# Beyond Boundaries: Leveraging Vision Foundation Models for Source-Free Object Detection

**Huizai Yao[1], Sicheng Zhao[2], Pengteng Li[1], Yi Cui[1], Shuo Lu[3], Weiyu Guo[1], Yunfan Lu[1], Yijie Xu[1], Hui Xiong[1,4*]**

[1] Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou), China
[2] Department of Psychological and Cognitive Sciences, Tsinghua University, China
[3] NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences, China
[4] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
huizai.yao@connect.hkust-gz.edu.cn

## Abstract

Source-Free Object Detection (SFOD) aims to adapt a source-pretrained object detector to a target domain without access to source data. However, existing SFOD methods predominantly rely on internal knowledge from the source model, which limits their capacity to generalize across domains and often results in biased pseudo-labels, thereby hindering both transferability and discriminability. In contrast, Vision Foundation Models (VFMs), pretrained on massive and diverse data, exhibit strong perception capabilities and broad generalization, yet their potential remains largely untapped in the SFOD setting. In this paper, we propose a novel SFOD framework that leverages VFMs as external knowledge sources to jointly enhance feature alignment and label quality. Specifically, we design three VFM-based modules: (1) Patch-weighted Global Feature Alignment (PGFA) distills global features from VFMs using patch-similarity-based weighting to enhance global feature transferability; (2) Prototype-based Instance Feature Alignment (PIFA) performs instance-level contrastive learning guided by momentum-updated VFM prototypes; and (3) Dual-source Enhanced Pseudo-label Fusion (DEPF) fuses predictions from detection VFMs and teacher models via an entropy-aware strategy to yield more reliable supervision. Extensive experiments on six benchmarks demonstrate that our method achieves state-of-the-art SFOD performance, validating the effectiveness of integrating VFMs to simultaneously improve transferability and discriminability.

**Extended version** — https://arxiv.org/abs/2511.07301

## 1 Introduction

Object detection aims to localize and classify objects in images, and modern detectors perform well when training and test data follow the same distribution. In practice, however, domain shifts between labeled source data and unlabeled target data often cause severe performance degradation. To alleviate this issue, Domain Adaptive Object Detection (DAOD) has been widely studied. DAOD aims to transfer a detector trained on labeled source data to unlabeled target data. However, their dependence on source data limits their applicability in scenarios involving privacy, security, or data transmission constraints (Li et al. 2024a; Lu et al. 2025b; Zhao et al. 2023). To address this, Source-Free Object Detection (SFOD) has emerged as a promising alternative, aiming to adapt a source-pretrained detector to a target domain using only unlabeled target data (VS, Oza, and Patel 2023; Chu et al. 2023; Yoon et al. 2024; Hao, Forest, and Fink 2024; Khanh et al. 2024; Zhao et al. 2024).

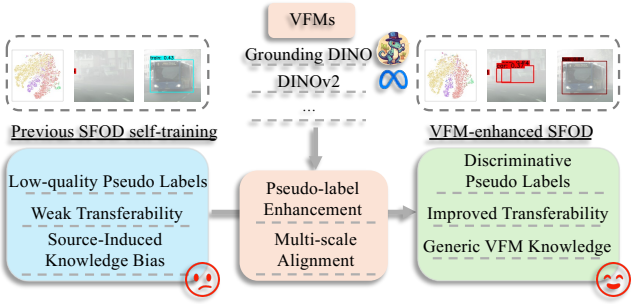


Figure 1: Illustration of our VFM-enhanced SFOD motivation. Our method leverages general VFM knowledge in VFMs such as DINOv2 (visual encoder) and Grounding DINO (vision-language detector) to address pseudo-label bias and multi-scale feature misalignment, resulting in improved transferability and discriminability compared with previous self-training pipelines of SFOD.

Previous research has identified two critical properties for effective cross-domain adaptation: transferability and discriminability (Kundu et al. 2022; Li et al. 2024a). Transferability ensures that feature representations remain domain-invariant, while discriminability guarantees separability between object categories. Despite this, current SFOD methods primarily focus on knowledge distillation within the teacher-student paradigm through feature alignment (Li et al. 2022a), adversarial learning (Chu et al. 2023), or pseudo-label refinement (Chen, Wang, and Zhang 2023; Zhao et al. 2024), without breaking out of the internal se-

*Corresponding Author (xionghui@ust.hk).

mantic space provided by the source-pretrained detector. This closed-loop design limits both transferability and discriminability under large domain shifts, as the internal representations often lack sufficient semantic richness and precise decision boundaries.

To address this, we turn to external knowledge sources, particularly Vision Foundation Models (VFMs). These large-scale models, pretrained on massive and diverse data, offer powerful generalization and strong visual perception. Yet, their potential remains underexplored in the SFOD setting. VFMs, including visual encoders such as DINOv2 (Oquab et al. 2023) and vision-language detectors such as Grounding DINO (Liu et al. 2024), provide rich, transferable features and robust semantic priors, making them well-suited to augment both feature representation and label quality under source-free constraints. Some recent methods have begun exploring the use of VFMs in domain adaptation. For example, DINO Teacher (DT) (Lavoie, Mahmoud, and Waslander 2025) aligns detector and VFM features using source data, while CODA (Li et al. 2024b) leverages external detections mainly to improve label discriminability. Thus, how to fully and effectively leverage VFMs for improving SFOD in terms of discriminability and transferability remains an open and underexplored problem.

To tackle this challenge, we propose a novel VFM-enhanced SFOD framework, as illustrated in Fig. 2. Unlike previous SFOD approaches that suffer from low-quality pseudo-labels, weak transferability, and source-induced knowledge bias, our method incorporates VFMs through three key modules to address these limitations. To enhance global feature-level transferability, we propose **Patch-weighted Global Feature Alignment (PGFA)**. This module improves feature transferability by aligning the student's feature space with that of a VFM such as DINOv2 (Oquab et al. 2023), guided by patch-wise similarity weights. It encourages the student to learn from VFM's rich, domain-agnostic representations. Considering the importance of instance features in object detection, we propose **Prototype-based Instance Feature Alignment (PIFA)**. This component constructs momentum-based prototypes from VFM features and performs contrastive learning at the instance level, enhancing fine-grained feature alignment and improving instance-level feature transferability and discriminability. To further improve pseudo-label discriminability in the self-training pipeline, we propose **Dual-source Enhanced Pseudo-label Fusion (DEPF)**. In DEPF, we fuse predictions from object detection VFMs such as Grounding DINO (Liu et al. 2024) and the teacher model using instance-level uncertainty-weighted box fusion. This produces cleaner, more reliable supervision signals without relying solely on the biased teacher.

