to

that level. This prevents MARD from mixing incompatible scales and directly targets the scale-specific features used for small/medium/large objects.

For each level ℓ, we collect a feature vector matrix $Z_\ell \in \mathbb{R}^{N_\ell \times C_\ell}$, containing foreground and background feature vectors. We then apply two complementary constraints: (i) a *variance* term to prevent feature collapse (ensuring each channel maintains a non-trivial spread), and (ii) a *covariance* term to reduce redundancy across channels.

$$\mathcal{L}_{\text{var}}(Z_\ell) = \frac{1}{C}\sum_{c=1}^{C} \max\left(0, \gamma - \sqrt{\text{Var}(Z_{\ell,:,c}) + \epsilon}\right), \tag{6}$$

$$\mathcal{L}_{\text{cov}}(Z_\ell) = \frac{1}{C(C-1)}\sum_{i\neq j}\left(\text{Cov}(\tilde{Z}_\ell)_{ij}\right)^2, \tag{7}$$

where $\tilde{Z}_\ell$ denotes channel-wise normalized tokens, γ is a hyperparameter and ϵ is a small constant for stability. While the variance and covariance terms share a functional form with collapse-prevention objectives used in self-supervised learning [3], MARD differs in that it operates on *multi-scale PAN features* rather than global embeddings, constructs feature vector sets via *pseudo-box guided foreground and background sampling* tied to detection structure, and enforces *scale-consistent* regularization via the level assignment in Eq. (5).
The MARD loss is defined as:

$$\mathcal{L}_{\text{mard}} = \sum_{\ell\in\{3,4,5\}} \left(\alpha\,\mathcal{L}_{\text{var}}(Z_\ell) + \beta\,\mathcal{L}_{\text{cov}}(Z_\ell)\right). \tag{8}$$

To avoid over-regularizing when pseudo-labels are unreliable, we use an adaptive weight

$$\lambda(t) = \lambda_0 \cdot \text{ramp}(t) \cdot \text{gate}(\bar{s}), \tag{9}$$

where $\text{ramp}(t)$ linearly increases from 0 to 1 over a warmup period, and $\text{gate}(\bar{s})$ increases with the batch-average pseudo-label confidence $\bar{s}$ (clipped to $[0, 1]$). We also compute MARD periodically (every I steps) to limit overhead.

3.3 Training Objective and Inference

Given the student outputs on x^s and pseudo-labels $\hat{Y}^w$ mapped to strong-view $\hat{Y}^s$, we first calculate the standard YOLOv10 detection loss (box regression, classification, and distribution focal loss), computed over *both* the O2O and O2M student heads to retain the dual-head training dynamics of YOLOv10:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dfl}}. \tag{10}$$

Total loss. Our final optimization objective is:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda(t)\,\mathcal{L}_{\text{mard}}. \tag{11}$$

Inference-time efficiency. Dual-Head Fusion and MARD are used only during *training-time* adaptation. After adaptation, we deploy the student f_{θ_S} as a standard YOLOv10 detector, preserving **end-to-end NMS-free inference** and incurring *no additional runtime cost*.

Table 1: Comparison with state-of-the-art on **C2F** (Cityscapes → Foggy Cityscapes). **Bold** denotes the best SFOD result; underline the second-best. Efficiency metrics are measured using each method's official public codebase. †FRANCK has no public implementation; its efficiency figures are adopted from DRU, which shares the same base detector. S = Source-only; UDAOD = Unsupervised DA OD; SFOD = Source-Free OD. All our results are averaged over three random seeds.

