

Expert-Teacher-Student Collaborative Learning for Domain Adaptive Object Detection

Yiming Cui^{1,2*} Liang Li^{2†} Haibing Yin¹ Yuhao Gao¹ Xichun Sheng³ Chenggang Yan^{1,4}

¹School of Communication Engineering, Hangzhou Dianzi University (HDU)

²Institute of Computing Technology, Chinese Academy of Sciences ³Macao Polytechnic University

⁴Zhejiang Provincial Key Laboratory of Low Altitude Ubiquitous Networking Technology, HDU

{cuiyiming, yhb, yuhangao, cgyan}@hdu.edu.cn, liang.li@ict.ac.cn, p2314922@mpu.edu.mo

Abstract

Domain adaptive object detection (DAOD) aims to generalize an object detector trained on a source domain to a target domain, where the domain gap degrades the adaptability. Recently, large-scale vision foundation models (VFMs), pretrained on web-scale datasets, exhibit such powerful generalization capabilities that many approaches leverage them to bridge the domain gap. However, their generalized knowledge is not tailored to the specific domain, which makes it difficult to offer precise guidance in the target domain. In this paper, we propose an **Expert-Teacher-Student collaborative learning (ETS)** framework to synergize the generalized knowledge from VFMs with the domain-specific knowledge from the teacher model. Concretely, we first design an **Expert-Teacher Collaborative Teaching (ETCT)** module, which leverages the complementary knowledge of expert and teacher models to collaboratively generate high-quality pseudo labels for supervising student model learning. Second, we devise an **Expert-Teacher Joint Consolidating (ETJC)** module, which introduces class-wise prototype alignment among expert, teacher, and student models, to jointly consolidate generalized and domain-specific knowledge within the student model. ETS leverages VFMs as the expert model in a free lunch manner, thus avoiding significant additional training costs. Extensive experiments exhibit that our method outperforms the existing SOTA methods on three benchmarks.

1. Introduction

Object detection plays a crucial role in visual applications such as autonomous driving and public safety. However, detectors trained on carefully curated datasets (*i.e.*, source domain) tend to degenerate when deployed in real-world en-

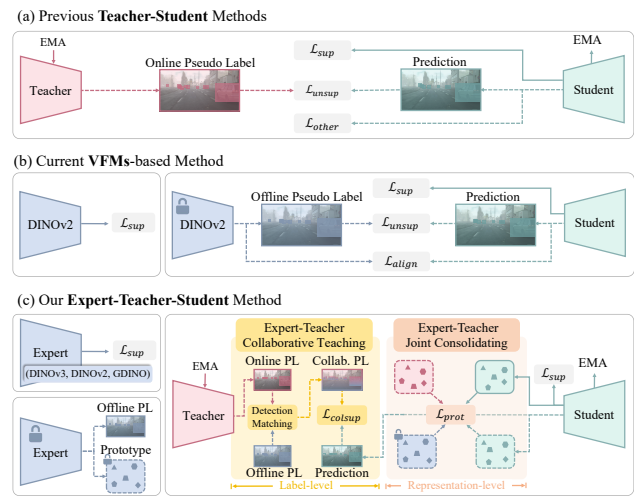


Figure 1. (a) Previous teacher-student DAOD methods utilize pseudo labels generated by the teacher model to supervise the student model. (b) The SOTA VFMs-based method DT [26] employs a frozen DINOv2 backbone for offline pseudo-labeling and online feature alignment. (c) Our ETS first trains an expert model with the source domain, then generates pseudo labels and prototypes for the target domain, and finally, expert and teacher models collaboratively supervise and consolidate the student model at both the label-level and representation-level.

vironments (*i.e.*, target domain). Domain Adaptive Object Detection (DAOD) is proposed to address this issue, which aims to generalize a detector trained on the source domain to accurately detect in the target domain [9].

The dominant paradigm of DAOD employs the teacher-student framework, where the teacher model generates pseudo labels to supervise the student model learning in the target domain (as shown in Fig. 1 (a)). The core challenge lies in generating high-quality pseudo labels to prevent self-training collapse [5]. To address this, one type of method calculates dynamic classification confidence thresholds to refine pseudo labels [14, 59, 67], while another type

*This work is done during the intern in VIPL group, ICT, CAS.

†Corresponding authors.

of method measures the consistency between localization and classification to filter unreliable predictions [7, 8, 16]. Despite this, their closed-loop internal knowledge and the inherent limitations of small-scale backbones leave them vulnerable to the domain gap, resulting in a performance bottleneck for domain adaptation.

To overcome these limitations, recent works leverage the pre-trained large-scale vision foundation models (VFMs, *e.g.*, DINOv2 [38], DINOv3 [45], Grounding DINO [37], *etc.*) to provide external generalized knowledge guidance [1, 20, 31]. Motivated by this, recent VFMs-based DAOD methods, such as DT [26], decouple the teacher-student framework by replacing the teacher model with an offline DINOv2-based labeller, and learning more generalizable features by aligning image representations between the student and the VFMs, as shown in Fig. 1 (b). The generalized knowledge makes VFMs excel at capturing domain-invariant cues (*e.g.*, using outlines to detect the obscured car in Fig. 2 (a)), which can provide universal guidance. However, relying solely on VFMs constrains the student model’s adaptability to the target domain, as their generalized knowledge is not tailored to the specific target domain. In contrast, the teacher model gradually distills target domain knowledge, specializing in domain-specific properties (*e.g.*, enabling the inference of cars covered by fog in Fig. 2 (b)). Both models have complementary knowledge and can jointly guide the student model toward better adaptation to the target domain.

Therefore, we propose an **Expert-Teacher-Student** collaborative learning (ETS) framework, which enables the student model to progressively learn the complementary knowledge from both expert and teacher models. As illustrated in Fig. 1(c), we leverage VFMs trained on the source domain as an expert model in a free lunch manner to generate offline pseudo labels and extract class-wise prototypes. Then, we introduce collaborative learning from dual levels via a teach-then-consolidate strategy: (1) At the label level, we design the Expert-Teacher Collaborative Teaching (ETCT) module to perform detection matching between the expert and teacher models, categorizing their pseudo labels into consistent, inconsistent, and unilateral detections. ETCT employs confidence-based fusion and selection mechanisms to integrate these pseudo labels, thereby generating high-quality collaborative pseudo labels that provide more robust supervision for the student model. (2) At the representation level, we devise the Expert-Teacher Joint Consolidating (ETJC) module, which enables the expert and teacher models to consolidate the student model’s learned representation. Specifically, we construct prototype networks for each model and align the student model’s target domain prototypes with those from expert and teacher models to consolidate the learned generalization and adaptability. Additionally, the student model