As shown in Fig. 1, the proposed framework bridges the gap between conventional SFOD and the rich representation space of VFMs. By integrating pseudo-label enhancement, global feature alignment, and instance-level alignment, our method significantly boosts both discriminability and transferability. The resulting detector benefits from the broad generalization capability of VFMs and the structure-aware learning of the student-teacher paradigm.

Our main contributions are summarized as follows:

- We identify a key limitation in current SFOD methods: the underutilization of Vision Foundation Models. To this end, we propose a unified framework that integrates VFMs to address both feature transferability and label discriminability.
- We design three lightweight but effective VFM-based components: **PGFA**, **PIFA**, and **DEPF**, which enhance global and instance-level alignment and improve pseudo-label quality through multi-source fusion, synergistically improve both discriminability and transferability.
- We conduct extensive experiments on multiple cross-domain object detection benchmarks. Results demonstrate that our method achieves state-of-the-art SFOD performance, validating the effectiveness of leveraging VFMs for cross-domain adaptation.

## 2 Related Work

### 2.1 Domain Adaptive Object Detection

Domain Adaptive Object Detection (DAOD) aims to transfer detection knowledge from labeled source domains to unlabeled target domains. Prior methods leverage self-training with teacher-student models (Chen et al. 2022; Kennerley et al. 2024; Yao et al. 2021), fine-grained alignment via graph matching (Li, Liu, and Yuan 2022; Li et al. 2023b), and data augmentation or image translation for improved generalization (Liu et al. 2021). More recently, Vision-Language or Foundation Models (VLMs/VFMs) have been introduced to enhance semantic robustness (Li et al. 2023a, 2024a; Lavoie, Mahmoud, and Waslander 2025). However, most DAOD methods require access to source data, which conflicts with privacy constraints. In contrast, our work operates in a source-free setting, without using any labeled source images.

### 2.2 Source-Free Object Detection

SFOD (Source-Free Object Detection) aims to transfer a source pretrained model to unsupervised target domain without access to any source images. Previous SFOD approaches perform efficient knowledge transfer by techniques such as feature alignment (Li et al. 2022a; VS, Oza, and Patel 2023; Yao et al. 2025b) or refined pseudo-labelling (Chen, Wang, and Zhang 2023; Zhang, Zhang, and Liu 2023; Zhao et al. 2024). However, these methods operate on internal knowledge and limited decision boundary of source pretrained model, making the target performance rather limited. VG-DETR (Han, Wang, and Chen 2025) similarly employs VFMs for semi-supervised SFOD, but uses feature clustering for pseudo-labeling, leading to lower efficiency and reliability. In contrast, our method considers the rapid development of VFMs and leverages the rich semantic and broad decision boundary from VFMs pretrained by large-scale data.

### 2.3 Vision Foundation Models

VFMs are large pre-trained models that provide a versatile foundation for vision tasks, offering broad generalization and easy adaptation via prompting or light tuning (Awais

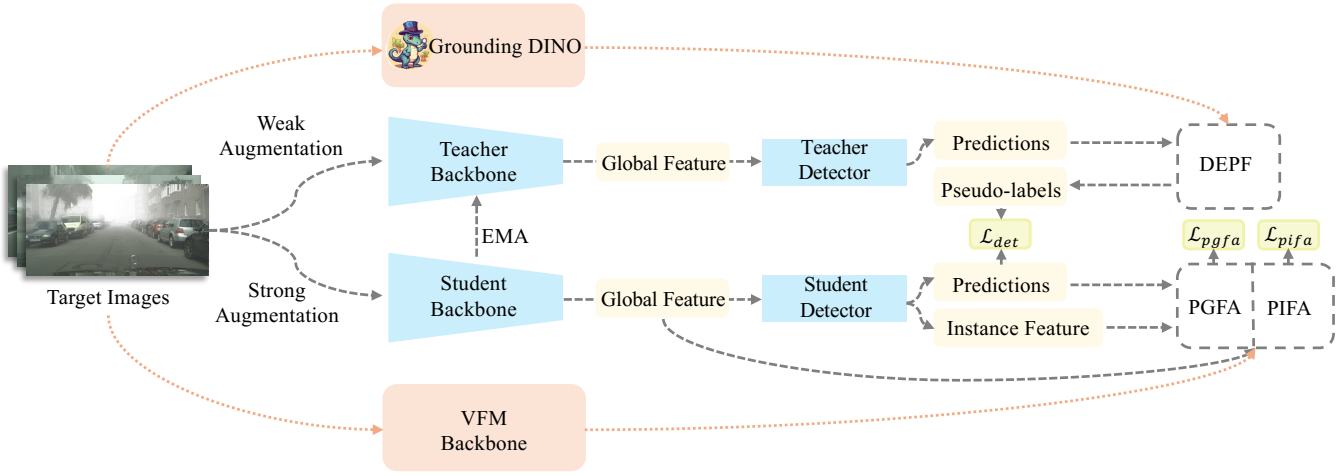


Figure 2: Overview of the proposed method. Unlabeled target images are fed into teacher and student models for self-training with detection loss. DEPF fuses teacher and Grounding DINO predictions to generate refined pseudo-labels, enhancing discriminability. PGFA and PIFA align multi-scale features from the student and VFM, leveraging the generic VFM space to improve transferability.

et al. 2025). Recent VFMs have shown promising performance on various vision tasks such as Segment Anything Model for segmentation (Kirillov et al. 2023), Grounding DINO for object detection (Liu et al. 2024), Depth Anything (Yang et al. 2024) and DINOv2 (Oquab et al. 2023) for depth estimation. Meanwhile, VLMs and MLLMs further extend this foundation to joint multimodal understanding (Lu et al. 2025a; Wu et al. 2025; Wang et al. 2025). In this work, we leverage the rich semantic prior and strong visual perception of VFMs to effectively enhance SFOD discriminability and transferability.

