DSD-DA: Distillation-based Source Debiasing for Domain Adaptive Object Detection

Yongchao Feng [1] **Shiwei Li** [2] **Yingjie Gao** [1] **Ziyue Huang** [1] **Yanan Zhang** [1] **Qingjie Liu** [1 3 2] **Yunhong Wang** [1 2]

Abstract

Though feature-alignment based Domain Adaptive Object Detection (DAOD) methods have achieved remarkable progress, they ignore the source bias issue, *i.e.,* the detector tends to acquire more source-specific knowledge, impeding its generalization capabilities in the target domain. Furthermore, these methods face a more formidable challenge in achieving consistent classification and localization in the target domain compared to the source domain. To overcome these challenges, we propose a novel Distillation-based Source Debiasing (DSD) framework for DAOD, which can distill domain-agnostic knowledge from a pre-trained teacher model, improving the detector's performance on both domains. In addition, we design a Target-Relevant Object Localization Network (TROLN), which can mine target-related localization information from source and target-style mixed data. Accordingly, we present a Domain-aware Consistency Enhancing (DCE) strategy, in which these information are formulated into a new localization representation to further refine classification scores in the testing stage, achieving a harmonization between classification and localization. Extensive experiments have been conducted to manifest the effectiveness of this method, which consistently improves the strong baseline by large margins, outperforming existing alignment-based works.

1. Introduction

State-of-the-art object detectors (Redmon & Farhadi, 2018; Ren et al., 2015; Tian et al., 2019) have demonstrated impressive performance when the training and testing data exhibit consistent distributions. However, their performance diminishes drastically when applied to novel domains, primarily due to domain shift (Chen et al., 2018), which impedes the generalization and transferability of the detectors across different scenes. The inability of object detectors to adapt to novel domains hampers their practical applicability in real-world scenarios.

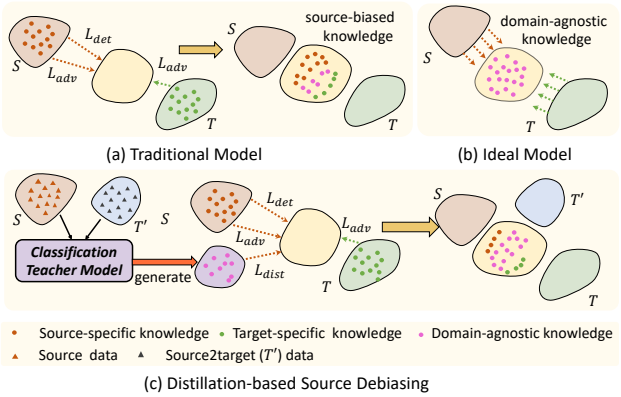

Figure 1. In (a) traditional alignment approaches, the detector tends to learn more source-specific knowledge due to the supervised detection loss L_{det}, rather than domain-agnostic knowledge (b). In (c) our proposed DSD framework, by introducing a distillation loss to the source data, the model can acquire more domain-agnostic knowledge than (a). The red and green dotted arrows represent the impact of source or target-related losses on knowledge transfer, respectively.

Extensive researches have been dedicated to address the challenge via Unsupervised Domain Adaption (UDA) methods, which aim to adapt to unlabeled target domain using the annotated source domains. One of the fashionable frameworks of UDA is to align the feature distributions between the source and target domains toward a cross-domain feature space. Early researches (Chen et al., 2018; Saito et al., 2019; Hsu et al., 2020; Jiang et al., 2021) aim to align image-level and instance-level features and achieve great margins over plain detectors. Recent works (Tian et al., 2021; Zhang et al., 2021b; Li et al., 2022b) devote to aligning class-conditional

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. [2]Hangzhou Innovation Institute, Beihang University, Hangzhou, China. [3]Zhongguancun Laboratory, Beijing, China. Correspondence to: Qingjie Liu <qingjie.liu@buaa.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