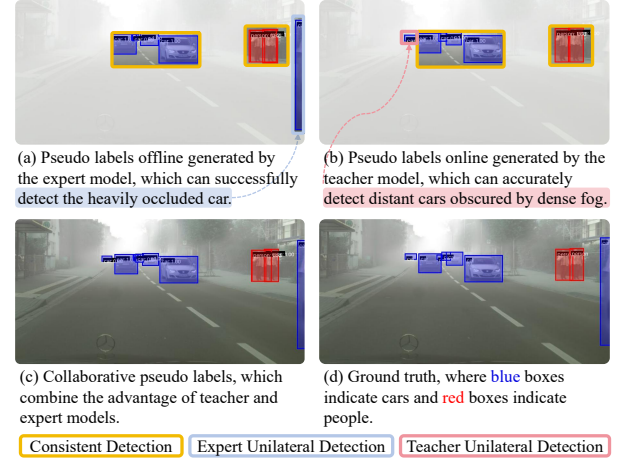


Figure 2. Visualization of detection results. We use different box colors to denote consistent and unilateral detection results. The expert model focuses on domain-invariant cues such as object outlines and semantics, while the teacher model specializes in domain-specific properties. Our framework leverages the complementary advantages of the teacher and expert models to generate high-quality collaborative pseudo-labels, thus better teaching the student model to adapt to the target domain.

aligns its source and target domain prototypes through a self-assessment way to encourage the domain-invariant representation. The experiments demonstrate the robust adaptation capability of our ETS over SOTA methods on three benchmarks.

Our main contributions can be summarized as follows:

- We propose ETS to leverage the expert model’s generalized knowledge and the teacher model’s domain-specific knowledge for guiding student model learning in the target domain.
- We design ETCT and ETJC modules, enabling the student model to inherit advantageous knowledge from both teacher and expert models through a progressive strategy of teach-then-consolidate.
- Comparison and ablation experiments demonstrate that ETS exhibits superior domain adaptation capability, outperforming existing methods on three benchmarks. *e.g.*, it reaches 49.8% mAP on Cityscapes→BDD100k, surpassing the state-of-the-art DT [26] by 2.0%.

2. Related Work

The rapid progress in deep learning has significantly advanced computer vision [10, 11, 17, 19, 30, 48, 52, 53, 58, 60, 63, 69] and multimedia understanding [35, 40, 47, 54, 55, 61, 62, 64, 68], particularly in DAOD [3, 7, 14, 36]. Next, we systematically review representative methods from the following perspectives:

2.1. Teacher-Student Framework for DAOD

The mainstream approaches in DAOD are to introduce the teacher-student framework. To be specific, AT [34] arms the teacher-student framework with weak-strong data augmentation, yielding a noticeable increase in accuracy. MIC [24] and SOCCER [13] introduce the masked image consistency between the teacher model and the student model to enhance the ability of context relations. The core challenge of the teacher-student framework lies in generating high-quality pseudo labels. Although the above approaches have achieved significant progress, the current framework remains constrained by its small-scale architecture and limited internal knowledge, making it vulnerable to domain gaps that result in noisy pseudo labels and ultimately lead to suboptimal adaptation performance.

2.2. Large Models Boost Object Detection

Recent advances in large-scale vision foundation models (VFMs), including vision-based models like DINOv3 [45] and SAM2 [42], and vision-language models like CLIP [41] and Grounding DINO [37], have injected new vitality into the development of object detection tasks. GT [1] employs Grounding DINO to implement an expert foundational branch to supervise the student model. COIN [31] introduces gradient-aligned knowledge distillation from cloud and CLIP models for better knowledge transfer. Furthermore, in DAOD tasks, DA-Pro [27] leverages CLIP with domain-adaptive prompts to construct dynamic detection heads, capturing domain-invariant and domain-specific knowledge. DT [26] decouples the teacher-student framework by replacing the teacher model with a frozen DINOv2 backbone-trained labeller to provide supervision signals. Different from the above works, we aim to inherit the generalization capability of rapidly developing VFMs into a lightweight model, while simultaneously maintaining its adaptability to the specific target domain.

3. Preliminaries

3.1. Problem Formulation

Given a labeled source domain $\mathcal{D}_s = (X_s, B_s, C_s)$ and an unlabeled target domain $\mathcal{D}_t = (X_t)$, where X_s and X_t present N_s source images and N_t target images, $B_s = \{B_i\}_{i=1}^{N_s}$ denotes the bounding box annotations and $C_s = \{C_i\}_{i=1}^{N_s}$ denotes corresponding object class labels. The ultimate goal is to design a detector with $\mathcal{D}_s$ and $\mathcal{D}_t$ that achieves robust adaptation within the unseen target domain.

3.2. Teacher-Student Framework

In the burn-in stage, the teacher model and the student model share the same structure in the self-training framework. The student model acquires the source domain

knowledge from source data by supervised training:

$$\mathcal{L}_{sup}(X_s, B_s, C_s) = \mathcal{L}_{sup}^{cls}(C'_s, C_s) + \mathcal{L}_{sup}^{reg}(B'_s, B_s), \quad (1)$$

where C'_s and B'_s denote classes and bounding boxes in X_s predicted by the student model.

In the mutual-learning stage, by forcing the student model to imitate the deep representations of the teacher model, the student model can learn the target domain knowledge from the pseudo labels $\hat{Y}(\hat{B}, \hat{C})$:

$$\mathcal{L}_{unsup}(X_t, \hat{B}, \hat{C}) = \mathcal{L}_{unsup}^{cls}(C'_t, \hat{C}) + \mathcal{L}_{unsup}^{reg}(B'_t, \hat{B}), \quad (2)$$

where C'_t and B'_t denote classes and bounding boxes in X_t predicted by the student model, $\hat{C}$ and $\hat{B}$ denote the categories and bounding boxes of objects in a pseudo label. In addition, to filter out the noisy pseudo label, we apply a confidence threshold δ to filter low-confidence bounding boxes. The teacher model is updated by Exponential Moving Average (EMA) from the weights of the student model:

$$\theta_t \leftarrow \alpha \theta_t + (1 - \alpha) \theta_s, \quad (3)$$

where θ_t and θ_s denote the model parameters of teacher and student, respectively, and α is an update hyper-parameter.

3.3. VFMs-based Expert Models

We use a frozen DINOv3 ViT-H+ [45] as the backbone of Faster R-CNN, and we train the object detection head (RPN and ROI head) using only the source domain data through Eq. (1). We utilize the above detector as the expert model to generate all pseudo labels $\hat{Y}(\hat{B}, \hat{C})$ in a single forward pass over the target training data. To filter uncertain detections, we apply a confidence threshold δ , the same as the teacher model. Both the pseudo labels generation and the prototype extraction (detailed in Sec. 4.3.1) are conducted offline, thereby incurring no computational cost during the ETS training. We exhibit the expert model's performance on each target domain in Tab. 9. Additionally, we employ DINOv2 [38] and MM Grounding DINO (MM GDINO) [66] as expert models discussed in Sec. 5.4.