# 3 Method

## 3.1 Preliminaries

**Problem Setup.** In SFOD approaches, the main goal is to transfer a detection model $\theta_S$ from source domain $D_S$ and target domain $D_T$ with unsupervised target dataset $X_T = \{x_T^i\}_{i=1}^{N_T}$. The source model is pretrained from source dataset $X_S = \{x_S^i, y_S^i\}_{i=1}^{N_S}$ where each $(x_S^i, y_S^i)$ is *i.i.d.* sampled from the source domain distribution $D_S$. Similarly, each $x_T^i$ is *i.i.d.* sampled from $D_T$ but the ground-truth label is unknown. Finally, the transferred detection model $\theta_T : x_T \rightarrow y_p$ is of good cross-domain generalization and perform well on target domain.

**Mean Teacher Self-Training Pipeline.** Following previous approaches (Li et al. 2022a; VS, Oza, and Patel 2023), we adopt Mean Teacher (Tarvainen and Valpola 2017) framework as the baseline. Firstly, a teacher model $\theta_{\text{te}}$ and a student model $\theta_{\text{st}}$ is initialized by the same parameter of the source pretrained model. In the adaptation stage, a batch of target samples is weakly and strongly augmented, and input to $\theta_{\text{te}}$ and $\theta_{\text{st}}$ respectively. $\theta_{\text{te}}$ generates pseudo-labels as supervision signals for $\theta_{\text{st}}$, and $\theta_{\text{st}}$ is updated with loss backpropagation. $\theta_{\text{te}}$ does not update from backpropagation but an exponential moving average (EMA) from $\theta_{\text{st}}$ parameters. This progress can be denoted as:

$$\theta_{\text{st}} \leftarrow \theta_{\text{st}} + \eta \frac{\partial \mathcal{L}_{\text{tot}}}{\partial \theta_{\text{st}}}, \quad \theta_{\text{te}} \leftarrow \alpha \theta_{\text{te}} + (1-\alpha)\theta_{\text{st}}. \tag{1}$$

where $\mathcal{L}_{\text{tot}}$ denotes the total loss, $\eta$ denotes learning rate, and $\alpha$ denotes EMA coefficient.

## 3.2 Dual-source Enhanced Pseudo-label Fusion

Starting from the mean teacher baseline, a central challenge in SFOD lies in generating reliable pseudo-labels, which directly affect the model's discriminability. Existing approaches attempt to refine low-confidence predictions or leverage soft-label learning (Yoon et al. 2024; Chen, Wang, and Zhang 2023), but they inherently rely on the biased knowledge of the source model, often leading to overconfident or inaccurate labels. DT (Lavoie, Mahmoud, and Waslander 2025) circumvents this by training a DINOv2-based labeler on source data to generate pseudo-labels for the target domain. However, this violates the core source-free constraint of SFOD and proves ineffective when adapted to noisy target data. To address this, we explore the use of powerful object detection VFMs such as Grounding DINO (Liu et al. 2024), which can act as zero-shot labelers in the target domain. However, while VFMs offer strong generalization, they often underperform under domain shifts (Zhang et al. 2024). Motivated by this, we propose **Dual-source Enhanced Pseudo-label Fusion (DEPF)**, which fuses the generalizable predictions from VFMs with the domain-adapted cues from the teacher model. This is achieved via an entropy-guided fusion strategy that dynamically balances contributions from both sources to produce more reliable pseudo-labels.

To fuse multi-source bounding boxes, Weighted Box Fusion (WBF) (Solovyev, Wang, and Gabruseva 2021) is a

Algorithm 1: DEPF

**Require:** Predictions from two models: $\mathcal{B}_1, \mathcal{B}_2$; IoU threshold $\beta$
**Ensure:** Fused predictions $\mathcal{P}$

Merge all predictions: $\mathcal{B} \leftarrow \mathcal{B}_1 \cup \mathcal{B}_2$
Group $\mathcal{B}$ into clusters using box IoU $> \beta$
**for all** cluster $\mathcal{C}_m = \{(b_k, p_k)\}_{k=1}^{n}$ **do**
  Compute entropy-based weights $\tilde{\mathbf{w}}_k \propto 1/H(p_k)$
  Fuse box: $\hat{b} = \sum_k \tilde{\mathbf{w}}_k b_k$; score: $\hat{p} = \sum_k \tilde{\mathbf{w}}_k p_k$
  Final label: $\hat{y} = \arg\max_c \hat{p}^{(c)}$
  Append $(\hat{b}, \hat{y})$ to $\mathcal{P}$
**end for**
**return** $\mathcal{P}$

straightforward approach that averages overlapping boxes based on confidence scores. However, WBF introduces two critical issues: (1) it requires tuning a global confidence weight, and (2) it merges boxes by class label. This can lead to errors when different sources assign conflicting labels to the same object, which is a frequent scenario due to differing domain perspectives.

To overcome these limitations, we discard class labels during box clustering and instead group overlapping boxes using a large IoU threshold. This allows us to handle label discrepancies more flexibly. For each cluster of overlapping boxes $\{(b_k, p_k)\}_{k=1}^{n}$, where $b_k \in \mathbb{R}^4$ denotes the box coordinates and $p_k \in [0,1]^C$ the class probability vector, we compute the Shannon entropy $H(p_k)$ of each prediction. We then assign inverse-entropy weights, giving higher influence to more certain predictions. The final fused box and class probability are computed via a weighted average using these normalized weights. The resulting pseudo-label set $\mathcal{P} = \{(\hat{b}_i, \hat{y}_i)\}_{i=1}^{n_p}$ is obtained by selecting the class with the highest fused probability for each box. The fusion procedure is outlined in Algorithm 1, while a more detailed version is provided in the extended version (Yao et al. 2025a). Through this entropy-aware design, our framework integrates predictions from both the teacher and VFM, yielding pseudo-labels that are more robust to label noise and domain shifts, leading to improved discriminability.

### 3.3 Patch-weighted Global Feature Alignment

While DEPF mainly improves discriminability by enhancing pseudo-label quality, we now focus on boosting transferability through feature alignment. Prior SFOD works align features between teacher and student networks using graph-based matching (Li et al. 2022a) or adversarial learning (Chu et al. 2023), but these strategies still operate solely within the teacher–student paradigm and thus inherit source-induced biases and limited decision boundaries. Vision Foundation Models (VFMs), particularly DINOv2, provide a broader and more transferable feature space due to large-scale pretraining. DT (Lavoie, Mahmoud, and Waslander 2025) explores aligning a student backbone with a frozen DINOv2 via patch-level cosine similarity, but assumes all patches contribute equally, ignoring the varying semantic importance and domain invariance across regions.