Category	Method	Base	Params (M)	FPS	Lat. (ms)	prsn	rider	car	truck	bus	train	mcycle	bicycle	mAP
S	Source only	YOLOv10S	7.2	233	4.3	41.0	46.1	49.6	12.0	27.0	4.9	17.1	39.0	29.6
		YOLOv10M	15.4	105	9.5	43.6	47.8	50.0	12.8	29.8	5.0	20.1	40.2	31.2
		YOLOv10L	24.4	67	15.0	47.1	52.4	58.5	18.6	35.1	10.8	24.5	44.8	36.5
UDAOD	AT (CVPR'22)	FRCNN	NA	NA	NA	43.7	54.1	62.3	31.9	54.4	49.3	35.2	47.9	47.4
	MRT (ICCV'23)	DETR	NA	NA	NA	52.8	51.7	68.7	35.9	58.1	54.5	41.0	47.1	51.2
	CAT (CVPR'24)	FRCNN	NA	NA	NA	44.6	57.1	63.7	40.8	66.0	49.7	44.9	53.0	52.5
	DATR (TIP'25)	DETR	NA	NA	NA	61.6	60.4	74.3	35.7	60.3	35.4	43.6	55.9	53.4
	SEEN-DA (CVPR'25)	FRCNN	NA	NA	NA	58.5	64.5	71.7	42.0	61.2	54.8	47.1	59.9	57.5
SFOD	IRG (CVPR'23)	FRCNN	34.0	51	19.5	37.4	45.2	51.9	24.4	39.6	25.2	31.5	41.6	37.1
	PETS (ICCV'23)		44.2	48	20.6	42.0	48.7	56.3	19.3	39.3	5.5	34.2	41.6	35.9
	Simple-SFOD (ECCV'24)		43.8	42	23.8	40.9	48.0	58.9	29.6	51.9	50.2	36.2	44.1	45.0
	FALCON-SFOD (CVPR'26)		43.8	42	23.8	41.0	48.3	58.7	33.6	54.8	**54.3**	38.6	46.2	46.9
	SF-YOLO (ECCV'24 ws)	YOLOv5S	7.2	172	5.8	48.6	51.8	64.5	24.7	48.1	37.6	23.1	44.8	42.9
	SF-YOLO (ECCV'24 ws)	YOLOv5L	46.5	52	19.1	55.5	58.0	**71.5**	36.6	53.7	46.1	**40.6**	50.5	51.6
	DRU (ECCV'24)	Def-DETR	40.0	29	35.0	48.3	51.5	62.5	26.2	43.2	34.1	34.2	48.6	43.6
	FRANCK (TIP'25)		40.0†	29†	35†	48.1	49.3	60.6	33.9	48.2	36.9	34.0	47.9	44.9
	VFM-SFOD (AAAI'26)		41.0	27	37.0	48.8	51.0	62.8	36.6	52.7	39.9	36.4	48.2	47.1
	RT-SFOD (Ours)	YOLOv10S	**7.2**	**233**	**4.3**	47.6	52.6	64.8	25.9	46.0	41.5	30.4	45.2	44.3
		YOLOv10M	**15.4**	**105**	**9.5**	53.7	57.4	67.3	26.1	46.3	44.9	35.8	47.3	47.4
		YOLOv10L	**24.4**	**67**	**15**	**58.4**	**64.7**	71.0	**38.5**	**57.0**	49.5	39.0	**52.1**	**53.8**

4 Experiments

Datasets and Metrics. Following existing works, we use five publicly available datasets covering four domain-shift scenarios: Cityscapes[9], Foggy Cityscapes[34], KITTI[11], Sim10k[16], and BDD100k[54]. Detailed description of datasets is provided in supplementary (Sec. S.4.2). Unlike existing SFOD methods, which only report target domain accuracy (mean average precision mAP at 0.5 IoU threshold), we report mAP along with *efficiency metics* like throughout (FPS), number of parameters, and latency (ms/img). As noted in [32], many efficiency-focused methods [59, 41, 44] report FP32 PyTorch accuracy but FP16 TensorRT efficiency, leading to an unfair comparison. Moreover, [32] shows that naive FP16 quantization can significantly degrade performance. Therefore, we report all primary results in FP32 using PyTorch, and provide FP16 TensorRT results in the supplementary (Sec. S.8) for comparability.

Implementation details. The student is trained using SGD with an initial learning rate of 1e-4 and cosine annealing, gradient clipping at norm 10, and batch size 16 for 60 epochs. The teacher is updated via Exponential Moving Average (EMA) with momentum 0.999 at the end of each epoch. All experiments are conducted on a single NVIDIA RTX A6000 GPU. We use the same hyperparameters across all domain-shifts. Details about hyperparameters are included in the supplementary (Sec. S.4.1).

4.1 Comparison with State-of-the-Art Methods

Following existing works, we compare our method with (i)UDA-based object detection (UDAOD), which retains source data during adaptation [23, 60, 17, 5, 20], and (ii) SFOD methods [43, 24, 12, 19, 52, 42, 51, 41].