4. Methodology

4.1. Overview

The proposed ETS introduces the VFM as the expert model to collaborate with the teacher model for better guiding student model learning. It consists of two modules, *i.e.*, ETCT and ETJC, as illustrated in Fig. 3. Concretely, we first design the ETCT module to exploit the complementary advantages of expert and teacher models for collaboratively generating high-quality pseudo labels to supervise the student model. Then, we devise the ETJC module that utilizes prototype alignment among models to bring the class feature spaces closer, consolidating the student model's generalization and adaptation capabilities.

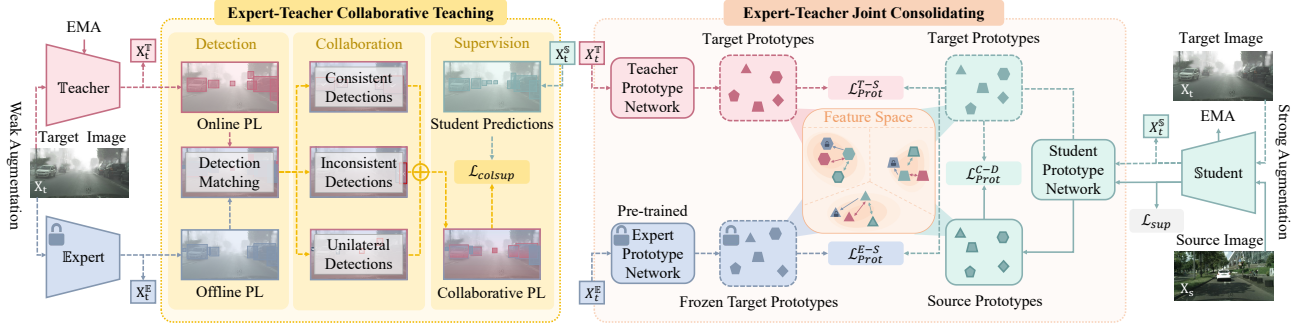


Figure 3. **Overview of the proposed ETS**, which implements dual-level collaboration between expert and teacher models to guide student model learning. 1) Expert-Teacher Collaborative Teaching (ETCT) integrates detections from both expert and teacher models to generate high-quality pseudo labels (PL), collaboratively supervising the student model’s learning. 2) Expert-Teacher Joint Consolidating (ETJC) aligns prototypes across models to consolidate the student model’s learned knowledge.

4.2. Expert-Teacher Collaborative Teaching

To better integrate knowledge from expert and teacher models, inspired by [31], we categorize their pseudo labels via detection matching into three types: consistent, inconsistent, and unilateral detections. For consistent detections, we reweight box coordinates to enhance the pseudo label localization quality. For inconsistent detections, we reduce model disagreement by reassigning classes based on confidence scores. For unilateral detections, we selectively incorporate high-confidence objects, maximizing the retention of generalized and specific knowledge.

4.2.1. Detection Matching.

Given the target image X_t , its offline pseudo label generated by the expert model is $\tilde{Y}(\tilde{B}, \tilde{C})$, and its online pseudo label generated by the teacher model is $\hat{Y}(\hat{B}, \hat{C})$. To find the matched boxes, we define a matching matrix $\mathcal{M}_{i,j}$, where $\mathcal{M}_{i,j} = 1$ if the $IoU \geq \tau$ between the i -th box $\tilde{B}_i$ from the expert and the j -th box $\hat{B}_j$ from the teacher, otherwise $\mathcal{M}_{i,j} = 0$. For the i -th expert-generated box $\tilde{B}_i$ (or the j -th teacher-generated box $\hat{B}_j$), the class confidence score $\tilde{s}_i$ (or $\hat{s}_j$) is obtained from the maximum probability of the predicted class distribution: $\tilde{s}_i = \max(\tilde{q}_i)$, $\hat{s}_j = \max(\hat{q}_j)$, where $\tilde{q}_i \in [0, 1]^K$ and $\hat{q}_j \in [0, 1]^K$ denote the K -dimensional class probability vectors for the respective detections. Then the initial consistent detections set $\mathcal{P}_c^{init}$, initial inconsistent detections set $\mathcal{P}_i^{init}$, and initial unilateral detections set $\mathcal{P}_u^{init}$ are defined as follows:

$$\begin{aligned} \mathcal{P}_c^{init} &= \{(\tilde{Y}_i, \hat{Y}_j) \mid \mathcal{M}_{i,j} = 1, \tilde{C}_i = \hat{C}_j\}, \\ \mathcal{P}_i^{init} &= \{(\tilde{Y}_i, \hat{Y}_j) \mid \mathcal{M}_{i,j} = 1, \tilde{C}_i \neq \hat{C}_j\}, \\ \mathcal{P}_u^{init} &= \{\tilde{Y}_i \mid \mathcal{M}_{i,\cdot} = 0\} \cup \{\hat{Y}_j \mid \mathcal{M}_{\cdot,j} = 0\}. \end{aligned} \quad (4)$$

4.2.2. Consistent Detections.

For one pair of matched boxes $\tilde{B}_i$ and $\hat{B}_j$ (including $\mathcal{P}_c^{init}$ and $\mathcal{P}_i^{init}$), we employ confidence scores to perform weighted box fusion, as the scores are generally propor-

tional to the localization accuracy of the boxes:

$$\bar{B}_n = \frac{\tilde{s}_i \cdot \tilde{B}_i + \hat{s}_j \cdot \hat{B}_j}{\tilde{s}_i + \hat{s}_j}, \quad (5)$$

where $\bar{B}_n$ denotes the n -th fused box. Thus, the final consistent detections set $\mathcal{P}_c$ can be denoted as $\mathcal{P}_c = (\bar{B}, \bar{C})$, where $\bar{C} = \hat{C}$, as the predicted classes are consistent.

4.2.3. Inconsistent Detections.

For inconsistent detections, the final class assignment follows the teacher model when its predicted class confidence exceeds the expert model’s by a confidence discrepancy threshold ϵ , otherwise adopts the expert model’s prediction:

$$\bar{C}_n = \begin{cases} \hat{C}_j & \text{if } \hat{s}_j - \tilde{s}_i \geq \epsilon \\ \tilde{C}_i & \text{otherwise} \end{cases}, \quad (6)$$

where $\bar{C}_n$ is finally reconciled class of n -th box. the final inconsistent detections set $\mathcal{P}_i$ can be denoted as $\mathcal{P}_i = (\bar{B}, \bar{C})$.

4.2.4. Unilateral Detections.

Finally, for the expert model’s unilateral detections, we retain all detections above the base confidence threshold δ . For the teacher model’s unilateral detections, we apply a stricter confidence threshold δ' (where $\delta' > \delta$) to ensure quality. The final unilateral detections set is $\mathcal{P}_u = (\bar{B}, \bar{C}) \cup \{(\hat{B}, \hat{C}) \mid \hat{s} \geq \delta'\}$.