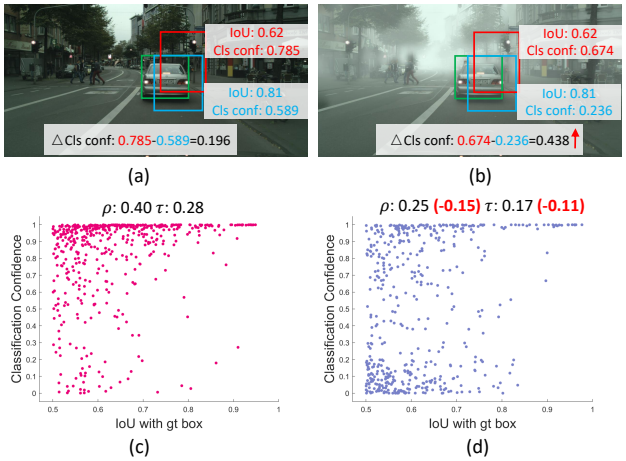

Figure 2. Demonstrative cases of the **exacerbated inconsistency** between classification and localization based on alignment-based detector DA-Faster (Chen et al., 2018). The upper row figures (a) and (b) show DA-Faster's detection results on *Cityscapes* and *FoggyCityscapes*, respectively. The lower row figures (c) and (d) display the correlation between localization ground-truth (x-axis, represented by the the IoU between the bounding box and its matched ground-truth) and classification scores (y-axis) for 500 randomly sampled DA-Faster's detected boxes in *Cityscapes* and *FoggyCityscapes*, respectively. The bounding boxes are filtered based on an IoU (≥ 0.5) with the corresponding ground-truth.

distribution across different domains, and have achieved fine-grained adaption in a category-wise manner. These approaches expect the model supervised by the labeled source domain to infer on the target domain effectively.

Despite of great success, there are still two challenges in existing alignment-based methods (Chen et al., 2018; Zheng et al., 2020; Li et al., 2022c). On one hand, as illustrated in Fig. 1(a), the supervision of the model originates from two aspects: 1) supervised detection loss L_{det} from source; 2) adversarial loss L_{adv} from both source and target. Compared with L_{adv}, L_{det} provides more explicit supervision signal and enforce the model to fit the source distribution. Thus, in this process, it is inevitable that detector will acquire more source-specific knowledge than target-specific knowledge. Simultaneously, the detector also gain domain-agnostic knowledge in the adversarial training process. As a result, the knowledge acquired by the detector is source-biased (*i.e.,* more source-specific knowledge) rather than ideal (Fig. 1(b)), hindering the model's generalization in the target domain. These observations motivate us to design a new paradigm that enables the model to learn more domain-agnostic knowledge compared to traditional methods.

On the other hand, we observed that the alignment-based detectors (e.g. DA-Faster (Chen et al., 2018)) face **exacerbated inconsistency** between classification and localization. Firstly, as shown in Fig. 2(a), inconsistency means the phenomenon (Jiang et al., 2018) that compared with the detected bounding box (blue), another detected bounding box (red) with higher classification scores could have lower IoU with ground truth boxes (green). **Exacerbated inconsistency** in this paper means that existing alignment-based detectors encounter more severe inconsistency issues on the target domain than that of source domain. From the perspective of detection visualization, the exacerbated inconsistency is reflected in that detection boxes (blue and red) in *FoggyCityscapes* (Fig. 2(b)) often exhibit larger differences in classification scores compared with ones located in the same position in *Cityscapes* (Fig. 2(a)). From the perspective of correlation metric, the exacerbated inconsistency means Spearman Rank Correlation Coefficient ρ and Kendall Rank Correlation Coefficient τ in the *FoggyCityscapes* (Fig. 2(d)) are lower than ones in *Cityscapes* (Fig. 2(c)). ρ and τ are the measures of rank correlation: the similarity of the ordings of the data when ranked by each of the quantities. As shown in Fig. 2(c) (d), the ρ and τ in the *FoggyCityscapes* are relatively lower (-0.15 and -0.11) compared to *Cityscapes*, indicating that the detector faces a more pronounced inconsistency issue in the target domain. In general detection pipelines, the classification scores are commonly employed as the metric for ranking the detected boxes, which can result in the suppression of accurate bounding boxes by less accurate ones during NMS procedure. In the cross-domain detection, this issue is amplified due to more severe inconsistency.

In this paper, to overcome the aforementioned constraints, we present a DSD-DA method that contains a novel Distillation-based Source Debiasing (DSD) framework and a Domain-aware Consistency Enhancing (DCE) strategy. Firstly, considering that current alignment-based DAOD methods tends to acquire more source-specific knowledge, we propose to perform knowledge distillation to guide detector to acquire more domain-agnostic knowledge and thus restrain the potential source bias issue. Specifically, as shown in Fig. 1(c), we first utilize CycleGAN (Zhu et al., 2017) to transform the source images into the target domain style, named source2target domain T' (light blue). Then, we train a classification-teacher via a supervised and adversarial loss, which is able to learn domain-agnostic knowledge from S and T' mix-style data. Finally, these knowledge is utilized to guide the detector's training, improving the performance of detector on both domains. Secondly, to mitigate exacerbated inconsistency on target data, we design a Target-Relevant Object Localization Network (TROLN), in which pixel and instance-level target affinity weights are proposed and incorporated into the loss function. TROLN is also trained on S and T' mix-style data, which is utilized to mine target relevant localization information. Then, in the testing stage, we conduct the DCE strategy, *i.e.,* formulating

the output of TROLN into a new localization representation. And we use this representation to adjust the classification scores of the detected boxes, enhancing the consistency and making sure that more accurate detected boxes are retained in NMS process.

