AsyFOD: An Asymmetric Adaptation Paradigm for Few-Shot Domain Adaptive Object Detection

Yipeng Gao[1,3*], Kun-Yu Lin[1,3*], Junkai Yan[1,3], Yaowei Wang[2], Wei-Shi Zheng[1,2,3†]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China, [2]Pengcheng Lab.
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{gaoyp23, linky5, yanjk3}@mail2.sysu.edu.cn, wangyw@pcl.ac.cn, wszheng@ieee.org

Abstract

In this work, we study few-shot domain adaptive object detection (FSDAOD), where only a few target labeled images are available for training in addition to sufficient source labeled images. Critically, in FSDAOD, the data scarcity in the target domain leads to an extreme data imbalance between the source and target domains, which potentially causes over-adaptation in traditional feature alignment. To address the data imbalance problem, we propose an asymmetric adaptation paradigm, namely AsyFOD, which leverages the source and target instances from different perspectives. Specifically, by using target distribution estimation, the AsyFOD first identifies the target-similar source instances, which serves to augment the limited target instances. Then, we conduct asynchronous alignment between target-dissimilar source instances and augmented target instances, which is simple yet effective for alleviating the over-adaptation. Extensive experiments demonstrate that the proposed AsyFOD outperforms all state-of-the-art methods on four FSDAOD benchmarks with various environmental variances, e.g., 3.1% mAP improvement on Cityscapes-to-FoggyCityscapes and 2.9% mAP increase on Sim10k-to-Cityscapes. The code is available at https://github.com/Hlings/AsyFOD.

1. Introduction

Object detection [4, 18, 47–50], which aims to localize and classify objects simultaneously, is widely used in real-world applications such as video surveillance [29,42,67,70] and autonomous driving [57, 69]. Unfortunately, detectors suffer a significant performance drop when deployed in an unseen domain due to the domain discrepancy between training and test data [36,37,43,63,66,68]. And usually, repeatedly collecting a large amount of labeled data in new domains requires expensive labor and time cost. In this work, we explore the Few-Shot Domain Adaptive Object Detection (FSDAOD) [16, 61, 72], which attempts to generalize detectors with minor cost. In addition to adequate labeled source images, the FSDAOD assumes that only a few (usually eight) labeled target images are available for adapting a detector in the target domain.

A critical challenge of the FSDAOD is the data scarcity in the target domain, which leads to an extreme data imbalance between the source and target domains. As shown in Figure 1, it is difficult to comprehensively describe the

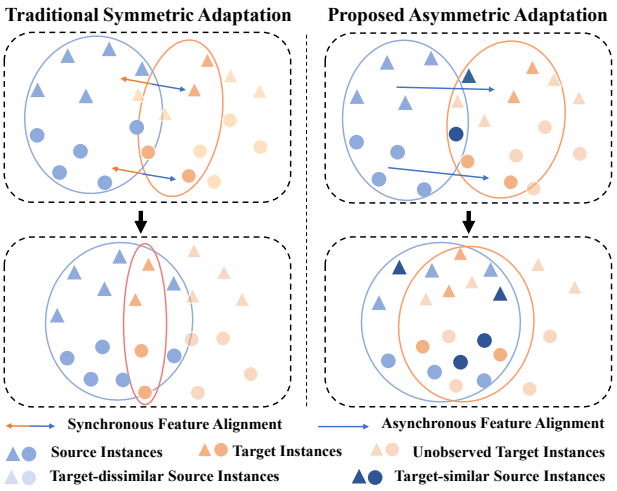

Figure 1. A few target instances are biased to represent the overall target data distribution, *i.e.*, many light orange target instances are not observed, as shown in the top left. And, the data scarcity in the target domain leads to a data imbalance between the source and target domains. Therefore, traditional symmetric adaptation (such as MMD [38, 40, 41, 56]) easily causes over-adaptation, *i.e.*, the detector concentrates on a small area for observed target instances but ignores many other unobserved ones, as shown in the top right. By contrast, our proposed asymmetric adaptation alleviates the over-adaptation via source instance division and asynchronous alignment.

* denotes the authors contributed equally to this work.
† denotes the corresponding author.

overall target data distribution by only a few target instances. Usually, in standard unsupervised domain adaptation [22, 24, 38, 53, 71, 73], alignment-based methods, *e.g.*, Maximum Mean Discrepancy (MMD) [38, 40, 41, 56], conduct synchronous feature alignment to mitigate the domain discrepancy, which is termed the *symmetric adaptation* paradigm in our work. However, without consideration of imbalanced distributions in FSDAOD, simply conducting synchronous feature alignment easily causes the over-adaptation problem, *i.e.*, the detector is prone to concentrate on limited observed target instances but hardly generalizes well on other unobserved ones [61]. Typically, existing FSDAOD methods attempt to alleviate the imbalance problem by reusing the same target samples, which yet overlooks the leverage of source samples [61, 72].

To address the extreme data imbalance problem, in this work, we propose a novel *asymmetric* adaptation paradigm, named AsyFOD, which leverages the source and target instances from different perspectives. The AsyFOD first divides the source instance set into two parts, namely target-similar and target-dissimilar instance sets. Such a division strategy is inspired by an observation that, some source instances are visually similar to the target instances (see Figure 5 for empirical verification). Accordingly, we identify the target-similar source instances by formulating a unified discrepancy estimation function, which serves to augment the limited target instances to alleviate the imbalanced amounts of data. The remaining source instances are regarded as target-dissimilar after identifying target-similar source instances. To further alleviate the data imbalance between domains, we propose conducting asynchronous alignment between the target-dissimilar source instances and augmented target instances. Unlike traditional methods, the AsyFOD aligns feature distributions in an asymmetric way, with a stop-gradient operation applied on target instance features when optimizing the detector. In this way, the proposed asynchronous alignment can better align the unobserved target samples.

The AsyFOD obtains the state-of-the-art performance on mitigating various types of domain discrepancy, such as background variations [9, 17], natural weather [54] and synthetic-to-real [1, 26, 35]. Also, the AsyFOD generalizes well on various few-shot settings of domain adaptive object detection, *i.e.*, FSDAOD with weak or strong augmentation [16] and Few-Shot Unsupervised Domain Adaptive Object Detection (FSUDAOD).