Cityscapes → Foggy Cityscapes (C2F). Table 1 presents results on the most widely used SFOD benchmark. Our RT-SFOD-L achieves **53.8 mAP**, surpassing the previous best SF-YOLO-L (51.6) by +2.2 mAP and remaining competitive with UDAOD methods (MRT: 51.2, DATR: 53.4) under the stricter source-free constraint. Our RT-SFOD-M (47.4 mAP) outperforms most Faster R-CNN and Deformable DETR based SFOD approaches. Compared to SF-YOLO-L, RT-SFOD-L is 1.3× faster (67 vs. 52 FPS) and 1.9× smaller (24.4M vs. 46.5M parameters). Against Deformable DETR-based methods (DRU, FRANCK, VFM-SFOD), which cluster around 27–29 FPS and 40–41M parameters, RT-SFOD-M matches their best accuracy at **3.6× higher throughput** and **2.6× fewer parameters**. Our smallest variant, RT-SFOD-S (44.3 mAP), is on par with DRU (43.6) and FRANCK (44.9) while running at **233 FPS** with only **7.2M** parameters. These trade-offs are visualized in Fig. 1.

Table 2: S2C and K2C comparison. Efficiency metrics are identical to those in Table 1

Cat.	Method	Base	S2C	K2C
S	Source only	YOLOv10S	51.2	37.4
		YOLOv10M	53.5	40.3
		YOLOv10L	56.5	45.0
UDA	DATR[5]	DETR	64.3	53.3
	SEEN-DA[20]	FRCNN	66.8	67.1
SFOD	IRG[43]		45.2	46.9
	PETS[24]		57.8	47.0
	LPLD[53]	FRCNN	49.4	51.3
	Simple[12]		55.4	46.2
	FALCON[42]		58.8	50.1
	SF-YOLO[41]	YOLOv5S	57.7	50.6
	SF-YOLO[41]	YOLOv5L	<u>69.8</u>	**63.7**
	DRU[19]		58.7	45.1
	FRANCK[52]	DefDETR	63.1	48.5
	VFM[51]		67.4	54.7
	RT-SFOD (Ours)	YOLOv10S	59.8	52.3
		YOLOv10M	66.4	54.2
		YOLOv10L	**71.2**	<u>60.9</u>

Sim10k→Cityscapes (S2C) and KITTI →Cityscapes (K2C). Following prior works, we report AP(on Car) for S2C and K2C. On S2C, our RT-SFOD-L achieves **71.2 mAP**, outperforming SF-YOLO-L (69.8) by +1.4 and VFM-SFOD (67.4) by +3.8. On K2C, it reaches <u>60.9 mAP</u>, second only to SF-YOLO-L (63.7), while using **1.9× fewer parameters** and **1.3× higher throughput**. We attribute the gap to SF-YOLO-L to model capacity: SF-YOLO-L uses 1.9× more parameters (46.5M vs. 24.4M), and at matched scale (both 7.2M), RT-SFOD-S outperforms SF-YOLO-S (52.3 vs. 50.6). Moreover, RT-SFOD-S/M outperform all five Faster R-CNN methods and two of three DETR-based methods on K2C while being faster and smaller, indicating that this camera-shift scenario particularly benefits from additional model capacity. The S→M→L scaling yields 59.8→66.4→71.2 on S2C and 52.3→54.2→60.9 on K2C, demonstrating consistent gains as model size increases.

Cityscapes → BDD100k (C2B). Table 3 presents a challenging large-scale benchmark spanning diverse cities, weather conditions, and times of day. Our RT-SFOD-L achieves **46.5 mAP**, exceeding VFM-SFOD (43.0) by +3.5 and FALCON-SFOD (36.9) by +9.6, and also beating both UDAOD baselines. Even RT-SFOD-S (41.1 mAP) is competitive with heavyweight Def-DETR methods such as FRANCK (40.7). The consistent gains across C2F (weather), C2B (large-

Table 3: Comparison with state-of-the-art on C2B (Cityscapes → BDD100k). Efficiency metrics are identical to those in Table 1.