4.2.5. Collaborative Supervising.

Therefore, the final collaborative pseudo labels can be represented as:

$$Y_{col} = \mathcal{P}_c \cup \mathcal{P}_i \cup \mathcal{P}_u. \quad (7)$$

We utilize the collaborative pseudo labels to supervise the student model for better target domain knowledge learning:

$$\mathcal{L}_{col sup} = \mathcal{L}_{unsup}(X_t, B_{col}, C_{col}), \quad (8)$$

where B_{col}, C_{col} denote the set of bounding boxes and classes in Y_{col} .

4.3. Expert-Teacher Joint Consolidating

To further consolidate the knowledge learned by the student, we leverage class-conditioned alignment to minimize prototype distance between the student and both the expert and teacher models in the target feature space. Concurrently, the student model enhances domain-invariant representations by self-aligning prototypes across source and target domains.

4.3.1. Prototype Formulation

We attach a lightweight prototype network to the ROI features of each model to obtain class-level prototypes. This network maps the ROI features to d -dimensional feature vectors. We use p_k^{new} to denote the prototype of the k -th class within the current mini-batch, which is a confidence-weighted average of all ROI features predicted as that class:

$$p_k^{new} = \frac{\sum_n s_{k,n} P(R(F(X), B_n))}{\sum_n s_{k,n}}, \quad (9)$$

where P is a 2-layer MLP prototype network, R is the Faster R-CNN ROI head, F is the backbone, B_n is the bounding box, and $s_{k,n}$ is the confidence score of the n -th object for class k . The prototype update follows a confidence-weighted EMA scheme:

$$\begin{aligned} p_k^{t+1} &= \frac{\alpha' m_k^t p_k^t + s_k p_k^{new}}{\alpha' m_k^t + s_k}, \\ m_k^{t+1} &= \alpha' m_k^t + s_k, \end{aligned} \quad (10)$$

where m_k denotes the accumulated confidence for class k , and $\alpha' = 0.999$ is the momentum coefficient. $s_k = \sum_n s_{k,n}$ is the aggregated confidence score for class k within the current batch. This confidence-weighted approach ensures that high-confidence predictions contribute more to the prototype formation, leading to more robust and representative class prototypes.

4.3.2. Expert-Student Prototype Alignment

To preserve the intrinsic generalization representation of expert prototypes, we extract and freeze expert model prototypes from 20,000-th training iterations as the alignment target of the student model. This enables the student model to inherit visual generalization representations from large-scale models. The expert-student alignment loss is computed only for classes whose prototypes have been sufficiently updated:

$$\mathcal{L}_{Prot}^{E-S} = \frac{1}{|V|} \sum_{k \in V} \left(1 - \cos \left(p_k^{exp_{tgt}}, p_k^{stu_{tgt}} \right) \right), \quad (11)$$

where $V = \{k \mid m_k^{exp_{tgt}} > 0 \wedge m_k^{stu_{tgt}} > 0\}$ represents the set of valid classes with updated prototypes in both models.

4.3.3. Teacher-Student Prototype Alignment

The teacher and student models progress together, gradually improving their capabilities to extract class-specific prototypes. By introducing alignment constraints between the teacher and student models, the student can better imitate the deep representations of the teacher model through a prototype-level alignment for robust class representation:

$$\mathcal{L}_{Prot}^{T-S} = \frac{1}{|V|} \sum_{k \in V} \left(1 - \cos \left(p_k^{tea_{tgt}}, p_k^{stu_{tgt}} \right) \right), \quad (12)$$

where $V = \{k \mid m_k^{tea_{tgt}} > 0 \wedge m_k^{stu_{tgt}} > 0\}$ denotes the valid class set for alignment.

4.3.4. Cross-Domain Prototype Alignment

We enhance the capability of the student model to extract domain-invariant features by aligning its source and target domain prototypes:

$$\mathcal{L}_{Prot}^{C-D} = \frac{1}{|V|} \sum_{k \in V} \left(1 - \cos \left(p_k^{stu_{src}}, p_k^{stu_{tgt}} \right) \right), \quad (13)$$

where $V = \{k \mid m_k^{stu_{src}} > 0 \wedge m_k^{stu_{tgt}} > 0\}$ ensures both domains provide valid prototypes for class k . tgt and src denote the target and source domains, respectively.

4.4. Overall Objective Function

The overall objective function is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{sup} + \lambda_2 \mathcal{L}_{colsup} + \lambda_3 \mathcal{L}_{Prot}^{E-S} + \lambda_4 \mathcal{L}_{Prot}^{T-S} + \lambda_5 \mathcal{L}_{Prot}^{C-D}, \quad (14)$$

where $\lambda_{1...5}$ are loss weights associated with each term. We provide the detailed pseudocode of ETS in Appendix G.

5. Experiments

5.1. Datasets

Pascal VOC to Clipart1k: Pascal VOC (P) [18] as the source domain, comprises 16,551 real-world images across 20 classes, for training. Clipart1k (CI) [25] as the target domain consists of 20 corresponding classes. We split Clipart1k into 500 training and 500 validation images.

Cityscapes to BDD100k: BDD100k (B) [56] is a large-scale driving dataset. We extract the daytime subset of BDD100k as the target domain, containing 36,728 training images and 5,258 validation images. Cityscapes with seven shared classes is adopted as the source domain.

Cityscapes to Foggy Cityscapes: Cityscapes (C) [12] is collected from 50 urban scenes of normal weather and contains 2,975 images. Foggy Cityscapes (F) [44] is a synthesized dataset that simulates foggy weather at three intensities (0.005, 0.01, 0.02) as the target domain. It contains 8,925 training images and 1,500 validation images.

Table 1. Results of Pascal VOC to Clipart1k (**P→CI**). The average precision (AP, %) for all classes and the mean AP (mAP) are reported.