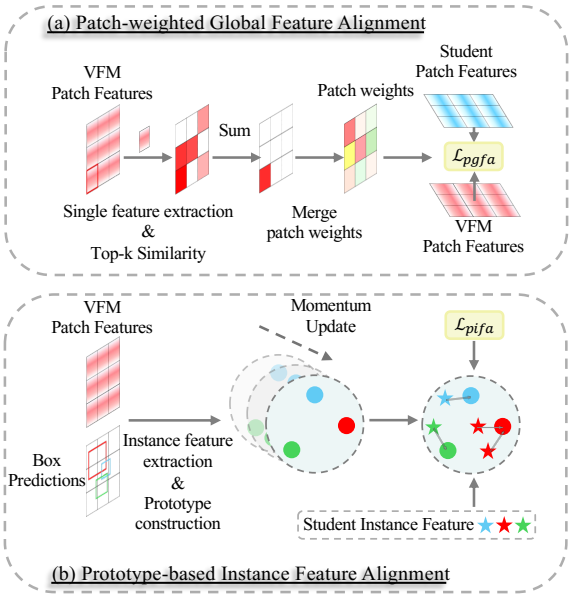


Figure 3: PIFA and PGFA. PGFA adopts similarity-based patch weights for fine-grained global feature fusion. PIFA extracts instance feature from VFM to construct a momentum-updated prototype for contrastive alignment with student instance feature.

To address this, we propose a **Patch-weighted Global Feature Alignment (PGFA)** module that introduces an adaptive patch-wise weighting scheme. We assign a weight to each patch based on its semantic coherence with other patches in the same image. The calculation proceeds in three steps: First, for each image in a minibatch, we compute the pairwise cosine similarity between all its $L_2$-normalized DINOv2 patch features ($\hat{\mathbf{F}}_b^{\mathrm{D}} = \mathbf{F}_b^{\mathrm{D}} / \|\mathbf{F}_b^{\mathrm{D}}\|_2$, where D stands for DINOv2 and $b$ stands for batch index). This results in a similarity matrix $\mathbf{S}_b \in \mathbb{R}^{N \times N}$, where $s_{b,i,j} = \langle \hat{\mathbf{F}}_{b,i}^{\mathrm{D}}, \hat{\mathbf{F}}_{b,j}^{\mathrm{D}} \rangle$ and $N \times N$ is the number of patches. Second, we apply a temperature-controlled softmax function row-wise to this matrix to obtain a probability distribution for each patch's similarity to all others. This yields a new matrix $\mathbf{P}_b \in \mathbb{R}^{N \times N}$:

$$\mathbf{P}_{b,i,j} = \frac{\exp(s_{b,i,j}/\tau)}{\sum_{l=1}^{N} \exp(s_{b,i,l}/\tau)}, \tag{2}$$

where $\tau = 0.07$ is a temperature parameter. Finally, the unnormalized weight $\mathbf{w}_{b,i}$ for patch $i$ is calculated by summing the probabilities of its top-$k$ most similar patches. Let $\mathcal{T}_k(i)$ be the set of indices for the top-$k$ values in the $i$-th row of $\mathbf{P}_b$. The weight $\mathbf{w}_{b,i} = \sum_{j \in \mathcal{T}_k(i)} p_{b,ij}$ is then normalized across all patches for each image:

$$\tilde{\mathbf{w}}_{b,i} = \frac{\mathbf{w}_{b,i}}{\sum_{j=1}^{N} \mathbf{w}_{b,j} + \varepsilon}, \quad i = 1, \ldots, N, \tag{3}$$

where $N$ is the number of patches and $\varepsilon$ ensures numerical stability. This process assigns higher weights to salient, semantically consistent patches, while down-weighting noisy

| Method | Source-Free | Detector | Truck | Car | Rider | Person | Train | Motor | Bicycle | Bus | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | - | DETR | 15.1 | 46.5 | 39.3 | 38.9 | 4.0 | 21.8 | 36.8 | 34.2 | 29.6 |
| CAT (Kennerley et al. 2024) | ✗ | Faster R-CNN | 40.8 | 63.7 | 57.1 | 44.6 | 49.7 | 44.9 | 53.0 | 66.0 | 52.5 |
| DT (Lavoie, Mahmoud, and Waslander 2025) | ✗ | | 47.2 | 65.4 | 60.0 | 48.5 | 52.9 | 46.2 | 56.7 | 66.5 | 55.4 |
| MTTrans (Yu et al. 2022) | ✗ | DETR | 25.8 | 65.2 | 49.9 | 47.7 | 33.9 | 32.6 | 46.5 | 45.9 | 43.4 |
| BiADT (He et al. 2023) | ✗ | | 31.7 | 69.2 | 58.9 | 52.2 | 45.1 | 42.6 | 51.3 | 55.0 | 50.8 |
| ACCT (Zeng, Ding, and Lu 2024) | ✗ | | 31.1 | 69.4 | 58.9 | 53.6 | 33.7 | 42.6 | 54.4 | 53.5 | 49.6 |
| IRG-SFDA (VS, Oza, and Patel 2023) | ✓ | Faster R-CNN | 24.4 | 51.9 | 45.2 | 37.4 | 25.2 | 31.5 | 41.6 | 39.6 | 37.1 |
| AASFOD (Chu et al. 2023) | ✓ | | 28.1 | 44.6 | 44.1 | 32.3 | 29.0 | 31.8 | 38.9 | 34.3 | 35.4 |
| BT (Deng, Li, and Duan 2024) | ✓ | | 24.3 | 52.7 | 47.1 | 38.4 | 36.3 | 30.2 | 40.1 | 44.6 | 39.2 |
| Simple-SFOD† (Hao, Forest, and Fink 2024) | ✓ | | 30.2 | 54.8 | 44.6 | 36.5 | 31.8 | 29.5 | 41.0 | 45.3 | 39.2 |
| LPLD (Yoon et al. 2024) | ✓ | | 29.6 | 56.6 | 49.1 | 39.7 | 26.4 | 36.1 | 43.6 | 46.3 | 40.9 |
| DRU (Khanh et al. 2024) | ✓ | DETR | 26.2 | 62.5 | **51.5** | 48.3 | 34.1 | 34.2 | **48.6** | 43.2 | 43.6 |
| **Ours** | ✓ | | **36.6** | **62.8** | 51.0 | **48.8** | **39.9** | **36.4** | 48.2 | **52.7** | **47.1** |
| Oracle | - | DETR | 31.3 | 71.8 | 52.9 | 52.9 | 41.0 | 41.4 | 44.0 | 53.9 | 48.7 |

† We report results without BN in the backbone following all other approaches for fair comparison.