2. Related Work

General Domain Adaptation. Unsupervised domain adaptation (UDA) [14] aims to transfer the recognition power from a label-sufficient source domain to a label-free target domain. Typically, UDA methods mitigate the domain discrepancy via domain-invariant feature learning. Specifically, some adversarial-based works [15, 39, 58] learn domain-invariant features by confusing an auxiliary domain discriminator. Also, some methods align feature distributions between different domains via minimizing well-defined statistical distances [20, 25, 38, 40, 41, 56], such as Maximum Mean Discrepancy (MMD). Considering the privacy or data protection issues, source-free domain adaptation (SFDA) assumes that the source data are not accessible during training. Previous methods are usually based on clustering samples in the target domain [12, 31]. In contrast, A^2Net [64] utilizes the target samples discriminatively and performs intra-domain alignment in the target domain. While most previous works assume a label-free target domain, some works consider supervised domain adaptation. One of the most representative settings is supervised few-shot domain adaptation (FDA) [44]. In FDA, only a few labeled target images are available, with the absence of any unlabeled target samples. Saeid *et al.* [44] construct pairs of samples using source and target samples for FDA.

Unsupervised Domain Adaptive Object Detection. Similar to general recognition, a source pre-trained detector suffers a performance drop when applied in the target domain. To address this problem, Unsupervised Domain Adaptive Object Detection (UDAOD) investigates the unsupervised domain adaptation problem in object detection. Existing UDAOD works can mainly be summarized into four types. The first one is pseudo labeling [27, 30, 46], which exploits pseudo labels of the target images. The second one uses an auxiliary model strategy, which trains an auxiliary model [7, 21, 27, 65] to assist the detector during the adaptation process, *e.g.*, mean teacher [3, 11, 33]. The third one utilizes data generation [28, 60], which aims to transfer the style of source and target images using a generative model [74]. The final one is domain alignment [6, 8, 22, 28, 53, 71, 75], which aligns different types of features at multiple levels [73].

Few-Shot Domain Adaptive Object Detection. Similar to FDA, some works study Few-Shot Domain Adaptive Object Detection (FSDAOD). In this scenario, adequate labeled source samples and only *a few labeled* target samples are available for bridging the domain gap. Therefore, the imbalanced data distributions between source and target domains easily cause over-adaptation. Wang *et al.* [61] adopt the pairing mechanism, which pairs source samples with target samples for multi-level alignment. PICA [72] exploits pixel-wise instance-level alignment coupled with moving average class centroids. In contrast, our work proposes an asymmetric adaptation paradigm that leverages source and target instances differently to address the data imbalance problem. In addition to alignment-based methods, recent work [16] shows the effectiveness of the domain-mix augmentation to augment limited target images, which is orthogonal to the work presented in this paper.

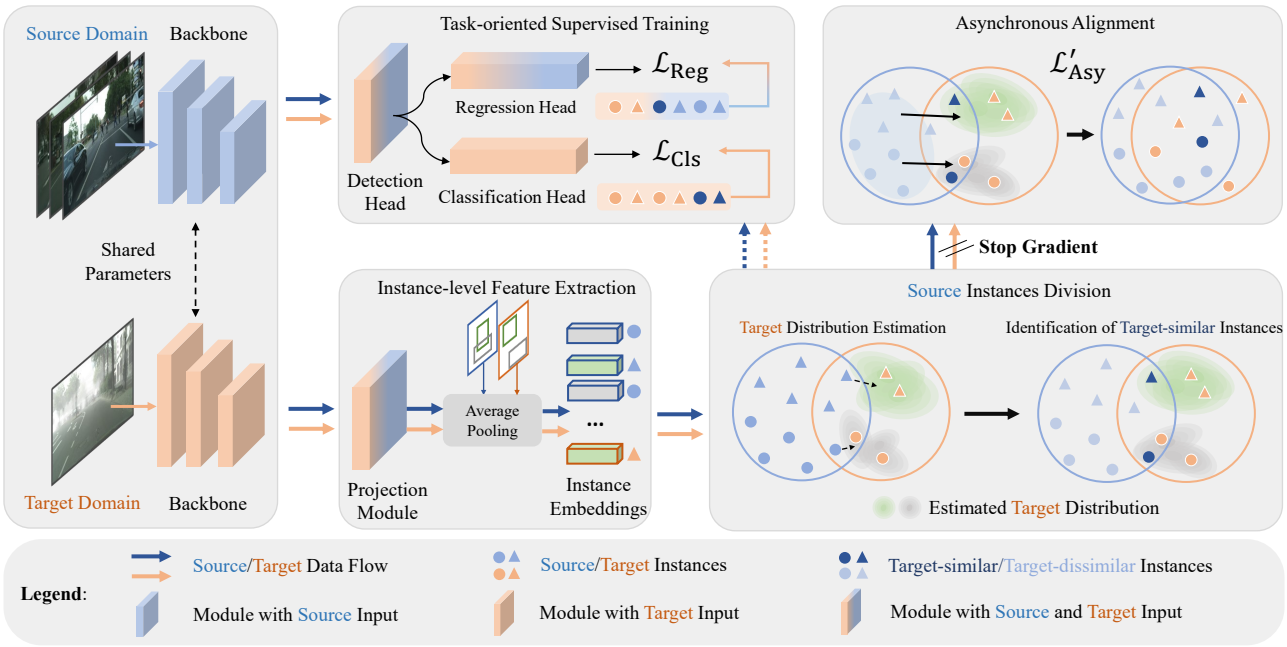

Figure 2. A brief description of the AsyFOD. After extracting the feature map of the input images via the backbone, the instance-level feature embeddings are extracted by the projection module and ground truths. Then, the AsyFOD divides the source instances into target-similar (dark blue ones) and target-dissimilar sets (light blue ones). The target-similar source instances are used to augment the target instances. With the above division, the asynchronous alignment is conducted between target-dissimilar source and augmented target instances, alleviating the premature over-adaptation due to extremely imbalanced data distributions. Also, the AsyFOD performs task-oriented supervised training for classification and localization tasks separately. Best viewed in color.

3. AsyFOD

The overall training pipeline is illustrated in Figure 2. We first recap the problem formulation and general supervised detector training in Section 3.1. Then, we present the proposed asymmetric adaptation paradigm in Section 3.2, including source instance division (Section 3.2.1) and asynchronous distribution alignment (Section 3.2.2). Finally, Section 3.3 summarizes the whole adapting process.