Method	Venue	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mtr	prsn	plant	shep	sofa	train	tv	mAP
UMT [15]	CVPR'21	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	67.6	21.4	29.5	48.2	75.9	70.5	56.7	25.9	28.9	39.4	43.6	44.1
SIGMA [32]	CVPR'22	40.1	55.4	37.4	31.1	54.9	54.3	46.6	23.0	44.7	65.6	23.0	22.0	42.8	55.6	67.2	55.2	32.9	40.8	45.0	58.6	44.5
AT [34]	CVPR'22	33.1	66.1	35.3	44.9	57.5	44.9	51.0	5.8	59.5	54.9	34.6	23.5	64.3	84.0	75.4	51.5	17.1	30.3	43.3	37.2	45.7
TIA [65]	CVPR'22	42.2	66.0	36.9	37.3	43.7	71.8	49.7	18.2	44.9	58.9	18.2	29.1	40.7	87.8	67.4	49.7	27.4	27.8	57.1	50.6	46.3
SIGMA++ [33]	TPAMI'23	36.3	54.6	40.1	31.6	58.0	60.4	46.2	33.6	44.4	66.2	25.7	25.3	44.4	58.8	64.8	55.4	36.2	38.6	54.1	59.3	46.7
CMT [4]	CVPR'23	39.8	56.3	38.7	39.7	60.4	35.0	56.0	7.1	60.1	60.4	35.8	28.1	67.8	84.5	80.1	55.5	20.3	32.8	42.3	38.2	47.0
TFD [49]	AAAI'24	27.9	64.8	28.4	29.5	25.7	64.2	47.7	13.5	47.5	50.9	50.8	21.3	33.9	60.2	65.6	42.5	15.1	40.5	45.5	48.6	41.2
GIBS [39]	ICASSP'25	32.5	60.1	28.2	37.0	38.3	49.5	37.5	22.7	38.2	66.7	27.8	16.6	42.5	86.2	63.6	49.9	35.8	22.6	45.0	31.8	41.6
CDMT [2]	TCSVT'25	50.0	60.4	38.1	42.3	57.9	42.2	54.9	8.3	63.3	50.6	40.3	26.3	58.5	80.5	79.6	58.8	25.0	35.7	34.4	46.6	47.7
SEEN-DA [29]	CVPR'25	44.1	73.4	54.7	47.1	45.1	76.0	51.6	20.4	51.7	53.0	18.5	17.3	61.8	86.8	72.2	22.8	37.7	21.1	58.9	52.7	47.9
ETS (Ours)	-	39.6	66.3	39.1	49.1	48.9	52.2	52.9	9.5	51.9	51.2	48.2	26.4	49.8	91.7	70.0	52.1	19.7	48.6	50.8	62.6	49.0

Table 2. Results of Cityscapes to BDD100k (**C→B**). The average precision (AP, %) on all classes is presented.

Method	Venue	person	rider	car	truck	bus	motorcycle	bicycle	mAP
MTTrans [57]	ECCV'22	44.1	30.1	61.5	25.1	26.9	17.1	23.0	32.6
BiADT [22]	ICCV'23	42.0	34.5	59.9	17.2	19.2	17.8	24.4	32.7
MIC [24]	CVPR'23	49.8	36.8	68.1	24.0	25.6	18.7	30.7	36.2
HT [16]	CVPR'23	53.4	40.4	63.5	27.4	30.6	28.2	38.0	40.2
MTM [50]	AAAI'24	53.7	35.1	68.8	23.0	28.8	23.8	28.0	37.3
SOCCER [13]	MM'24	56.8	42.2	73.1	31.0	29.5	26.1	33.8	41.8
DATR [7]	TIP'25	58.5	42.8	73.4	26.9	39.9	24.2	37.3	43.3
MGCAMT [6]	TIP'25	61.2	44.9	75.3	32.5	30.0	31.9	37.8	44.8
DA [23]	AAAI'25	61.4	45.4	75.4	33.0	36.2	29.5	36.7	45.8
DT [26]	CVPR'25	51.6	47.0	66.6	44.3	45.9	38.3	40.8	47.8
ETS (Ours)	-	53.9	49.2	67.2	47.1	47.2	40.4	43.5	49.8

Table 3. Results of Cityscapes to Foggy Cityscapes (**C→F**). The average precision (AP, %) on all classes is presented.

Method	Venue	person	rider	car	truck	bus	train	motorcycle	bicycle	mAP
AT [34]	CVPR'22	45.5	55.1	54.3	35.0	56.3	54.3	38.5	51.9	50.9
BiADT [22]	ICCV'23	52.2	58.9	69.2	31.7	55.0	45.1	42.6	51.3	50.8
CMT [4]	CVPR'23	47.0	55.7	64.5	39.4	63.2	51.9	40.3	53.1	51.9
MTM [50]	AAAI'24	51.0	53.4	67.2	37.2	54.4	41.6	38.4	47.7	48.9
REACT [28]	TIP'24	51.4	57.9	67.4	32.7	55.9	49.1	40.1	50.3	50.4
SOCCER [13]	MM'24	51.7	57.7	68.6	38.2	51.6	47.5	41.6	51.7	51.1
PACF [36]	arXiv'25	45.6	57.1	63.2	38.7	62.5	53.6	44.2	53.6	52.0
CDMT [2]	TCSVT'25	46.3	56.5	64.1	39.0	63.0	53.7	41.3	55.6	52.5
DATR [7]	TIP'25	61.6	60.4	74.3	35.7	60.3	35.4	43.6	55.9	53.4
DT [26]	CVPR'25	48.5	60.0	65.4	47.2	66.5	52.9	46.2	56.7	55.4
ETS (Ours)	-	49.0	61.8	65.8	41.7	69.9	56.8	46.7	57.7	56.2

5.2. Implementation Details

We use Faster R-CNN [43] as the object detector. We employ DINOv3_{vit-H+} [45] as the default expert model backbone. Input images are resized to 592-pixel height for expert model training (DINOv3's patch size is 16) and 600-pixel height for teacher-student training. We set the confidence thresholds at $\delta=0.8$ and $\delta'=1.0$, and confidence discrepancy threshold $\epsilon=0.15$. The expert model trains 40,000 iterations, while the teacher-student framework undergoes a 20,000-iteration burn-in stage and an 80,000-iteration mutual-learning stage. We begin prototype alignment at the 25,000-th iteration. For both **C→F** and **C→B**, we employ VGG16 [46] as the teacher/student backbone with a learning rate of 0.04. For **P→CI**, following [2, 4, 34], we employ ResNet101 [21] as the teacher/student backbone with a learning rate of 0.01. All experiments are conducted on 4 RTX3090 GPUs using Detectron2 [51] with 8 source and 8 target images per batch. Additional experiment details and parameter settings are provided in Appendix D.

5.3. Comparison with SOTA

We use the mean Average Precision (mAP) with a threshold of 0.5 to evaluate the detection performance. Following other teacher-student methods [4, 24, 26, 34], we report the performance of the teacher as the final result of our method. **Real to Artistic Adaptation (P→CI)** In Tab. 1, due to the large number of classes and the significant performance

variation of different methods across them, our focus remains on the average performance. Our method outperforms the SOTA CDMT [2] by 1.3% and SEEN-DA [29] by 1.1%. Further, our ETS outperforms conventional teacher-student methods by +3.3% over AT [34] and +2.0% over CMT [4], which demonstrates the necessity of incorporating experts into the teacher-student framework.

Small to Large-scale Adaptation (C→B) In Tab. 2, ETS achieves 49.8% mAP, outperforming all other models by a considerable margin. Leveraging the fundamental generalization knowledge from VFMs, our ETS effectively distinguishes between easily confused classes such as “motorcycle” and “bicycle”. Furthermore, the prototype alignment enhances ETS’s class representations, enabling superior performance over DT across all classes.