Category	Method	Base	truck	car	rider	person	motor	bicycle	bus	mAP
S	Source only	YOLOv10S	17.5	59.6	29.1	42.4	14.6	19.3	20	28.9
		YOLOv10M	18.8	61.8	32	44.5	15.8	22.6	21.5	31
		YOLOv10L	23.1	68.3	35.4	49.2	17.8	24	25.3	34.7
UDA	MRT[60] (ICCV'23)	Def-DETR	24.7	63.7	30.9	48.4	20.2	22.6	25.5	33.7
	DATR[5] (TIP'25)	Def-DETR	26.9	73.4	42.8	58.5	24.2	37.3	39.9	43.3
SFOD	IRG[43] (CVPR'23)	FRCNN	31.4	59.7	32.8	39.9	16.7	26.9	21.5	32.7
	PETS[24] (ICCV'23)	FRCNN	19.3	62.4	34.5	42.6	17.0	26.3	16.9	31.3
	Simple-SFOD[12] (ECCV'24)	FRCNN	32.0	60.0	33.4	40.2	19.7	29.9	24.9	34.3
	FALCON-SFOD[42]	FRCNN	32.6	59.8	34.0	40.0	25.7	**35.7**	30.5	36.9
	DRU[19] (ECCV'24)	Def-DETR	27.1	62.7	36.9	45.8	22.7	32.5	28.1	36.6
	FRANCK[52]	Def-DETR	29.5	65.6	43.3	55.0	30.0	28.0	33.6	40.7
	VFM-SFOD[51]	Def-DETR	33.2	72.3	44.2	54.9	32.9	29.0	34.8	43.0
	RT-SFOD (Ours)	YOLOv10S	28.8	71.6	44.8	53.8	28.7	26.5	33.6	41.1
		YOLOv10M	33.7	75.9	47.3	57.5	33	28.6	35.8	44.5
		YOLOv10L	**36.4**	**77.0**	**49.3**	**58**	**34.1**	31.4	**39.6**	**46.5**

scale diversity), S2C (rendering gap), and K2C (camera shift) indicate that our modules generalize across domain gap characteristics.

4.2 Ablation Study

Table 4: Ablation study of the proposed components with RT-SFOD-M across domain-shifts C2F, S2C, K2C.

Module	Components		C2F									S2C	K2C
	DHF	MARD	prsn	rider	car	truck	bus	train	mcycle	bicycle	mAP	AP	AP
Source-trained	×	×	45.6	48.8	50	7.48	27	3	20.1	42.2	30.5	52.5	40.3
AdaBN	×	×	50.9	51.4	57.9	15.7	35.8	15.5	26.6	45.1	37.4	54.2	45.2
MT (O2M)	×	×	50.3	54.5	63.7	24.3	45.0	41.1	33.5	45.3	44.7	61.4	49.9
MT (O2O)	×	×	51.5	55.0	63.8	24.9	45.2	41.7	33.8	45.9	45.2	62.1	50.6
MT (O2O) + DHF	✓	×	52.9	56.4	67.0	25.5	45.6	44.3	35.2	46.6	46.7	65.1	53.0
MT (O2O) + MARD	×	✓	51.8	55.7	66.2	25.2	45.9	44.7	35.0	46.2	46.3	65.3	52.5
MT (O2O) + DHF + MARD	✓	✓	53.7	57.4	67.3	26.1	46.3	44.9	35.8	47.3	**47.4**	**66.4**	**54.2**

Table 4 presents a component-wise ablation with RT-SFOD-M across three benchmarks. Starting from the source-trained model (30.5/52.5/40.3 mAP on C2F/ S2C/K2C), AdaBN provides a better starting point for the teacher by aligning batch normalization statistics before training. Adding mean-teacher (MT) self-training yields further gains: MT with one-to-many (O2M) pseudo-labels achieves 44.7 mAP on C2F, while MT with one-to-one (O2O) obtains 45.2. The O2O head's slight advantage is consistent with our precision-recall analysis in Fig. 2a: while O2M with standard post-processing provides broader coverage (recall 0.469 vs. 0.446), it introduces additional noise (precision 0.847 vs. 0.923) that can compound during iterative self-training. However, despite these gains, the vanilla MT framework remains sub-optimal.