3.1. Preliminaries

Problem Formulation. Given labeled source data $\mathcal{D}^s = \{(x_i^s, y_i^s, b_i^s)\}_{i=1}^{N_s}$ and labeled target data $\mathcal{D}^t = \{(x_j^t, y_j^t, b_j^t)\}_{j=1}^{N_t}$, Few-Shot Domain Adaptive Object Detection (FSDAOD) aims to adapt detector from the data-sufficient source domain to the data-scarce target domain (*i.e.*, $N_s \gg N_t$). Specifically, $x_i^s, x_j^t \in \mathcal{X}$ denote the i^{th}, j^{th} images and $(y_i^s, b_i^s), (y_j^t, b_j^t) \in \mathcal{Y}$ are corresponding labels consisting of object categories y_i^s, y_j^t and position coordinates b_i^s, b_j^t. We focus on transfer scenarios with domain discrepancy between the source distribution $\mathcal{P}^s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ and the target distribution $\mathcal{P}^t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. Without loss of generality, we assume that the source and target domains share identical label space but with different data distributions ($\mathcal{P}^s(x) \neq \mathcal{P}^t(x)$).

General supervised object detection. Modern detectors [2, 13, 18, 47–50] consist of a backbone network f and a detect head h. To pre-train the detector in $\mathcal{D}^s$, the overall loss function is as follows:

$$\mathcal{L}_{\text{Det}} = \mathbb{E}_{(x^s, y^s, b^s) \in \mathcal{D}^s} \mathcal{L}_{\text{Reg}} + \mathcal{L}_{\text{Cls}}, \tag{1}$$

where $\mathcal{L}_{\text{Reg}}$ and $\mathcal{L}_{\text{Cls}}$ denote the bounding box regression and classification loss, respectively. Our proposed AsyFOD aims to transfer the detector $h \circ f$ with limited labeled images available in the target domain.

3.2. The Asymmetric Adaptation Paradigm

The data scarcity in FSDAOD leads to inherent imbalanced distribution between the source and target domains. If performing traditional feature alignment (*e.g.*, Maximum Mean Discrepancy (MMD) [38, 40, 41, 56]), the detector is prone to concentrate on limited observed target instances but hardly generalizes well on other unobserved ones.

To this end, we propose an asymmetric adaptation paradigm to leverage a large amount of source instances and a few target instances from different perspectives. Firstly, we divide the source instance set into target-similar/-dissimilar sets and use the former to augment limited target instances. Secondly, we propose to conduct asynchronous alignment between the target-dissimilar source instance set

and augmented target instance set, which is a simple yet effective method specific to FSDAOD. Furthermore, we propose task-oriented supervised training to alleviate the extreme data imbalance in the supervised training scheme.

3.2.1 Source Instances Division

We first extract the feature embeddings of instances from $\mathcal{D}^s \bigcup \mathcal{D}^t$. The feature map is produced by the projection module g and backbone network f, $z_i^s = g \circ f(x_i^s), z_j^t = g \circ f(x_j^t)$. The instance-level embedding o_j^t of j^{th} instance in $\mathcal{D}^t$ is extracted via average pooling, $o_j^t = \text{AvgPooling}(S(z_j^t, b_j^t))$. The S denotes the function that slices the z_j^t according to the bounding box b_j^t [5]. With z_i^s, b_i^s, the embeddings of source instances o_i^s are also obtained in a similar way.

Then, we estimate the target instances distribution via observed $\mathbf{o}^t = \{o_j^t\}$. Specifically, we design a discrepancy estimation function for getting the target-conditional probability density function $\mathcal{I}(\gamma^t)$ parameterized by γ^t through $\mathbf{o}^t$. Without loss of generality, the γ^t is determined by maximizing the log likelihood over all the instances $o_j^t \in \mathbf{o}^t$:

$$\gamma^t = \underset{\gamma}{\text{argmax}} \sum\nolimits_{o_j^t \in \mathbf{o}^t} \log(\mathcal{I}(o_j^t|\gamma)), \tag{2}$$

where $\mathcal{I}(o_j^t|\gamma^t)$ denotes the probability of o_j^t in $\mathcal{I}(\gamma^t)$.

Due to imbalanced distributions between domains, the number of target instances $|\mathbf{o}^t|$ is relatively small, so the γ^t is hard to represent the overall target domain. Inspired by the observation that some source instances partially describe the $\mathcal{P}^t$, we augment $\mathbf{o}^t$ with o_i^s that are close to $\mathcal{D}^t$ with minimal introduced target risk. The set of such source instances is denoted as *target-similar* instance set $\mathbf{o}_{t'}^s$. Specifically, we use $\mathcal{I}(\gamma^t)$ to estimate the target distribution $\mathcal{P}^t$. Then, we identify $\mathbf{o}_{t'}^s$ in $\mathbf{o}^s = \{o_j^s\}$ via the reciprocal of probability that o_i^s belongs to $\mathcal{D}^t \sim \mathcal{P}^t$ as the distance d_i^s:

$$d_i^s = 1/p(o_i^s|\mathbf{o}^t, \gamma^t) = 1/\mathcal{I}(o_i^s|\gamma^t). \tag{3}$$

According to the increasing order of d_i^s, we select the top $|\mathbf{o}_{t'}^s| = \beta * |\mathbf{o}^s|$ source instances with the proportion β. The β controls the introduced estimation error of $\mathbf{o}_{t'}^s$. The remaining source instances belong to the *target-dissimilar* instances set $\mathbf{o}_{s'}^s$.

We provide three discrepancy estimation functions for estimating $\mathcal{I}(\gamma^t)$, including the L_2 distance with average pooling, K-means and Gaussian Mixture Model (GMM).

- L_2 distance. We first calculate the target prototype via average pooling $\gamma^t = \sum_{j=1}^{|\mathbf{o}^t|} o_j^t$. Then, we use the L_2 distance as d_i^s in Eq. (3). For each $o_i^s \in \mathbf{o}^s$, the d_i^s is as follows:

$$d_i^s = ||o_i^s - \gamma^t||_2. \tag{4}$$

- K-means. We first cluster $\mathbf{o}^t$ by K-means and use the clustered centers $\gamma^t = \{\gamma_n^t\}$ to represent $\mathcal{I}$. Then, we utilize the distance between o_i^s and its nearest neighbor in γ^t as d_i^s:

$$d_i^s = \min\{||o_i^s - \gamma_n^t||_2\}_{n=1}^{|\gamma^t|}. \tag{5}$$

- Gaussian Mixture Model (GMM). GMM assumes that $\mathcal{I}(\gamma^t)$ is a weighted mixture of M multivariate Gaussian distributions $\{\mathcal{N}(\mu_m, \Sigma_m)\}_{m=1}^M$. μ_m and Σ_m denote the mean vector and covariance matrix of component m. γ^t is optimized by Expectation-Maximization [10] algorithm. The Eq. (3) is rewritten as follows:

$$d_i^s = 1/\sum\nolimits_{m=1}^{M} \pi_m \mathcal{N}(o_i^s; \mu_m, \Sigma_m), \tag{6}$$

where π_m is the weight of $\mathcal{N}(\mu_m, \Sigma_m)$ constrained to $\sum_{m=1}^M \pi_m = 1$ and $\gamma^t = \{\pi_m, \mu_m, \Sigma_m\}_{m=1}^M$. We compare and discuss Eqs (4)-(6) in Section 4.3.