Normal to Foggy Adaptation (C→F) In Tab. 3, our ETS achieves the best mAP of 56.2%. Benefiting from the generalized knowledge transferred of VFMs, our ETS demonstrates improvements on rare classes such as “bus”, “train”, “motorcycle”, and “bicycle”. Moreover, our superior performance over DT on most classes validates the effectiveness of the collaborative learning mechanism.

5.4. Ablation Studies and Component Analysis

Ablation on components. We conduct ablation experiments of each component in Tab. 4. *Row 1* is our method’s baseline, where the student is only supervised by the

Table 4. Ablation study of ETS on $C \rightarrow B$. CD, UD, and ID denote consistent, unilateral, and inconsistent detections, respectively. CD, TS, and ES denote cross-domain, teacher-student, and expert-student prototype alignment, respectively.

row	ETCT			ETJC			mAP
	CD	UD	ID	CD	TS	ES	
1							36.4
2	✓						43.9
3	✓	✓					47.9
4	✓	✓	✓				48.1
5	✓	✓	✓	✓			48.6
6	✓	✓	✓	✓	✓		49.2
7	✓	✓	✓	✓	✓	✓	49.8

Table 5. Hyperparameter ablation of IoU threshold τ and stricter confidence threshold δ' on $C \rightarrow F$. When adjusting one, the other in the gray background remains unchanged. The symbol $\emptyset$ denotes excluding the teacher model's unilateral detections.

τ	0.25	0.5	0.75	1.0
mAP	55.1	56.2	55.5	54.3
δ'	0.8	0.9	0.95	1.0
mAP	54.1	54.6	55.2	55.5

teacher. The performance began to decline after reaching 36.4% mAP, indicating a training collapse. From rows 2-4, we conduct component ablations on ETCT. In row 2, introducing consistency detections effectively prevents training collapse and achieves a continuous performance improvement of 7.5% mAP. This indicates that the introduction of the expert model guides the optimization of the teacher-student framework in the right direction. In row 3, the unilateral detections further enhance performance by 4% mAP through the complementary knowledge provided by the expert and teacher models. With the complete ETCT module (row 4), our model achieves 48.1% mAP, surpassing all existing SOTA methods. From rows 5-7, we perform component ablations on ETJC. In row 5, aligning source and target domain prototypes in the student model results in a 0.5% mAP gain, while student-teacher prototype alignment (row 6) further enhances domain-specific representations, bringing an additional 0.6% mAP improvement. In row 7, after incorporating expert prototype alignment, the ETS achieves the best performance of 49.8% mAP.

Ablation on hyperparameters. In Tab. 5 top, we observe that a minimal τ increases recall for consistent or inconsistent detections, leading to redundant pseudo labels. While an overly stringent τ causes many correct matches to be missed. $\tau = 0.5$ achieves optimal balance. In Tab. 5 bottom, our experiments demonstrate that increasing the confidence threshold effectively filters prediction noise from the teacher model. Specifically, setting $\delta' = 1.0$ yields best

Table 6. Ablation on supervision from different expert models on $C \rightarrow B$. We also compare each expert model's parameters, runtime, and memory consumption.

Expert	Parameters	Throughput	GPU Memory	mAP _{Expert}	mAP _{Ours}
DINOv2 _{vit-L}	300M	1.57 it/s	11.40 GB	45.3	47.7
DINOv3 _{vit-L}	300M	1.99 it/s	8.29 GB	48.8	48.9
DINOv2 _{vit-G}	1,100M	0.58 it/s	21.80 GB	50.5	49.5
DINOv3 _{vit-H+}	840M	1.02 it/s	15.62 GB	51.4	49.8
GDINO _{swin-L}	341M	0.72 it/s	10.36 GB	51.6	50.0

Table 7. Comparison of different DAOD methods in terms of computational efficiency and memory consumption on $C \rightarrow B$.

Method	Student	Teacher	Expert	Throughput	GPU Memory	mAP
AT [34]	✓	✓		2.03 it/s	6.98 GB	35.5
DT [26]	✓		✓	1.79 it/s	8.38 GB	47.8
ETS	✓	✓	✓	1.82 it/s	10.88 GB	49.8

Table 8. Ablation of different expert prototypes on $P \rightarrow Cl$.

Prototype From	Dimension	Layer	mAP _{0.5:0.95}	mAP ₇₅	mAP
DINOv2 _{vit-G}	128	2	22.5	17.7	48.5
DINOv3 _{vit-H+}	128	2	22.9	19.0	49.0

performance compared to without the teacher's unilateral detections (*i.e.*, $\emptyset$).

Ablation on different expert models. In Tab. 6, we report the source-only performance of various expert models and our ETS's performance supervised by different expert models, respectively. In this experiment, the ETJC module remains unchanged. MM GDINO [66], fine-tuned on the source domain (see Appendix A for details), achieves the best performance of 51.6% mAP. DINOv3 outperforms its DINOv2 counterpart with comparable parameters by 1-3%. As the parameter count of the expert model increases, their performance continues to improve; however, this comes at the cost of reduced training efficiency and higher GPU memory consumption. Fortunately, by leveraging experts in a free lunch manner, our ETS framework preserves their strong performance without sacrificing computation efficiency, as shown in Tab. 7 last row. We also make two observations: (1) The performance of the student is positively correlated with that of the expert. When using GDINO as the expert, our performance achieves 50.0%. (2) As seen in the first two rows, the expert's performance is not the upper bound for the student, because when the expert's performance is not enough to support the student's target domain learning, the teacher plays a greater teaching role.

Computational efficiency comparison. In Tab. 7, we compare our ETS with AT [34] and DT [26]. Although our framework consumes more GPU memory, it achieves higher throughput *i.e.*, higher training efficiency, than DT

Table 9. The performance of the expert model DINOv3_{VIT-H+} on each benchmark.

row	Source → Target	mAP
1	Pascal VOC → Clipart1k	63.8
2	Cityscapes → BDD100k	51.4
3	Cityscapes → Foggy	58.8

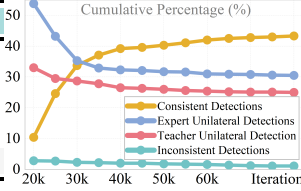
Table 10. Ablation of different prototype alignment losses on C→B.

row	Loss Type	mAP
1	L2 loss	49.0
2	Contrastive loss	49.2
3	Cosine loss	49.8

Table 11. Ablation on different supervision strategies.

row	Supervision (C→B)	mAP
1	Only Teacher	38.1
2	Only Expert	47.6
3	Expert → Teacher	44.9
4	Expert ⇌ Teacher	47.1
5	Collaborative teaching	49.8

Figure 4. Visualization of the collaborative learning training.



by leveraging offline-stored prototypes. This advantage stems from DT’s reliance on DINOv2_{VIT-B} for online feature alignment, which incurs significant computational overhead. While our memory usage and throughput are not as advantageous as those of AT, our performance outperforms AT by 14% mAP.