Effect of DHF. Adding Dual-Head Fusion to MT (O2O) yields consistent gains: +1.5 on C2F (45.2 → 46.7), +3.0 on S2C (62.1 → 65.1), and +2.4 on K2C (50.6 → 53.0), with the largest per-class gains on car (+3.2) and train (+2.6).
Effect of MARD. Adding MARD loss (w/o DHF) yields +1.1 on C2F (45.2 → 46.3), +3.2 on S2C (62.1 → 65.3), and +1.9 on K2C (50.6 → 52.5). As shown in Fig. 2b, MARD recovers 76–84% of the effective rank lost to domain-shift at each PAN-feature level (vs. only 26–48% without it), restoring feature discriminability.
Complementarity of DHF and MARD. Combining both modules achieves the best results: **47.4/66.4/54.2** on C2F/S2C/K2C, with total gains of +2.2/ +4.3/+3.6 over MT (O2O). The combined improvement exceeds either module in isolation, confirming they address *complementary bottlenecks*: DHF improves the supervisory signal while MARD improves learned representations.

Table 5: Ablation studies on RT-SFOD-M. Left: MARD sub-objectives. Right: Pseudo-label strategy comparison.

(a) Ablation of MARD components. Baseline: MT (O2O) + DHF.

Variance	Covariance	C2F	S2C	K2C
×	×	46.7	65.1	53.0
✓	×	47.2	65.9	53.8
×	✓	47.0	65.7	53.5
✓	✓	**47.4**	**66.4**	**54.2**

(b) Pseudo-label strategy comparison (RT-SFOD-M, C2F).

Strategy	Prec	Recall	F1	mAP
O2O only	0.923	0.446	0.602	45.2
O2M+NMS	0.847	0.469	0.603	45.2
O2O ∪ O2M+NMS	0.834	0.512	0.634	45.5
DHF (Ours)	0.919	0.525	**0.668**	**46.7**

4.3 Analysis

MARD sub-component analysis. Table 5a ablates the two MARD components using MT (O2O) + DHF as the fixed baseline. The variance term (preventing channel collapse) provides the larger individual contribution (+0.5/+0.8/+0.8 on C2F/S2C/K2C), consistent with channel collapse being the primary mode of rank degradation in Fig. 2b. The covariance term (reducing inter-channel redundancy) provides a complementary gain (+0.3/+0.6/+0.5). Combining both yields the full MARD improvement (+0.7/+1.3/+1.2), confirming they address distinct failure modes.
Pseudo-label fusion strategy comparison. Table 5b compares four pseudo-label strategies on label-quality metrics and downstream mAP. O2O yields high precision (0.923) but low recall (0.446), while O2M + NMS provides broader coverage (0.469) at lower precision (0.847), giving a comparable F1 (0.603). Directly merging O2O with O2M + NMS improves recall (0.512) but reduces precision (0.834), yielding a suboptimal F1 (0.634). Our DHF uses O2O boxes as high-confidence anchors and selectively adds non-redundant, high-scoring O2M boxes with low overlap w.r.t. O2O anchors, retaining near-O2O precision (0.919) while boosting recall to 0.525, achieving the best F1 (0.668) and mAP (46.7). This confirms that pseudo-label selection must balance precision and recall in

self-training: the direct union improves recall but reduces mAP, as noisy pseudo-labels can compound across iterations.

Hyperparameter sensitivity. Fig. 4 shows C2F mAP of RT-SFOD-M as a function of four key hyperparameters, with performance stable across all tested ranges ($\leq$1.5 mAP variation). Both confidence thresholds (τ_{o2o}, τ_{o2m}) peak at 0.5; the τ_{o2m} curve is notably flatter (46.0–47.4), indicating DHF's IoU-based deduplication provides robustness against the O2M threshold choice. The optimal MARD weight is λ_0=0.05, with performance degrading gracefully at both extremes; the effective weight is further modulated by a linear warmup and confidence gate (Sec. 3.2). EMA momentum μ=0.999 balances teacher stability and responsiveness. No hyperparameter requires per-benchmark tuning, and the same values transfer across detectors and scales, likely because they depend on relative statistics (calibrated confidence, geometric IoU, and feature distributions) rather than architecture-specific factors. The full set of values used is: τ_{o2o}=τ_{o2m}=0.5, τ_{no}=0.2, τ_{dup}=0.7, λ_0=0.05, K=8, M=128, μ=0.999; full sensitivity analysis is in the supplementary (Sec. S.4.1).