3.2.2 Asymmetric Adaptation

The optimization objective of detector adaptation is to bridge the cross-domain discrepancy between $\mathcal{D}^s \sim \mathcal{P}^s$ and $\mathcal{D}^t \sim \mathcal{P}^t$. To this end, we develop asymmetric adaptation from two aspects: asynchronous distribution-level alignment and task-oriented supervised training.

Asynchronous Alignment. Previous methods [61, 72] adopt the pairing mechanism to reuse the target instances $\mathbf{o}^t$ and align with the source instances $\mathbf{o}^s$. Generally, if using such a technique, the instance-level distribution feature alignment can first be defined by minimizing the loss $\mathcal{L}_{\mathbf{Sym}}$ as follows:

$$\mathcal{L}_{\mathbf{Sym}} = \mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s, \mathbf{o}^t), \tag{7}$$

where $\mathcal{J}_{\mathbf{Ali}}$ is the metric function that measures the distribution discrepancy, *e.g.*, mean maximum discrepancy [20, 38]. The $\mathcal{L}_{\mathbf{Sym}}$ performs in a symmetric way for aligning $\mathbf{o}^s$ and $\mathbf{o}^t$. Unfortunately, such optimization objective meets obstacle if there exists an extreme imbalance between $\mathbf{o}^t$ and $\mathbf{o}^s$. Due to limited observed target instances o_j^t in $\mathcal{D}^t$, simply aligning the feature distributions of $\mathbf{o}^s$ and $\mathbf{o}^t$ ($|\mathbf{o}^s| \gg |\mathbf{o}^t|$) easily causes over-adaptation [61], *i.e.*, the detector is prone to concentrate on the limited observed target instances $\mathbf{o}^t$ but hardly generalizes well on other unobserved ones.

To this end, we propose the instance-level asynchronous distribution alignment. Specifically, we first detach the gradient of $\mathbf{o}^t$ in the optimization process. As a result, we formulate a new alignment criterion as follows:

$$\mathcal{L}_{\mathbf{Asy}} = \mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s, sg(\mathbf{o}^t)), \tag{8}$$

where $sg(\cdot)$ denotes the stop-gradient operation. The few target instances $o_j^t \in \mathbf{o}^t$ are biased to represent the target distribution $\mathcal{P}^t$ (shown in Figure 1). With $\mathcal{L}_{\mathbf{Asy}}$, we relatively stabilize the feature alignment process between $\mathbf{o}^s$ and $\mathbf{o}^t$ to alleviate the interference caused by the data bias.

Moreover, we adopt target-similar source instances $\mathbf{o}^s_{t'}$ (obtained via Eq. (2)-(3) in Section 3.2.1) to augment the $\mathbf{o}^t$, which alleviates the imbalanced amounts of data. As a result, the asynchronous instance-level distribution alignment is given as follows:

$$\mathcal{L}^{'}_{\mathbf{Asy}} = \mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s_{s'}, sg(\mathbf{o}^s_{t'} \cup \mathbf{o}^t)). \tag{9}$$

Task-oriented Supervised Training. Object detection requires spatially different activated areas in the feature map for regression and classification tasks [55]. Specifically, models are prone to focus on some salient areas for classification, while they should be aware of the boundary areas for localization [55]. Motivated by this, we propose to exploit different instance sets to optimize classification and localization tasks separately.

Classification. Due to the domain shift between the source and target domains, classification supervision of target-dissimilar instances negatively affects target-specific feature learning. Therefore, we propose to force the classifier to focus on target-specific information. Specifically, in the source domain, we use the target-similar instance data $\mathcal{D}^s_{\text{tar}}$ for classification supervision. With the observed target data $\mathcal{D}^t$, the loss is given as follows:

$$\mathcal{L}^{'}_{\mathbf{Cls}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}^s_{\text{tar}}\bigcup\mathcal{D}^t}\mathcal{J}_{\mathbf{Cls}}(h_{\text{Cls}}\circ f(\mathbf{x}), \mathbf{y}), \tag{10}$$

where $\mathcal{J}_{\mathbf{Cls}}$ is the loss function for classification such as cross-entropy and focal loss [34], and h_{Cls} is the classification head in detect head h.

Localization. For better awareness of boundary information, we utilize all the bounding boxes for localization without filtering. The regression objective is given as follows:

$$\mathcal{L}^{'}_{\mathbf{Reg}} = \mathbb{E}_{(\mathbf{x},\mathbf{b})\in\mathcal{D}^s\bigcup\mathcal{D}^t}\mathcal{J}_{\mathbf{Reg}}(h_{\text{Reg}}\circ f(\mathbf{x}), \mathbf{b}), \tag{11}$$

where $\mathcal{J}_{\mathbf{Reg}}$ is the loss function that serves for bounding box regression, *e.g.*, GIoU [52], and h_{Reg} is the localization head in detect head h.

3.2.3 The Optimization Objective

The AsyFOD performs asymmetric adaptation via source instances division following both asynchronous distribution alignment and task-oriented supervised training perspectives. The detector $h_{\theta_2}\circ f_{\theta_1}$ and additional projection module g_{θ_3} parameterized by $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3\}$ are encouraged to overcome the domain gap between $\mathcal{D}^t$ and $\mathcal{D}^s$:

$$\min_{\boldsymbol{\theta}} \mathcal{L}^{'}_{\mathbf{Reg}} + \mathcal{L}^{'}_{\mathbf{Cls}} + \alpha\mathcal{L}^{'}_{\mathbf{Asy}}, \tag{12}$$

where α is the trade-off hyperparameter for balancing the asymmetric feature alignment and supervised losses.