Ablation on different expert prototypes. In Tab. 8, the ETCT module remains unchanged (*i.e.*, supervised by DINOv3_{VIT-H+}). The results indicate that DINOv3 is more effective than DINOv2 at guiding student learning the generalized class representation in the target domain.

Ablation on different alignment losses. In Tab. 10, we compare common alignment losses, *i.e.*, L2, contrastive, and cosine losses. While all these losses achieve performance improvements compared to using no alignment loss (in Tab. 4 row 4 48.1% mAP), the cosine similarity loss demonstrates the most significant gains (+1.7%).

Ablation on different supervision strategies. As shown in Tab. 11, we study different supervision strategies to validate the effectiveness of collaborative teaching. Beyond relying solely on the teacher (row 1) or expert (row 2) for supervision, we follow DT’s setting by employing the teacher exclusively after 25,000 iterations (row 3), as well as alternately utilizing both the expert and teacher models (row 4). The results demonstrate that neither single model can effectively supervise student learning, while simple replacement or alternation fails to inherit knowledge from both models.

The proportion of different detections. Fig. 4 illustrates the proportion of different detections during the training process. Initially, unilateral detections are most numerous, particularly from the expert (>50%). As training progresses, the number of consistent detections increases while unilateral and inconsistent detections decrease, demonstrating that the disagreement between the expert and teacher diminishes and the models optimize in the correct direction.

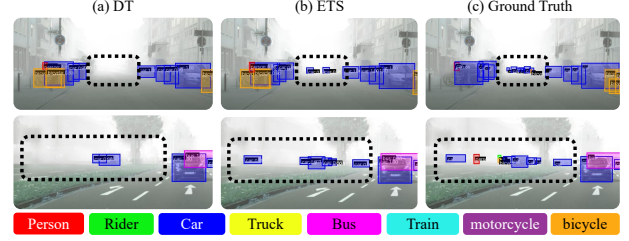


Figure 5. Qualitative results comparison on C→F.

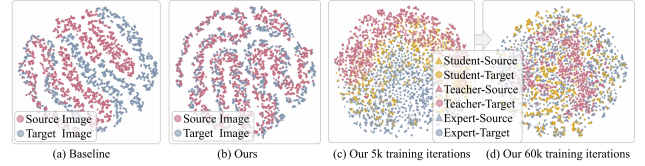


Figure 6. T-SNE visualization of extracted image features.

5.5. Visualization Analysis

Qualitative Visualization The qualitative results are displayed in Fig. 5. Compared to DT, which discards the teacher model, our ETS leverages domain-specific knowledge distilled from the teacher model to identify objects even in dense fog regions accurately (as indicated by the dashed boxes). We provide more detailed qualitative analysis in Appendix E and Appendix F.

T-SNE Visualization As shown in Fig. 6 (a) and (b), the self-training collapsed baseline (*i.e.*, Tab. 4 row 1) fails to bridge the domain gap, while ETS effectively extracts domain-invariant image features. Fig. 6 (c) and (d) illustrate the feature distributions extracted by each model’s backbone during early and late training iterations, respectively, on C→B. The results demonstrate that the student model gradually acquires knowledge from both expert and teacher models, with its feature distribution progressively intermediate between both during training.

6. Conclusions

In this paper, we reveal that teacher models are as essential as visual foundation models for domain adaptive object detection. To integrate their complementary strengths, we propose Expert-Teacher-Student Collaborative Learning (ETS), which follows a teach-then-consolidate strategy. Specifically, the expert model and the teacher model first collaborate to generate pseudo labels, transferring both generalized knowledge and domain-specific knowledge to the student model. Subsequently, the teacher and expert models jointly consolidate the student’s class representations via prototype alignment, enhancing generalized feature representation from source to target domain. Through these operations, ETS achieves superior adaptation performance, and we conduct comprehensive experiments to validate the cross-domain effectiveness.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62322211, 62336008), the National Key R&D Program of China under Grant (2023YFB4502803), the “Pioneer” and “Leading Goose” R&D Program of Zhejiang Province (2024C01023), and the Key R&D Program Projects in Zhejiang Province (2024C01232).

References

- [1] Tajamul Ashraf, Rajes Manna, Partha Sarathi Purkayastha, Tavaheed Tariq, and Janibul Bashir. Context aware grounded teacher for source free object detection. *arXiv preprint arXiv:2504.15404*, 2025. 2, 3
- [2] Yunfei Bai, Yiqiang Wu, Bin Zhu, and Xiaomao Li. Contrastive-domain mean teacher for domain adaptive object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 6
- [3] Atif Belal, Akhil Meethal, Francisco Perdigon Romero, Marco Pedersoli, and Eric Granger. Multi-source domain adaptation for object detection with prototype-based mean teacher. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1277–1286, 2024. 2
- [4] Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, and Yu-Xiong Wang. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23839–23848, 2023. 6
- [5] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debaised self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35:32424–32437, 2022. 1
- [6] Jiangming Chen, Li Liu, Wanxia Deng, Zhen Liu, Yu Liu, Yingmei Wei, and Yongxiang Liu. Refining pseudo labeling via multi-granularity confidence alignment for unsupervised cross domain object detection. *IEEE Transactions on Image Processing*, 2025. 6
- [7] Liang Chen, Jianhong Han, and Yupei Wang. Datr: Unsupervised domain adaptive detection transformer with dataset-level adaptation and prototypical alignment. *IEEE Transactions on Image Processing*, 2025. 2, 6
- [8] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yuet-ing Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*, 2022. 2
- [9] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1
- [10] Gaoxiang Cong, Liang Li, Jiadong Pan, Zhedong Zhang, Amin Beheshti, Anton van den Hengel, Yuankai Qi, and Qingming Huang. Flowdubber: Movie dubbing with llm-based semantic-aware learning and flow matching based voice enhancing. In *ACM MM*, pages 905–914, 2025. 2
- [11] Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Hengel, Jian Yang, and Qingming Huang. Emodubber: Towards high quality and emotion controllable movie dubbing. In *CVPR*, pages 15863–15873, 2025. 2
- [12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5
- [13] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Heng Jin, and Li Wu. Stochastic context consistency reasoning for domain adaptive object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1331–1340, 2024. 3, 6
- [14] Yiming Cui, Liang Li, Haibing Yin, Yuhua Gao, Yaoqi Sun, and Chenggang Yan. Debaised teacher for day-to-night domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2577–2587, 2025. 1, 2
- [15] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 6
- [16] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23829–23838, 2023. 2, 6
- [17] Peihua Deng, Jiehua Zhang, Xichun Sheng, Chenggang Yan, Yaoqi Sun, Ying Fu, and Liang Li. Multi-granularity class prototype topology distillation for class-incremental source-free unsupervised domain adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30566–30576, 2025. 2
- [18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5
- [19] Boyang Guo, Liang Li, Jiehua Zhang, Yaoqi Sun, Chenggang Yan, and Xichun Sheng. Prompt learning with knowledge regularization for pre-trained vision-language models. *IEEE Transactions on Multimedia*, 2025. 2
- [20] Jianhong Han, Yupei Wang, and Liang Chen. Vfm-guided semi-supervised detection transformer under source-free constraints for remote sensing object detection. *arXiv preprint arXiv:2508.11167*, 2025. 2
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [22] Liqiang He, Wei Wang, Albert Chen, Min Sun, Cheng-Hao Kuo, and Sinisa Todorovic. Bidirectional alignment for domain adaptive detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18775–18785, 2023. 6