Table 1: Results of cross-weather adaptation (Cityscapes to Foggy Cityscapes).

or domain-specific regions. For alignment, following DT, we adopt a weighted cosine loss between $L_2$-normalized patch features from DINOv2 ($\mathbf{F}^{\mathrm{D}}$) and the student model ($\mathbf{F}^{\mathrm{S}}$):

$$\mathcal{L}_{\mathrm{pgfa}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N}\tilde{\mathbf{w}}_{b,i}\big(1-\cos(\mathbf{F}^{\mathrm{D}}_{b,i},\mathbf{F}^{\mathrm{S}}_{b,i})\big), \quad (4)$$

where $B$ is the batch size. Note that in our implementation, we first input student model patch features into a lightweight alignment module *e.g.* MLP. This loss encourages the student to align more strongly with transferable regions in the VFM feature space, improving global feature transferability in a fine-grained manner.

### 3.4 Prototype-based Instance Feature Alignment

While PGFA enhances transferability by aligning global features, accurate object detection further relies on instance-level representations. Given the rich semantic priors in VFMs, we propose leveraging them also as semantic anchors to guide instance recognition and complement global alignment. Specifically, we design **Prototype-based Instance Feature Alignment (PIFA)**, which constructs class-wise prototypes from VFM features and applies prototype-based contrastive learning to enforce semantic consistency, facilitating more robust instance-level alignment.

**Prototype update.** We first extract a global feature map $\mathbf{F}^{\mathrm{D}} \in \mathbb{R}^{C\times H\times W}$ from the VFM backbone. Given pseudo-labeled boxes $(b_{k'}, y_{k'})_{k'=1}^{n'}$, we apply RoIAlign (He et al. 2017) to obtain instance features:

$$\mathbf{f}_c^t = \frac{1}{N_c}\sum_{y_{k'}=c}\mathrm{RoIAlign}(\mathbf{F}^{\mathrm{D}}, b_{k'}), \quad (5)$$

where $\mathbf{f}_c^t$ is the mean feature for class $c$ in the current batch.

Inspired by momentum prototype (Li, Xiong, and Hoi 2020), to maintain stable and evolving class-level semantics, we update the prototype $\mathbf{p}_c^t$ using EMA:

$$\mathbf{p}_c^t = \mu\mathbf{p}_c^{t-1} + (1-\mu)\mathbf{f}_c^t, \quad (6)$$

where $\mu = 0.999$ following previous approaches (Li, Xiong, and Hoi 2020). This momentum update smooths temporal fluctuations in the class-wise feature representation while gradually incorporating new semantic information.

**Contrastive learning.** To align the feature spaces of the VFM and the student model's instance features, we first pass the patch-level features from the student model through a lightweight alignment module (e.g., an MLP). After projection, both the resulting features $\mathbf{f}_i$ and the prototypes $\mathbf{p}_c$ are $L_2$-normalized to ensure consistent feature magnitudes and facilitate effective similarity computation. Given pseudo label $\hat{y}_i$ obtained in DEPF, the contrastive loss is computed using InfoNCE (Oord, Li, and Vinyals 2018):

$$\mathcal{L}_{\mathrm{pifa}} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp\left(\mathbf{f}_i^\top\mathbf{p}_{\hat{y}_i}/\tau\right)}{\sum_{c=1}^{K}\exp\left(\mathbf{f}_i^\top\mathbf{p}_c/\tau\right)}, \quad (7)$$

where $\tau$ is a temperature hyperparameter and $M$ is the number of non-empty prototypes. This loss encourages each instance-level feature to align closely with its corresponding class prototype while being repelled from other class prototypes. The use of VFM-guided prototypes as anchors ensures that the student model learns instance features that are semantically aligned and more domain-invariant, leading to improved instance-level feature transferability and class-wise discriminability.

### 3.5 Overall Objective

For model optimization, the student is optimized to minimize the weighted summation of the original detection loss $\mathcal{L}_{\mathrm{det}}$ calculated with pseudo labels, and the proposed $\mathcal{L}_{\mathrm{pgfa}}$ and $\mathcal{L}_{\mathrm{pifa}}$ losses:

$$\mathcal{L}_{\mathrm{tot}} = \mathcal{L}_{\mathrm{det}} + \lambda(\mathcal{L}_{\mathrm{pgfa}} + \mathcal{L}_{\mathrm{pifa}}). \quad (8)$$

The teacher model is then updated through an exponential moving average of student parameters as stated in Eq. 1.

| Method | Source-Free | Detector | C2B | | | | | | | | S2R | K2C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Truck | Car | Rider | Person | Bicycle | Motor | Bus | mAP | mAP | mAP |
| Source | - | DETR | 17.5 | 57.0 | 29.4 | 43.7 | 15.6 | 17.7 | 17.6 | 28.3 | 50.8 | 33.9 |
| SFA (Wang et al. 2021) | ✗ | DETR | 19.1 | 57.5 | 27.6 | 40.2 | 19.2 | 15.4 | 23.4 | 28.9 | 52.6 | 46.7 |
| DA-DETR (Zhang et al. 2023) | ✗ | DETR | - | - | - | - | - | - | - | - | 54.7 | 48.9 |
| IRG-SFDA (VS, Oza, and Patel 2023) | ✓ | Faster R-CNN | - | - | - | - | - | - | - | - | 46.9 | 45.2 |
| AASFOD (Chu et al. 2023) | ✓ | Faster R-CNN | 26.6 | 50.2 | 36.3 | 33.2 | 22.5 | 28.2 | 24.4 | 31.6 | 44.9 | 44.0 |
| SF-UT (Hao, Forest, and Fink 2024) | ✓ | Faster R-CNN | - | - | - | - | - | - | - | - | 55.4 | 46.2 |
| LPLD (Yoon et al. 2024) | ✓ | Faster R-CNN | - | - | - | - | - | - | - | - | 49.4 | 51.3 |
| BT (Deng, Li, and Duan 2024) | ✓ | Faster R-CNN | 24.2 | 50.4 | 34.6 | 32.7 | 28.5 | 24.7 | 24.9 | 31.4 | 48.6 | 48.7 |
| DRU (Khanh et al. 2024) | ✓ | DETR | 27.1 | 62.7 | 36.9 | 45.8 | **32.5** | 22.7 | 28.1 | 36.6 | 58.7 | 45.1 |
| **Ours** | ✓ | DETR | **33.2** | **72.3** | **44.2** | **54.9** | 29.0 | **32.9** | **34.8** | **43.0** | **67.4** | **54.7** |
| Oracle | - | DETR | 66.9 | 87.9 | 56.4 | 74.9 | 53.8 | 68.3 | 55.0 | 66.2 | 75.9 | 75.9 |