Algorithm 1 The training pipeline of the AsyFOD

Input: Initialized detector with the additional projection module $\boldsymbol{\theta}^{ini}$, the source domain $\mathcal{D}^s$, the target domain $\mathcal{D}^t$, total epochs T, amount of steps for every epoch N, discrepancy estimation function F, proportion $0 < \beta \leq 1$, localization/classification loss function $\mathcal{J}_{\mathbf{Reg}}/\mathcal{J}_{\mathbf{Cls}}$.
Output: Domain Adaptive Detector $h \circ f$ consists of the backbone f_{θ_1} and detect head h_{θ_2}

```
Initialize projection module g_{θ3}
Initialize θ = {θ1, θ2, θ3} ← θ^{ini}
for epoch ← 1, ..., T do
    for step ← 1, ..., N do
        Sample batch B^s = {(x_i^s, y_i^s, b_i^s)}_{i=1}^{n_s} ∈ D^s
        Sample batch B^t = {(x_j^t, y_j^t, b_j^t)}_{j=1}^{n_t} ∈ D^t
        Get source and target instance sets o^s, o^t of B^s, B^t with g_{θ3} ∘ f_{θ1}
        Divide o^s to o^s_{s'} and o^s_{t'} via F, β and Eq. (2)-(3)
        Predictions p = h_{θ2} ∘ f_{θ1}({x_i^s}_{i=1}^{n_s} ⋃ {x_j^t}_{j=1}^{n_t})
        Asynchronous alignment loss L'_Asy between o^s_{s'} and (o^s_{t'} ∪ o^t) in Eq. (9)
        Task-oriented supervised training L'_Reg + L'_Cls with J_Reg and J_Cls, p and D^s_tar ⋃ D^t in Eq. (10)-(11)
        Update θ to minimize Eq. (12)
    f_{θ1}, h_{θ2}, g_{θ3} ← θ
```

3.3. The Overall Training Pipeline

The detailed training pipeline of our AsyFOD is summarized in **Algorithm** 1. After feature extraction, the target-similar source instances $\mathbf{o}^s_{t'}$ will be identified according to the discrepancy estimation function. All parameters used for training will be updated after every iteration. After training, the projection module g can be dropped. Therefore, the AsyFOD does not introduce additional inference costs when applying the adapted detector.

4. Experiments

We present the main results of the proposed AsyFOD for addressing the Few-shot Domain Adaptation Object Detection (FSDAOD) task in various scenarios in Section 4.2. Then, we analyze multiple parts of the AsyFOD in Section 4.3. Furthermore, we conduct a qualitative analysis, including the instance-level feature distribution and retrieved target-similar source instances in Section 4.4.

4.1. Datasets and Experimental Details

- Cityscapes→ Foggy Cityscapes. The Cityscapes [9] contains 3,475 real urban images, with 2,975 images used for training and 500 for validation. Based on the Cityscapes, the Foggy Cityscapes [54] is a synthesized dataset with hand-crafted foggy modification. The highest fog intensity

Setting	Method	Architecture	person	rider	car	truck	bus	train	mcycle	bicycle	mAP50	SO/GAIN
UDAOD	DA-Faster [8]	F-RCNN V	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6	19.9/7.7
	FAFRCNN [61]	F-RCNN V	29.1	39.7	42.9	20.8	37.4	24.1	26.5	29.9	31.3	19.9/11.4
	ViSGA [51]	F-RCNN R	38.8	45.9	57.2	29.9	50.2	51.9	31.9	40.9	43.3	22.8/20.5
	SIGMA [32]	F-RCNN R	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2	24.2/20.0
FSDAOD	FAFRCNN [61]	F-RCNN V	$27.9_{\pm 0.6}$	$37.8_{\pm 0.6}$	$42.3_{\pm 0.7}$	$20.1_{\pm 0.5}$	$31.9_{\pm 1.1}$	$13.1_{\pm 1.5}$	$24.9_{\pm 1.3}$	$30.6_{\pm 0.9}$	$28.6_{\pm 0.5}$	19.9/8.7
	PICA [72]	F-RCNN V	$28.3_{\pm 2.2}$	$41.3_{\pm 0.3}$	$43.0_{\pm 0.4}$	$23.8_{\pm 2.2}$	$38.1_{\pm 1.5}$	$24.3_{\pm 0.8}$	$25.4_{\pm 1.4}$	$33.7_{\pm 0.4}$	$32.2_{\pm 0.8}$	20.3/11.9
	SimRoD [46]	YOLOv5 X	$34.3_{\pm 1.3}$	$35.8_{\pm 0.3}$	$55.9_{\pm 0.8}$	$9.6_{\pm 1.8}$	$18.0_{\pm 0.6}$	$5.9_{\pm 0.3}$	$10.6_{\pm 0.2}$	$29.2_{\pm 0.8}$	$24.9_{\pm 0.2}$	21.9/5.0
	FsDet [62]	YOLOv5 X	$32.3_{\pm 1.2}$	$29.8_{\pm 1.2}$	$44.0_{\pm 1.7}$	$14.1_{\pm 2.2}$	$24.2_{\pm 1.4}$	$8.4_{\pm 1.2}$	$22.9_{\pm 1.6}$	$26.2_{\pm 2.2}$	$25.2_{\pm 1.1}$	21.9/3.3
	AcroFOD [16]	YOLOv5 X	$46.2_{\pm 0.5}$	$47.3_{\pm 0.6}$	$63.5_{\pm 0.4}$	$20.1_{\pm 1.6}$	$41.5_{\pm 0.8}$	$34.2_{\pm 1.8}$	$36.1_{\pm 0.7}$	$39.6_{\pm 0.9}$	$41.1_{\pm 0.8}$	21.9/19.2
	AsyFOD	YOLOv5 X	$46.9_{\pm 0.7}$	$48.7_{\pm 1.1}$	$66.8_{\pm 0.7}$	$26.3_{\pm 1.8}$	$45.1_{\pm 1.2}$	$40.6_{\pm 0.9}$	$40.6_{\pm 0.6}$	$39.2_{\pm 1.3}$	**44.3**$_{\pm 1.0}$	21.9/**22.4**

Table 1. Results (%) on Cityscapes → Foggy Cityscapes. "V"/"R" stand for VGG16/ResNet50 backbone networks. "X" stands for a type of yolov5 model. SO denotes the source-only results, and GAIN represents gains after adaptation compared with the source-only model.

Setting	Method	S→ C	SO/GAIN	K→ C	SO/GAIN
UDAOD	SWDA [53]	44.6	31.9/12.7	43.2	32.5/10.7
	SSAL [45]	51.8	38.0/13.8	45.6	34.9/10.7
	SIGMA [32]	53.7	39.8/13.9	45.8	34.4/11.4
FSDAOD	FAFRCNN [61]	$41.2_{\pm 0.6}$	33.5/7.7	-	-
	PICA [72]	$42.1_{\pm 0.7}$	34.6/7.5	-	-
	FsDet [62]	$52.9_{\pm 1.2}$	49.0/3.9	$52.9_{\pm 1.2}$	47.4/5.5
	SimRoD [46]	$54.2_{\pm 0.5}$	49.0/5.2	$55.8_{\pm 0.6}$	47.4/8.4
	AcroFOD [16]	$62.5_{\pm 1.6}$	49.0/13.5	$62.6_{\pm 2.1}$	47.4/15.2
	AsyFOD	**65.4**$_{\pm 0.9}$	49.0/**16.4**	**64.1**$_{\pm 1.1}$	47.4/**16.7**

Table 2. Comparison results (%) on Sim10K → Cityscapes (S → C) and KITTI →Cityscapes (K → C).

images of 8 classes are used in our experiments.