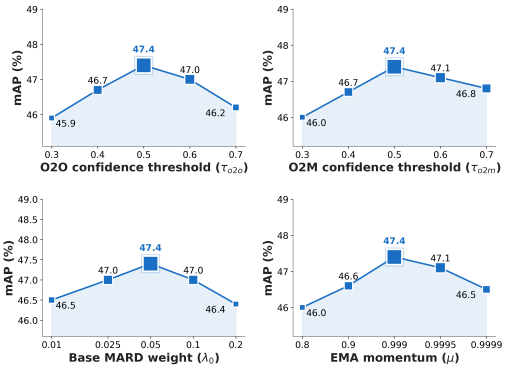


Fig. 4: **Hyperparameter sensitivity** on C2F (RT-SFOD-M). mAP (%) vs. O2O threshold τ_{o2o}, O2M threshold τ_{o2m}, MARD weight λ_0, and EMA momentum μ. Variation is $\leq$1.5 mAP in all cases.

Feature Discriminability Analysis. Fig. 5 illustrates the average cosine similarity between each anchor point's extracted features and all others across PAN scales (P3, P4, and P5) on the Foggy Cityscapes validation set. We compare the Source model, Vanilla Mean-Teacher, and the proposed MARD approach. Notably, MARD consistently yields lower cosine similarity across all scales, indicating more structured feature representations and enhanced discriminability.

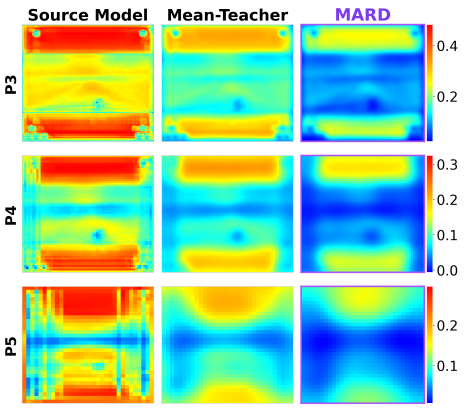


Fig. 5: The average cosine similarity of each anchor point's extracted features with all others across PAN scales(P3, P4, P5) on Foggy Cityscapes validation set.

Generality across Dual-Head Detectors. While our main experiments use YOLOv10, RT-SFOD can be applied to any NMS-free dual-head de-

tector. DHF requires only a precise O2O head and a dense, high-recall O2M head, and MARD applies to any multi-scale feature pyramid. Table 6 presents the analysis of RT-SFOD applied with the same hyperparameters to three additional dual-head detectors on C2F.

As shown in Table 6, RT-SFOD improves MT (O2O) consistently across YOLOv26S/M/L [40] (+2.5/+2.8/+2.1), MS-DETR [58] (+2.0), and Mr. DETR [56] (+2.3), matching the gains observed on YOLOv10. This confirms that the benefits of DHF and MARD are properties of the dual-head paradigm rather than of a specific architecture. For single-head detectors, DHF has no second candidate set and reduces to confidence-threshold pseudo-labeling as in SF-YOLO [41], while MARD remains applicable whenever pseudo-labels and multi-scale features are available.

Table 6: Generality of RT-SFOD across dual-head detectors on C2F domain-shift.

Detector	MT (O2O)	RT-SFOD
YOLOv26S [40]	42.1	**44.6** (+2.5)
YOLOv26M [40]	46.0	**48.8** (+2.8)
YOLOv26L [40]	50.6	**52.7** (+2.1)
MS-DETR [58]	42.3	**44.3** (+2.0)
Mr. DETR [56]	43.8	**46.1** (+2.3)

5 Conclusion

We present **RT-SFOD**, a source-free object detection framework for NMS-free dual-head detectors, instantiated primarily on YOLOv10, that improves the accuracy-speed-model size trade-off for domain-adaptive detection. We identify two challenges in directly applying mean-teacher self-training to NMS-free dual-head detectors: (i) suboptimal pseudo-label supervision from unconditioned use of O2O and O2M predictions, and (ii) feature rank degradation under domain-shift. To address these, we propose **Dual-Head Fusion**, which uses high-precision O2O predictions as anchors and selectively incorporates non-redundant O2M candidates with low overlap relative to O2O anchors to increase recall while maintaining near-O2O precision. We also introduce **Multi-scale Adaptive Representation Diversification**, which enforces structured variance and covariance constraints on PAN features, recovering 76–84% of rank lost under domain-shift compared to 26–48% without it. Across domain-shift benchmarks, RT-SFOD improves mAP by 1.4–3.5% over prior state-of-the-art methods, while the 1.3× higher throughput and smaller parameter count are inherited from the efficient YOLO backbone. We also demonstrate that our modules generalize across other dual-head detectors. Failure-case analysis and limitations are provided in the supplementary.