- [23] Xinyu He, Xinhui Li, and Xiaojie Guo. Differential alignment for domain adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17150–17158, 2025. 6
- [24] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 3, 6
- [25] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 5
- [26] Marc-Antoine Lavoie, Anas Mahmoud, and Steven L Waslander. Large self-supervised models bridge the gap in domain adaptive object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4692–4702, 2025. 1, 2, 3, 6, 7
- [27] Haochen Li, Rui Zhang, Hantao Yao, Xinkai Song, Yifan Hao, Yongwei Zhao, Ling Li, and Yunji Chen. Learning domain-aware detection head with prompt tuning. *Advances in Neural Information Processing Systems*, 36:4248–4262, 2023. 3
- [28] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, and Ling Li. React: Remainder adaptive compensation for domain adaptive object detection. *IEEE Transactions on Image Processing*, 2024. 6
- [29] Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, Shaohui Peng, Yongwei Zhao, Chen Zhao, Yanjun Wu, et al. Seen-da: Semantic entropy guided domain-aware attention for domain adaptive object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25465–25475, 2025. 6
- [30] Liang Li, Gaoxiang Cong, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Quan Z Sheng, Qingming Huang, and Ming-Hsuan Yang. Dubbing movies via hierarchical phoneme modeling and acoustic diffusion denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [31] Shuaifeng Li, Mao Ye, Lihua Zhou, Nianxin Li, Siying Xiao, Song Tang, and Xiatian Zhu. Cloud object detector adaptation by integrating different source knowledge. *Advances in Neural Information Processing Systems*, 37:25251–25283, 2024. 2, 3, 4
- [32] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5300, 2022. 6
- [33] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9022–9040, 2023. 6
- [34] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7581–7590, 2022. 3, 6, 7
- [35] Zeyi Li, Pan Wang, and Zixuan Wang. Flowganomaly: Flow-based anomaly network intrusion detection with adversarial learning. *Chinese Journal of Electronics*, 33(1):58–71, 2024. 2
- [36] Chenguang Liu, Yongchao Feng, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Pacf: Prototype augmented compact features for improving domain adaptive object detection. *arXiv preprint arXiv:2501.08605*, 2025. 2, 6
- [37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2, 3
- [38] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3
- [39] Haiming Peng, Dingkan Yang, Mingxu Wang, Weilong Lin, and Xinhua Zeng. Guiding inter-domain class balancing with salient features for domain adaptive object detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 6
- [40] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibao Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Authorea Preprints*, 2025. 2
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3
- [42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3
- [43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6
- [44] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 5
- [45] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2, 3, 6

- [46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [47] Jingyi Tang, Li Liang, Beichen Zhang, and Qingming Huang. Lmda: Llm-guided marginal distribution alignment for open-set active learning. *Chinese Journal of Electronics*, 2026. 2
- [48] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4926–4943, 2024. 2
- [49] Haoan Wang, Shilong Jia, Tiejong Zeng, Guixu Zhang, and Zhi Li. Triple feature disentanglement for one-stage adaptive object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5401–5409, 2024. 6
- [50] Weixi Weng and Chun Yuan. Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5912–5920, 2024. 6
- [51] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6
- [52] Zhicong Wu, Hongbin Xu, Gang Xu, Ping Nie, Zhixin Yan, Jinkai Zheng, Liangqiong Qu, Ming Li, and Liqiang Nie. Textsplat: Text-guided semantic fusion for generalizable gaussian splatting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8478–8487, 2025. 2
- [53] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. 2
- [54] Jiong Yin, Liang Li, Jiehua Zhang, Chenggang Yan, Lei Zhang, and Zunjie Zhu. Reducing intrinsic and extrinsic data biases for moment localization with natural language. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4584–4594, 2023. 2
- [55] Jiong Yin, Liang Li, Jiehua Zhang, Yuhao Gao, Chenggang Yan, and Xichun Sheng. Progressive homeostatic and plastic prompt tuning for audio-visual multi-task incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2022–2033, 2025. 2
- [56] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 5
- [57] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mtrnns: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, pages 629–645. Springer, 2022. 6
- [58] Beichen Zhang, Liang Li, Shuhui Wang, Shaofei Cai, Zheng-Jun Zha, Qi Tian, and Qingming Huang. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9780–9796, 2024. 2
- [59] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3252–3261, 2022. 1
- [60] Tao Zhang, Ying Fu, and Jun Zhang. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1):303–312, 2024. 2
- [61] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhao Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: Movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pages 7523–7532, 2024. 2
- [62] Zhedong Zhang, Liang Li, Jiehua Zhang, Zhenghui Hu, Hongkui Wang, Chenggang Yan, Jian Yang, and Yuankai Qi. Generating high-quality symbolic music using fine-grained discriminators. In *International Conference on Pattern Recognition*, pages 332–344. Springer, 2024. 2
- [63] Zhedong Zhang, Liang Li, Chenggang Yan, Chunshan Liu, Anton Van Den Hengel, and Yuankai Qi. Prosody-enhanced acoustic pre-training and acoustic-disentangled prosody adapting for movie dubbing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 172–182, 2025. 2
- [64] Zhedong Zhang, Liang Li, Gaoxiang Cong, Chunshan Liu, Yuhao Gao, Xiaowan Wang, Tao Gu, and Yuankai Qi. Instructdubber: Instruction-based alignment for zero-shot movie dubbing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12988–12996, 2026. 2
- [65] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14217–14226, 2022. 6
- [66] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 3, 7
- [67] Zijiang Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, and Yang Liu. Masked retraining teacher-student framework for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19039–19049, 2023. 1
- [68] Zhiqian Zhao, Liang Li, Lei Shen, Xichun Sheng, Yaoqi Sun, Fang Kang, and Chenggang Yan. Temporal calibrating and distilling for scene-text aware text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13323–13331, 2026. 2
- [69] Jinkai Zheng, Xinchun Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20228–20237, 2022. 2