- SIM10k→Cityscapes and ViPeD→COCO. SIM10k [26] is a simulated dataset from a video game, which contains 10k synthetic images and 58,701 car bounding boxes. The ViPeD [1] contains 200K frames of the person class. We select one frame per 10 frames and a total of 20K frames during training. The subset of the COCO [35] containing only person annotations is chosen as the target domain.

- KITTI→Cityscapes. We use the KITTI [17] as our source data in this scenario. The KITTI contains 7,481 images of the car class, which exists the cross-camera domain discrepancy with Cityscapes (on-board cameras).

Implementation details. We adopt single-stage detector YOLOv5 [13] as the baseline and compare with both Unsupervised Domain Adaptive object detection (UDAOD) and Few-Shot Domain Adaptive object detection (FSDAOD) methods. For the UDAOD setting, the results are based on all target data. For the FSDAOD setting, we randomly select **60/8** fully labeled target images on ViPeD→COCO/the other three scenarios [16].

The projection module g is applied by a 3x3 convolution layer following ReLU [19] and batchnorm [23]. The α/β are set to 1/0.2 by default. In the main results, we use strong augmentation [16] by default for a fair comparison. We report average precision with an IoU threshold of 0.5 as AP50/mAP50 for single/multi classes, and AP or mAP for 10 averaged IoU thresholds of 0.5:0.05:0.95 [35]. We report the mean and deviation of three random rounds for all results.

Method	AP50	SO/GAIN	AP	SO/GAIN
Pre+FT	$43.2_{\pm 0.8}$	30.4/13.2	$21.0_{\pm 0.5}$	13.0/8.0
SimRoD [46]	$42.8_{\pm 1.0}$	30.4/12.4	$19.5_{\pm 0.7}$	13.0/6.5
AcroFOD [16]	$45.8_{\pm 0.6}$	30.4/15.4	$22.5_{\pm 0.4}$	13.0/9.5
AsyFOD	**47.4**$_{\pm 0.7}$	30.4/**17.0**	**23.1**$_{\pm 0.5}$	13.0/**10.1**

Table 3. Results (%) on ViPeD → COCO with YOLOv5 X.

4.2. Comparisons with State-of-the-arts

Results on Cityscapes→ Foggy Cityscapes. As summarized in Table 1, our AsyFOD performs better than other compared FDA methods in almost all categories. Besides, the AsyFOD obtains 44.3% in terms of mAP50, which is 3.2% higher than previous state-of-the-art AcroFOD [16] by a large margin. Figure 3 presents some detection results. It can be observed that the AsyFOD helps the detector to detect more targets accurately.

Results on other three scenarios. As shown in Table 2, the proposed AsyFOD performs better than previous methods on synthetic-to-real and cross-camera scenarios. In Table 3, our AsyFOD outperforms the pretrain-then-finetune paradigm (Pre+FT) about 3.8% AP50 and 2.1% AP on ViPeD → COCO. In summary, the performance improvements shown by our model across three scenarios demonstrate that our AsyFOD alleviates the domain discrepancy with very limited target data.

4.3. Ablation Studies

In this part, we conduct ablation studies on Sim10K → Cityscapes and Cityscapes → Foggy Cityscapes with YOLOv5 X.

Stop-gradient operation. To verify the effectiveness of the proposed asynchronous alignment, we compare Eq. (7) and

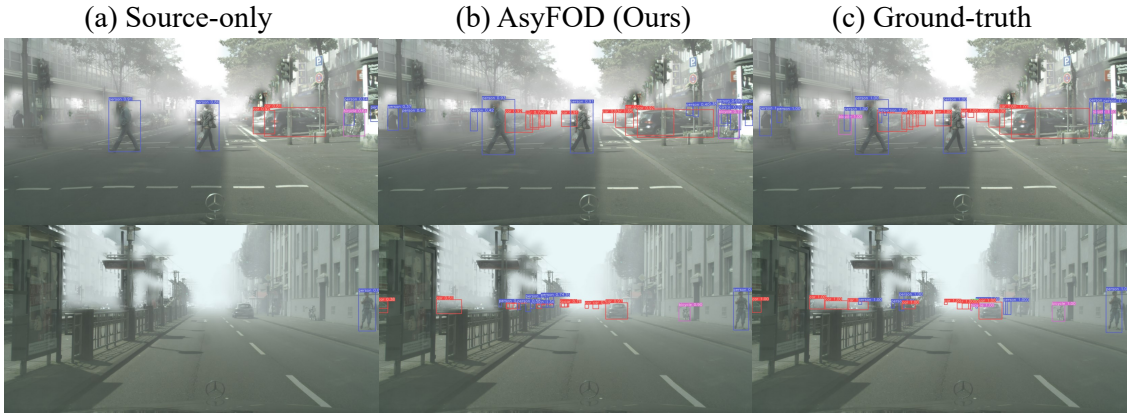

Figure 3. Qualitative results on Cityscapes → Foggy Cityscapes adaptation scenario among (a) the source-only model, (b) the proposed AsyFOD, and (c) Ground-truth. Please zoom in and view it in color.

Method	FSDAOD		FSDAOD*		FSUDAOD*	
	AP50	AP	AP50	AP	AP50	AP
Baseline	62.5	38.1	54.1	27.7	44.1	21.7
$\mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s, \mathbf{o}^t)$	62.2	37.9	55.5	29.1	48.6	24.4
$\mathcal{J}_{\mathbf{Ali}}(sg(\mathbf{o}^s), \mathbf{o}^t)$	61.7	37.4	54.8	28.3	48.0	23.6
$\mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s, sg(\mathbf{o}^t))$	**63.4**	**38.6**	**56.3**	**29.8**	**49.5**	**24.9**

Table 4. Analysis of the asynchronous alignment on Sim10K → Cityscapes in terms of AP50 and AP (%). * denotes results without strong data augmentation.