Acknowledgements

We gratefully acknowledge the support of the Ministry of Electronics and Information Technology and the Ministry of Education, Government of India, as well as IIT Hyderabad (India) and MBZUAI (Abu Dhabi) for their support of this project. We also sincerely thank the anonymous reviewers, area chairs, and program chairs for their valuable feedback, which helped improve the quality and presentation of the paper.

References

1. Aljalbout, E., Xing, J., Romero, A., Akinola, I., Garrett, C.R., Heiden, E., Gupta, A., Hermans, T., Narang, Y., Fox, D., et al.: The reality gap in robotics: Challenges, solutions, and best practices. Annual Review of Control, Robotics, and Autonomous Systems **9** (2025)
2. Ashraf, T., Bashir, J.: Titan: Query-token based domain adaptive adversarial learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 250–262 (2025)
3. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)
4. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8869–8878 (2020)
5. Chen, L., Han, J., Wang, Y.: Datr: Unsupervised domain adaptive detection transformer with dataset-level adaptation and prototypical alignment. IEEE Transactions on Image Processing **34**, 982–994 (2025)
6. Chen, Q., Su, X., Zhang, X., Wang, J., Chen, J., Shen, Y., Han, C., Chen, Z., Xu, W., Li, F., et al.: Lw-detr: A transformer replacement to yolo for real-time detection. arXiv preprint arXiv:2406.03459 (2024)
7. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
8. Chu, Q., Li, S., Chen, G., Li, K., Li, X.: Adversarial alignment for source free object detection. Proceedings of the AAAI Conference on Artificial Intelligence **37**(1), 452–460 (Jun 2023)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
10. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4091–4101 (June 2021)
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
12. Hao, Y., Forest, F., Fink, O.: Simplifying source-free domain adaptation for object detection: Effective self-training strategies and performance insights. In: European Conference on Computer Vision. pp. 196–213. Springer (2024)
13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Huo, Y., Wu, T., Shen, Y., Li, X., Tao, Z., Yang, D.: Qrt-detr: Post-training quantization for real-time detection transformer. Neurocomputing **661**, 131957 (2026)
15. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), `https://github.com/ultralytics/ultralytics`
16. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016)
17. Kennerley, M., Wang, J.G., Veeravalli, B., Tan, R.T.: Cat: Exploiting inter-class dynamics for domain adaptive object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16541–16550 (2024)

18. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
19. Khanh, T.L.B., Nguyen, H.H., Pham, L.H., Tran, D.N.N., Jeon, J.W.: Dynamic retraining-updating mean teacher for source-free object detection. In: European Conference on Computer Vision. pp. 328–344. Springer (2024)
20. Li, H., Zhang, R., Yao, H., Zhang, X., Hao, Y., Song, X., Peng, S., Zhao, Y., Zhao, C., Wu, Y., et al.: Seen-da: Semantic entropy guided domain-aware attention for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25465–25475 (2025)
21. Li, S., Ye, M., Zhu, X., Zhou, L., Xiong, L.: Source-free object detection by learning to overlook domain style. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8014–8023 (June 2022)
22. Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., Zhuang, Y.: A free lunch for unsupervised domain adaptive object detection without source data. Proceedings of the AAAI Conference on Artificial Intelligence **35**, 8474–8481 (05 2021)
23. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7581–7590 (2022)
24. Liu, Q., Lin, L., Shen, Z., Yang, Z.: Periodically exchange teacher-student for source-free object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
25. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: International Conference on Learning Representations (2022)
26. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 9905, pp. 21–37. Springer (2016)
27. Ma, C., Wang, N., Zhao, Z., Chen, Q.A., Shen, C.: Slowperception: Physical-world latency attack against visual perception in autonomous driving. arXiv preprint arXiv:2406.05800 (2024)
28. Naufaldihanif, R., Kurniawan, D., Tania, K.: Performance analysis of yolo, faster r-cnn, and detr for automated personal protective equipment detection. Journal of Applied Informatics and Computing **9**, 3810–3820 (12 2025)
29. Noori, M., Hakim, G.A.V., Osowiechi, D., Shakeri, F., Bahri, A., Yazdanpanah, M., Dastani, S., Ben Ayed, I., Desrosiers, C.: Histopath-c: Towards realistic domain shifts for histopathology vision-language adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4890–4900 (2026)
30. RangiLyu: Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model. (2021)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
32. Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: Rf-detr: Neural architecture search for real-time detection transformers (2025)
33. Safdari, R., Nikouei Mahani, M.A., Koohi-Moghadam, M., Bae, K.T.: Mixstyleflow: Domain generalization in medical image segmentation using normalizing flows. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376–385. Springer (2025)

34. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**, 973–992 (2018)
35. Sinha, S., Dwivedi, S., Azizian, M.: Towards deterministic end-to-end latency for medical ai systems in nvidia holoscan. arXiv preprint arXiv:2402.04466 (2024)
36. Sun, T., Segu, M., Postels, J., Wang, Y., Van Gool, L., Schiele, B., Tombari, F., Yu, F.: Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21371–21382 (2022)
37. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
38. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
39. Ultralytics: Ultralytics yolov5. urlhttps://github.com/ultralytics/yolov5 (Dec 2020), accessed: [Insert date here]
40. Ultralytics: YOLO26 (2025), `https://github.com/ultralytics/ultralytics`
41. Varailhon, S., Aminbeidokhti, M., Pedersoli, M., Granger, E.: Source-free domain adaptation for yolo object detection. In: European Conference on Computer Vision. pp. 218–235. Springer (2024)
42. VCR, S., Lalla, R., Dayal, A., Kulkarni, T., Lalla, A., Balasubramanian, V.N., Khan, M.H.: Foundation model priors enhance object focus in feature space for source-free object detection. arXiv preprint arXiv:2512.17514 (2025)
43. Vibashan, V., Oza, P., Patel, V.M.: Instance relation graph guided source-free domain adaptive object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2023)
44. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. Advances in neural information processing systems **37**, 107984–108011 (2024)
45. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
46. Wang, S., Xia, C., Lv, F., Shi, Y.: Rt-detrv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In: WACV. pp. 1628–1636 (2025)
47. Wang, W., Cao, Y., Zhang, J., He, F., Zha, Z.J., Wen, Y., Tao, D.: Exploring sequence feature alignment for domain adaptive detection transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 1730–1738. MM '21, Association for Computing Machinery, New York, NY, USA (2021)
48. Weng, W., Yuan, C.: Mean teacher detr with masked feature alignment: a robust domain adaptive detection transformer framework. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24, AAAI Press (2024)
49. Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9342–9351 (2021)
50. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. Advances in Neural Information Processing Systems **35**, 4203–4217 (2022)

51. Yao, H., Zhao, S., Li, P., Cui, Y., Lu, S., Guo, W., Lu, Y., Xu, Y., Xiong, H.: Beyond boundaries: Leveraging vision foundation models for source-free object detection. arXiv preprint arXiv:2511.07301 (2025)
52. Yao, H., Zhao, S., Lu, S., Chen, H., Li, Y., Liu, G., Xing, T., Yan, C., Tao, J., Ding, G.: Source-free object detection with detection transformer. IEEE Transactions on Image Processing **34**, 5948–5963 (2025)
53. Yoon, I., Kwon, H., Kim, J., Park, J., Jang, H., Sohn, K.: Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In: European Conference on Computer Vision. pp. 337–353. Springer (2024)
54. Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T., et al.: Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687 **2**(5), 6 (2018)
55. Yu, J., Liu, J., Wei, X., Zhou, H., Nakata, Y., Gudovskiy, D., Okuno, T., Li, J., Keutzer, K., Zhang, S.: Mttrans: Cross-domain object detection with mean-teacher transformer. In: ECCV (2024)
56. Zhang, C.B., Zhong, Y., Han, K.: Mr. detr: Instructive multi-route training for detection transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9933–9943 (June 2025)
57. Zhang, J., Huang, J., Luo, Z., Zhang, G., Zhang, X., Lu, S.: Da-detr: Domain adaptive detection transformer with information fusion. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 23787–23798 (2021)
58. Zhao, C., Sun, Y., Wang, W., Chen, Q., Ding, E., Yang, Y., Wang, J.: Ms-detr: Efficient detr training with mixed supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17027–17036 (June 2024)
59. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16965–16974 (2024)
60. Zhao, Z., Wei, S., Chen, Q., Li, D., Yang, Y., Peng, Y., Liu, Y.: Masked retraining teacher-student framework for domain adaptive object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19039–19049 (2023)
61. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)