Table 2: Results on cross-scene adaptation (Cityscapes to BDD100K, C2B; KITTI to Cityscapes, K2C) and synthetic-to-real adaptation (Sim10k to Cityscapes, S2R). We report AP@50 for each category on C2B and car AP@50 for S2R and K2C.

# 4 Experiments

## 4.1 Datasets and Settings

We evaluate our method on six widely-used object detection datasets: Cityscapes, Foggy Cityscapes, Sim10k, KITTI, BDD100K, and ACDC. We conclude the settings into three domain adaptation scenarios.

**Cross-weather Adaptation.** Cityscapes (Cordts et al. 2016) contains 2,975 training and 500 validation images of urban scenes. Its synthetic variant, Foggy Cityscapes (Sakaridis, Dai, and Van Gool 2018), overlays fog at varying densities; we follow prior work and adopt the 0.02 fog level. Additionally, ACDC (Sakaridis, Dai, and Van Gool 2021) includes four challenging conditions (snow, rain, night, and fog) to further evaluate robustness under diverse weather.

**Synthetic-to-Real Adaptation.** Sim10k (Johnson-Roberson et al. 2017), generated from the GTA V engine, provides 9,000 training and 1,000 validation images and serves as the synthetic source. We use Cityscapes as the real-world target to assess synthetic to real generalization.

**Cross-scene Adaptation.** We consider two cross-scene transfers: First, following prior work, we transfer from Cityscapes to the daytime split of BDD100K (Yu et al. 2020) (36,728 training and 5,258 validation images). Second, we transfer from KITTI (Geiger, Lenz, and Urtasun 2012) (7,481 labeled images) to Cityscapes. These two settings test the model's ability to adapt across cities, camera perspectives, and environmental variations.

## 4.2 Implementation Details

We set $\mu = 0.9$ in Eq. 6 and $\lambda = 1$ in Eq. 8. $\alpha$ in Eq. 1 is set to $0.999$ with EMA applied every 5 iterations. We train for 30 epochs with a learning rate of $5 \times 10^{-5}$ and batch size of 8. See extended version (Yao et al. 2025a) for more details.

## 4.3 Comparisons with State-of-the-art

We evaluate our method under various SFOD scenarios using Deformable DETR as the base detector. Results are shown in Tab. 1, Tab. 2, and Tab. 4. In the cross-weather setting, our approach achieves 47.1% mAP, surpassing DRU (Khanh et al. 2024) (43.6%) and LPLD (Yoon et al. 2024) (40.9%), with especially large gains in challenging classes such as `Truck` (+10.4%) and `Bus` (+9.5%). On the Cityscapes-to-ACDC benchmark, our method consistently outperforms all baselines across all weather conditions. For cross-scene adaptation, we achieve a 6.4% mAP gain over DRU on Cityscapes-to-BDD100K, effectively leveraging the large-scale unlabeled target data. Our performance is also competitive with the non-source-free DT (Lavoie, Mahmoud, and Waslander 2025), trailing by only 4.8% despite not using source data. On Cityscapes-to-KITTI, we outperform all prior SFOD and DAOD methods by a notable margin. In the Sim10k-to-Cityscapes (synthetic-to-real) setting, our method improves car AP by 8.5% over DRU and 12.7% over DA-DETR, demonstrating strong domain adaptation capability across diverse settings.

| Mean Teacher | PGFA | PIFA | DEPF | mAP / Gain |
|---|---|---|---|---|
| ✓ | | | | 42.3 / - |
| ✓ | ✓ | | | 43.4 / +1.1 |
| ✓ | | ✓ | | 43.9 / +1.6 |
| ✓ | ✓ | ✓ | | 45.0 / +2.7 |
| ✓ | | | ✓ | 45.9 / +3.6 |
| ✓ | ✓ | | ✓ | 46.3 / +4.0 |
| ✓ | | ✓ | ✓ | 46.5 / +4.2 |
| ✓ | ✓ | ✓ | ✓(w/o $\tilde{\mathbf{w}}_k$) | 46.8 / +4.5 |
| ✓ | ✓(w/o $\tilde{\mathbf{w}}_b$) | ✓ | ✓ | 46.5 / +4.2 |
| ✓ | ✓ | ✓ | ✓ | **47.1** / +**4.8** |

Table 3: Component effectiveness. Mean Teacher baseline is reproduced with our hyperparameters for fair comparison.

## 4.4 Ablation Study

We conduct ablation experiments mainly on the Cityscapes-to-Foggy Cityscapes setting unless otherwise specified.