$(\mathbf{o}^s, \mathbf{o}^t)$	$(\mathbf{o}^s_{s'}, \mathbf{o}^s_{t'} \cup \mathbf{o}^t)$	$(\mathbf{o}^s_{s'}, \mathbf{o}^t)$	$(\mathbf{o}^s_{t'}, \mathbf{o}^t)$	$(\mathbf{o}^s_{s'}, \mathbf{o}^s_{t'})$
$56.3_{\pm 1.1}$	**57.4**$_{\pm 0.5}$	$56.9_{\pm 0.6}$	$56.6_{\pm 0.4}$	$55.9_{\pm 0.4}$

Table 5. Ablation study of the source division on Sim10K → Cityscapes in terms of AP50 (%). $(\mathbf{o}^s, \mathbf{o}^t)$ means $\mathcal{L}^{'}_{\mathbf{Asy}} = \mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s, sg(\mathbf{o}^t))$.

Eq. (8) on various settings in Table 4, including the default FSDAOD setting, FSDAOD setting without strong augmentation and an unsupervised setting where only 8 *unlabeled* target images available (FSUDAOD). The results are evaluated without the proposed source instances division and task-oriented supervised training. Our proposed asynchronous alignment obtains consistent improvements on all of the settings. With the strong augmentation, the traditional synchronous alignment ($\mathcal{J}_{\mathbf{Ali}}(\mathbf{o}^s, \mathbf{o}^t)$) causes worse results than the baseline. Also, the detector suffers from an obvious performance drop if stopping the gradient of $\mathbf{o}^s$, due to the effect of the imbalanced data distributions. We use stop-gradient operation by default in the following experiments.

Discrepancy estimation function and source instance division for $\mathcal{L}^{'}_{\mathbf{Asy}}$. We analyze the effect of source instance division in Table 5. Augmenting the target instance set $\mathbf{o}^t$ with the target-similar source instance set $\mathbf{o}^s_{t'}$ for asynchronous alignment performs best. It demonstrates that appropriately utilizing source instances can help the detector generalize better.

Scenario	Baseline	L_2 distance	K-means	GMM
S → C*	$56.3_{\pm 1.1}$	$57.1_{\pm 0.6}$	**57.4**$_{\pm 0.3}$	$57.0_{\pm 0.4}$
S → C	$62.5_{\pm 1.6}$	$64.5_{\pm 0.7}$	**65.4**$_{\pm 0.9}$	$65.1_{\pm 1.1}$
C → F	$41.1_{\pm 0.8}$	$43.0_{\pm 0.9}$	$43.6_{\pm 0.7}$	**44.3**$_{\pm 1.0}$

Table 6. Comparison results (%) on Sim10K → Cityscapes (S → C) and Cityscapes → Foggy Cityscapes (C → F) with different estimation functions. * denotes results without strong data augmentation.

Type	$\mathcal{L}_{\mathbf{Reg}}$	$\mathcal{L}_{\mathbf{Cls}}$	AP50	AP
Source-only	$\mathcal{D}^s$	$\mathcal{D}^s$	49.0	26.5
Baseline	$\mathcal{D}^s \bigcup \mathcal{D}^t$	$\mathcal{D}^s \bigcup \mathcal{D}^t$	$62.5_{\pm 1.6}$	$38.1_{\pm 1.8}$
Strategy A (Ours)	$\mathcal{D}^s \bigcup \mathcal{D}^t$	$\mathcal{D}^s_{\text{tar}} \bigcup \mathcal{D}^t$	**64.2**$_{\pm 1.1}$	**39.3**$_{\pm 0.8}$
Strategy B	$\mathcal{D}^s_{\text{tar}} \bigcup \mathcal{D}^t$	$\mathcal{D}^s_{\text{tar}} \bigcup \mathcal{D}^t$	$63.4_{\pm 0.9}$	$38.7_{\pm 0.7}$

Table 7. Ablation study of the task-oriented supervised training on Sim10K → Cityscapes in terms of AP50 or AP (%). The $\mathcal{D}^s$, $\mathcal{D}^t$ and $\mathcal{D}^s_{\text{tar}}$ denote using source instances, target instances, and target-similar source instances, respectively.

In Table 6, we show the results of different discrepancy estimation functions. We find that simply using L_2 distance obtains improvement over the baseline while the K-means or GMM usually obtains further improvement. On the complex scenario Cityscapes → Foggy Cityscapes, the GMM gets the best results. We notice that the strong augmentation [16] may produce target instances in the training batch of the source dataset, which affects the analysis of the division process. Therefore, we also compare the estimation strategies without strong augmentation on Sim10K → Cityscapes.

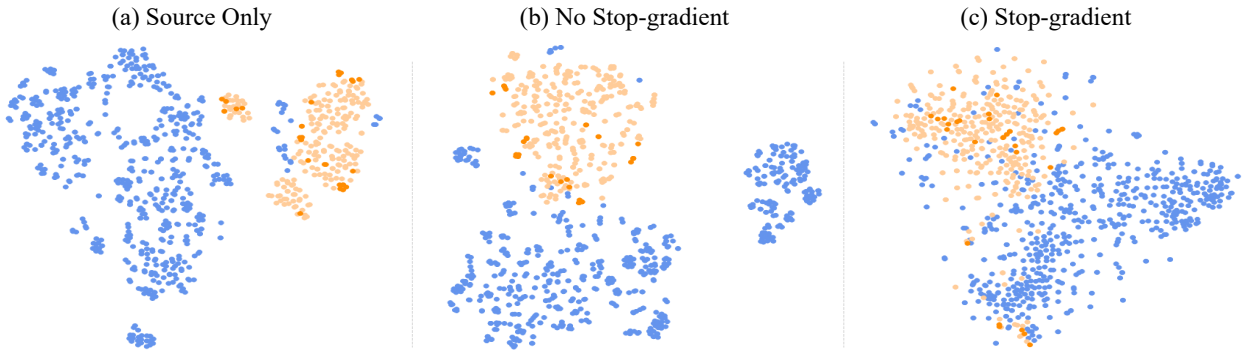

Figure 4. Instance-level feature visualization by t-SNE on the Sim10K→Cityscapes scenario (Blue: Source instances; Deep orange: Observed target instances; Light orange: Unobserved target instances): (a) Features of source-only detector; (b) Features of the proposed AsyFOD *with* the stop-gradient operation; (c) Features of the proposed AsyFOD *without* the stop-gradient operation. Please zoom in and view it in color.

Task-oriented supervised training. As shown in Table 7, the source-dissimilar instance set $\mathbf{o}_{s'}^{s}$ is crucial for localization regression loss $\mathcal{L}_{\mathbf{Reg}}$, while causing performance drop in $\mathcal{L}_{\mathbf{Cls}}$. Therefore, it is necessary to consider the relationship between source instances and limited target instances on $\mathcal{L}_{\mathbf{Reg}}$ and $\mathcal{L}_{\mathbf{Cls}}$ separately.

4.4. Qualitative Analysis

Feature distribution visualization. We visualize by t-SNE [59] in Figures 4 (a)-(c) the instance-level representations of the source-only detector and the AsyFOD with/without the stop-gradient operation. In Figure 4 (a), some source instances are very close to some target instances, partially representing the target distribution, and such observation verifies our motivation. If conducting synchronous alignment between the limited observed target instances (deep orange points) and adequate source instances (blue points), many observed target instances would be pushed closer to the source instances. In contrast, the unobserved target instances (light orange points) are not aligned. In contrast, as shown in Figure 4 (c), the unobserved target instances are better aligned than those in a traditional alignment, which is attributed to our proposed asynchronous alignment method with the stop-gradient operation.

Visualization of target-similar source instances. For each target instance, we retrieve the top-3 nearest source instances as target-similar instances by L_2 distance, whose results are shown in Figure 5. We find that the retrieved target-similar source instances (dark blue rectangles) are relatively blurred compared with those target-dissimilar source instances (light blue rectangles), which is similar to the blur caused by fog in the target domain. The results verify our assumption that some source instances are similar to target instances, which are used for augmenting the data-scarce target domain in our method.

Figure 5. Visualization of the top-3 nearest target-similar source instances for the corresponding target instance on Cityscapes → Foggy Cityscapes. Also, we show target-dissimilar source instances for comparison. Best viewed in color.

5. Conclusion

This paper presents the AsyFOD, an asymmetric adaptation paradigm for alleviating the over-adaptation problem due to the imbalanced data distribution on Few-Shot Domain Adaptive Object Detection (FSDAOD). Extensive experiments verify the effectiveness of the AsyFOD in mitigating the domain discrepancy with only a few labeled target images. We find insights into how the proposed asymmetric adaptation paradigm works in the FSDAOD task through ablation studies and visualizations. We hope the study will further inspire the community to address the FSDAOD problem.

Acknowledgement

This work was supported partially by the NSFC (U21A20471, U1911401, U1811461), Guangdong NSF Project (No. 2023B1515040025, 2020B1515120085). We also thank Jia-Chang Feng for his helpful comments.

References

[1] Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Nicola Messina. Learning pedestrian detection from virtual worlds. In *ICIAP*, 2019. 2, 6

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[5] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, 2021. 4

[6] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, 2020. 2

[7] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, 2022. 2

[8] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018. 2, 6

[9] 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 *CVPR*, 2016. 2, 5

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977. 4

[11] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2021. 2

[12] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *CVPR*, 2022. 2

[13] Glenn Jocher et al. ultralytics/yolov5: v3.0 - third release. In *Zenodo*, December 2020. 3, 6

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 2

[16] Yipeng Gao, Lingxiao Yang, Yunmu Huang, Song Xie, Shiyong Li, and Wei-Shi Zheng. Acrofod: An adaptive method for cross-domain few-shot object detection. *ECCV*, 2022. 1, 2, 6, 7

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 6

[18] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 1, 3

[19] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011. 6

[20] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012. 2, 4

[21] Mengzhe He, Yali Wang, Jiaxi Wu, Yiru Wang, Hanqing Li, Bo Li, Weihao Gan, Wei Wu, and Yu Qiao. Cross domain object detection by target-perceived dual branch distillation. In *CVPR*, 2022. 2

[22] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019. 2

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6

[24] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. In *ICLR*, 2022. 2

[25] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, 2020. 2

[26] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *ICRA*, 2017. 2, 6

[27] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *ICCV*, 2019. 2

[28] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *CVPR*, 2019. 2

[29] Junjie Li, Yichao Yan, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. Domain adaptive person search. In *ECCV*, 2022. 1

[30] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *AAAI*, 2021. 2

[31] Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *CVPR*, 2022. 2

[32] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, 2022. 6

[33] 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 *CVPR*, 2022. 2

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 6

[36] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *AAAI*, 2022. 1

[37] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *CVPR*, 2022. 1

[38] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 1, 2, 3, 4

[39] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 2

[40] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, 2016. 1, 2, 3

[41] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017. 1, 2, 3

[42] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022. 1

[43] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *CVPR*, 2022. 1

[44] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *NeurIPS*, 2017. 2

[45] Muhammad Akhtar Munir, Muhammad Haris Khan, M Sarfraz, and Mohsen Ali. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. In *NeurIPS*, 2021. 6

[46] Rindra Ramamonjison, Amin Banitalebi-Dehkordi, Xinyu Kang, Xiaolong Bai, and Yong Zhang. Simrod: A simple adaptation method for robust object detection. In *ICCV*, 2021. 2, 6

[47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1, 3

[48] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 1, 3

[49] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 3

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 3

[51] Farzaneh Rezaeianaran, Rakshith Shetty, Rahaf Aljundi, Daniel Olmeda Reino, Shanshan Zhang, and Bernt Schiele. Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection. In *ICCV*, 2021. 6

[52] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[53] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019. 2, 6

[54] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018. 2, 5

[55] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *CVPR*, 2020. 5

[56] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NeurIPS*, 2007. 1, 2, 3

[57] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1

[58] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 2

[59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[60] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *TIP*, 2021. 2

[61] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *CVPR*, 2019. 1, 2, 4, 6

[62] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020. 6

[63] Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *CVPR*, 2022. 1

[64] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *ICCV*, 2021. 2

[65] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 2020. 2

[66] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *CVPR*, 2022. 1

[67] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021. 1

[68] Jayeon Yoo, Inseop Chung, and Nojun Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, 2022. 1

[69] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of deep vision-based autonomous driving systems: Review and challenges. *IJCV*, 2022. 1

[70] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 1

[71] Liang Zhao and Limin Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, 2022. 2

[72] Chaoliang Zhong, Jie Wang, Cheng Feng, Ying Zhang, Jun Sun, and Yasuto Yokota. Pica: Point-wise instance and centroid alignment based few-shot domain adaptive object detection with loose annotations. In *WACV*, 2022. 1, 2, 4, 6

[73] Wenzhang Zhou, Dawei Du, Libo Zhang, Tiejian Luo, and Yanjun Wu. Multi-granularity alignment domain adaptation for object detection. In *CVPR*, 2022. 2

[74] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[75] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *CVPR*, 2019. 2