**Component Effectiveness.** We evaluate the effectiveness of each proposed module. As shown in Tab. 3, each compo-

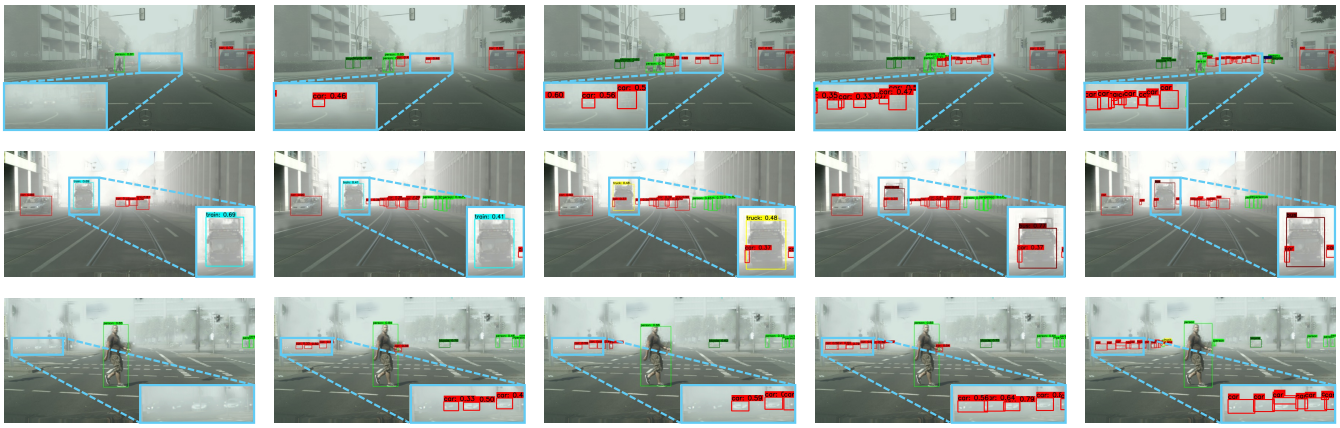

Figure 4: Detection visualization on the Cityscapes-to-Foggy Cityscapes adaptation scenario. Each column from left to right corresponds to: **Source only**, **Mean Teacher**, **DRU**, **Ours**, and **Ground Truth**. We zoom in on the discriminative regions.

| Method | Source-Free | Detector | Snow | Rain | Night | Fog |
|---|---|---|---|---|---|---|
| AT | ✗ | Faster R-CNN | 55.2 | 37.7 | 29.5 | 62.2 |
| DT | ✗ | | 56.8 | 39.0 | 36.4 | 68.6 |
| DRU | ✓ | DETR | 37.9 | 26.3 | 16.5 | 45.4 |
| **Ours** | ✓ | | **47.9** | **32.1** | **23.0** | **54.0** |

Table 4: Results of cross-weather adaptation (Cityscapes-to-ACDC). AT (Li et al. 2022b) and DT (Lavoie, Mahmoud, and Waslander 2025) are non-source-free methods only for reference. We report mAP@50 for each weather scenario.

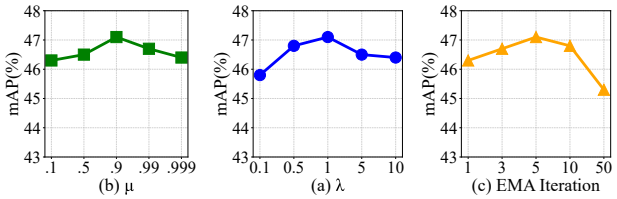


Figure 5: Hyperparameter sensitivity. $\mu$ is the momentum for prototype update, $\lambda$ is the loss balancing factor, EMA iteration is the number of iterations for every EMA update.

nent brings noticeable gains over the Mean Teacher (MT) baseline, and their combination leads to 47.1% mAP, a 4.8% improvement over MT. This confirms their complementary roles in improving transferability and discriminability. Additionally, we assess the effect of the proposed weights: patch-wise distillation weight $\tilde{w}_b$ in PGFA and label fusion weight $\tilde{w}_k$ in DEPF. Including these weights yields further gains of 0.6% and 0.3% mAP, respectively, highlighting the importance of emphasizing informative regions and predictions.

**Hyperparameter Sensitivity.** We evaluate three key hyperparameters: prototype momentum $\mu$ (Eq. 6), loss balancing factor $\lambda$ (Eq. 8), and the EMA update interval, *i.e.*, the number of iterations between teacher updates. As shown in Fig. 5, our method remains robust across a broad range of settings. However, updating prototypes too frequently or too rarely degrades performance. An EMA interval of 5 performs best, balancing stability and timely adaptation, consistent with prior studies (Chu et al. 2023; Zhao et al. 2024).

### 4.5 Visualization

To qualitatively assess the effectiveness of our approach, we conduct visualization experiments on detection results. As shown in Fig. 4, our method demonstrates clear advantages over the baselines. It successfully detects challenging objects and produces correct classifications for visually ambiguous instances (line 2, where *bus* is misclassified as *train* by MT and as *truck* by DRU), highlighting the benefits of our method and the improved optimization enabled by reliable pseudo-labels. We also present pseudo-label visualizations in the extended version (Yao et al. 2025a).

### 4.6 Additional Experiments

Additional experiments, such as detector variants, zero-shot Grounding DINO performance, efficiency analysis, etc., are provided in the extended version (Yao et al. 2025a).

## 5 Conclusion

This paper explores an under-investigated direction of enhancing Source-Free Object Detection (SFOD) using Vision Foundation Models (VFMs) to improve both discriminability and transferability in self-training pipelines. We introduce a simple yet effective framework consisting of three modules: PGFA, PIFA, and DEPF, each focusing on global feature alignment, instance-level feature alignment, and pseudo-label refinement, respectively. By effectively integrating generic knowledge from VFMs, our framework achieves consistent improvements in SFOD performance, as validated by extensive experiments. These results indicate that integrating VFMs into more scalable and robust DAOD and SFOD frameworks is a promising direction for future research. In future work, we aim to explore more about richer vision-language multimodal feature spaces to enhance alignment effectiveness and to incorporate VFMs into more challenging cross-domain object detection scenarios.

## Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities(2024ZDZX1007), in part by the Education Bureau of Guangzhou.

## References

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, M.; Chen, W.; Yang, S.; Song, J.; Wang, X.; Zhang, L.; Yan, Y.; Qi, D.; Zhuang, Y.; Xie, D.; et al. 2022. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*.

Chen, Z.; Wang, Z.; and Zhang, Y. 2023. Exploiting low-confidence pseudo-labels for source-free object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5370–5379.

Chu, Q.; Li, S.; Chen, G.; Li, K.; and Li, X. 2023. Adversarial alignment for source free object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 452–460.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Deng, J.; Li, W.; and Duan, L. 2024. Balanced teacher for source-free object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7231–7243.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Han, J.; Wang, Y.; and Chen, L. 2025. VFM-Guided Semi-Supervised Detection Transformer under Source-Free Constraints for Remote Sensing Object Detection. *arXiv preprint arXiv:2508.11167*.

Hao, Y.; Forest, F.; and Fink, O. 2024. Simplifying source-free domain adaptation for object detection: Effective self-training strategies and performance insights. In *European Conference on Computer Vision*, 196–213. Springer.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, L.; Wang, W.; Chen, A.; Sun, M.; Kuo, C.-H.; and Todorovic, S. 2023. Bidirectional alignment for domain adaptive detection with transformers. In *IEEE/CVF International Conference on Computer Vision*, 18775–18785.

Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S. N.; Rosaen, K.; and Vasudevan, R. 2017. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *IEEE International Conference on Robotics and Automation*, 746–753.

Kennerley, M.; Wang, J.-G.; Veeravalli, B.; and Tan, R. T. 2024. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16541–16550.

Khanh, T. L. B.; Nguyen, H.-H.; Pham, L. H.; Tran, D. N.-N.; and Jeon, J. W. 2024. Dynamic retraining-updating mean teacher for source-free object detection. In *European Conference on Computer Vision*, 328–344. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Kundu, J. N.; Kulkarni, A. R.; Bhambri, S.; Mehta, D.; Kulkarni, S. A.; Jampani, V.; and Radhakrishnan, V. B. 2022. Balancing discriminability and transferability for source-free domain adaptation. In *International conference on machine learning*, 11710–11728. PMLR.

Lavoie, M.-A.; Mahmoud, A.; and Waslander, S. L. 2025. Large Self-Supervised Models Bridge the Gap in Domain Adaptive Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4692–4702.

Li, H.; Zhang, R.; Yao, H.; Song, X.; Hao, Y.; Zhao, Y.; Li, L.; and Chen, Y. 2023a. Learning domain-aware detection head with prompt tuning. *Advances in Neural Information Processing Systems*, 36: 4248–4262.

Li, J.; Xiong, C.; and Hoi, S. C. 2020. Mopro: Webly supervised learning with momentum prototypes. *arXiv preprint arXiv:2009.07995*.

Li, J.; Yu, Z.; Du, Z.; Zhu, L.; and Shen, H. T. 2024a. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5743–5762.

Li, P.; He, Y.; Yu, F. R.; Song, P.; Yin, D.; and Zhou, G. 2023b. IGG: Improved graph generation for domain adaptive object detection. In *Proceedings of the 31st ACM international conference on multimedia*, 1314–1324.

Li, S.; Ye, M.; Zhou, L.; Li, N.; Xiao, S.; Tang, S.; and Zhu, X. 2024b. Cloud object detector adaptation by integrating different source knowledge. *Advances in Neural Information Processing Systems*, 37: 25251–25283.

Li, S.; Ye, M.; Zhu, X.; Zhou, L.; and Xiong, L. 2022a. Source-free object detection by learning to overlook domain style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8014–8023.

Li, W.; Liu, X.; and Yuan, Y. 2022. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5291–5300.

Li, Y.-J.; Dai, X.; Ma, C.-Y.; Liu, Y.-C.; Chen, K.; Wu, B.; He, Z.; Kitani, K.; and Vajda, P. 2022b. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7581–7590.

Liu, F.; Zhang, X.; Wan, F.; Ji, X.; and Ye, Q. 2021. Domain contrast for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8227–8237.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Lu, S.; Chen, Y.; Feng, W.; Fan, J.; Li, F.; Zhang, Z.; Lv, J.; Shen, J.; Law, C.; and Liang, J. 2025a. Uni-layout: Integrating human feedback in unified layout generation and evaluation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7709–7718.

Lu, S.; Wang, Y.; Sheng, L.; He, L.; Zheng, A.; and Liang, J. 2025b. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58(2): 1–39.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126: 973–992.

Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10765–10775.

Solovyev, R.; Wang, W.; and Gabruseva, T. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

VS, V.; Oza, P.; and Patel, V. M. 2023. Instance relation graph guided source-free domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3520–3530.

Wang, W.; Cao, Y.; Zhang, J.; He, F.; Zha, Z.-J.; Wen, Y.; and Tao, D. 2021. Exploring sequence feature alignment for domain adaptive detection transformers. In *ACM International Conference on Multimedia*, 1730–1738.

Wang, Y.; Huang, T.; He, C.; Li, Q.; and Gao, J. 2025. Simple and Efficient Heterogeneous Temporal Graph Neural Network. *arXiv preprint arXiv:2510.18467*.

Wu, J.; Feng, M.; Zhang, S.; Jin, R.; Che, F.; Wen, Z.; and Tao, J. 2025. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv e-prints*, arXiv–2502.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10371–10381.

Yao, H.; Zhao, S.; Li, P.; Cui, Y.; Lu, S.; Guo, W.; Lu, Y.; Xu, Y.; and Xiong, H. 2025a. Beyond Boundaries: Leveraging Vision Foundation Models for Source-Free Object Detection. *arXiv preprint arXiv:2511.07301*.

Yao, H.; Zhao, S.; Lu, S.; Chen, H.; Li, Y.; Liu, G.; Xing, T.; Yan, C.; Tao, J.; and Ding, G. 2025b. Source-Free Object Detection With Detection Transformer. *IEEE Transactions on Image Processing*, 34: 5948–5963.

Yao, X.; Zhao, S.; Xu, P.; and Yang, J. 2021. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3273–3282.

Yoon, I.; Kwon, H.; Kim, J.; Park, J.; Jang, H.; and Sohn, K. 2024. Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In *European Conference on Computer Vision*, 337–353. Springer.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2633–2642.

Yu, J.; Liu, J.; Wei, X.; Zhou, H.; Nakata, Y.; Gudovskiy, D.; Okuno, T.; Li, J.; Keutzer, K.; and Zhang, S. 2022. MTTrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, 629–645. Springer.

Zeng, Z.; Ding, Y.; and Lu, H. 2024. Enhancing cross-domain detection: adaptive class-aware contrastive transformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6670–6674.

Zhang, H.; Su, Y.; Xu, X.; and Jia, K. 2024. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23385–23395.

Zhang, J.; Huang, J.; Luo, Z.; Zhang, G.; Zhang, X.; and Lu, S. 2023. Da-detr: Domain adaptive detection transformer with information fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23787–23798.

Zhang, S.; Zhang, L.; and Liu, Z. 2023. Refined pseudo labeling for source-free domain adaptive object detection. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Zhao, S.; Hong, X.; Yang, J.; Zhao, Y.; and Ding, G. 2023. Toward Label-Efficient Emotion and Sentiment Analysis. *Proceedings of the IEEE*, 111(10): 1159–1197.

Zhao, S.; Yao, H.; Lin, C.; Gao, Y.; and Ding, G. 2024. Multi-source-free domain adaptive object detection. *International Journal of Computer Vision*, 132(12): 5950–5982.