In summary, our contributions are as follows:

- We propose a novel DSD framework for DAOD, which utilizes an unbiased classification-teacher to guide the detector to learn more domain-agnostic feature representations. To the best of our knowledge, this is the first study to analyze and solve source bias issue in alignment-based methods.
- We reveal the **exacerbated inconsistency** issue between the classification and localization existing in traditional alignment-based method. We design TROLN and conduct DCE strategy to refine classification scores in the testing stage, which enhances the consistency between the classification and localization.
- Extensive experiments demonstrate that our method consistently outperforms the strong baseline by significant margins, highlighting its superiority compared to existing alignment-based methods.

2. Related Work

Domain adaptation for object detection. Several approaches have been proposed for DAOD, which can be categorized into alignment-based (Chen et al., 2018; Saito et al., 2019; Li et al., 2022b;a; Xu et al., 2022) and self-training (Deng et al., 2021; Ramamonjison et al., 2021; Deng et al., 2023) methods. However, regardless of various technological approaches, the source bias issue persists. Self-training methods alleviate the detector's bias towards the source domain by continuously improving the quality of pseudo-labels during training stage. The development of alignment-based methods often involves modeling features from coarse to fine. DA-Faster (Chen et al., 2018) implements feature alignment at both the image-level and instance-level. H^2FA (Xu et al., 2022) enforces two image-level alignments for the backbone features, as well as two instance-level alignments for the RPN and detection head. SIGMA (Li et al., 2022b) constructs the feature distributions of the source and target domains as graphs and reformulates the adaption with graph matching. However, regardless of the granularity of modeling, it is unable to change the asymmetry in the losses in the alignment stage and thus cannot effectively address the source bias problem. In this paper, we propose a novel distillation-based alignment (DSD) framework, where a distillation loss is constructed to guide the learning process of detector. To the best of our knowledge, this is the first method directly optimizing source bias in alignment-based approaches.

Representation of localization quality. The conflict between classification and localization tasks is a well-known problem (Jiang et al., 2018; Tychsen-Smith & Petersson, 2018; Li et al., 2020; Zhang et al., 2021a; Pu et al., 2023; Zhang et al., 2023; Feng et al., 2021) in the object detection field. Existing works have focused on finding more accurate localization representation to guide the learning of the classification head, addressing the inconsistency issue. IoU-Net (Jiang et al., 2018) introduces an extra head to predict IoU and use it to rank bounding boxes in NMS. Fitness NMS (Tychsen-Smith & Petersson, 2018) and IoU-aware RetinaNet (Wu et al., 2020) multiply the predicted IoU or IoU-based ranking scores by the classification score as the ranking basis. Instead of predicting the IoU-based score, FCOS (Tian et al., 2019) predicts centerness scores to suppress the low-quality detections. However, for exacerbated inconsistency in cross-domain scenarios, how to incorporate target-relevant information into localization representation designing has become a new challenge. Unfortunately, few methods explore and solve this challenge. Therefore, we first propose a target-relevant OLN to mine target-related localization information from style-mixed data. Then we integrate these target relevant information into a novel localization representation to refine the classification scores, enhancing the consistency.

3. Method

3.1. Problem Formulation

In the cross-domain object detection, we have a labeled source domain $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s and $y_i^s = (b_i^s, c_i^s)$ denote the i_{th} image and its corresponding labels, *i.e.*, the coordinates of the bounding box b and its associated category c, respectively. In addition, we have access to an unlabeled target domain $\mathcal{D}_T = \{x_i^t\}_{i=1}^{N_t}$. In this work, we employ CycleGAN (Zhu et al., 2017) to convert the source images into the target domain style, creating a new domain named source2target domain $\mathcal{D}_{T'} = \{(x_i^{t'}, y_i^{t'})\}_{i=1}^{N_s}$, which shares labels with the source domain data. We assume that the source and target samples come from different distributions (*i.e.*, $\mathcal{D}_S \neq \mathcal{D}_T$) but the categories are exactly the same. The objective is to enhance the performance of the detector in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$.

3.2. Framework Overview

As shown in Fig. 3, our method involves two training stages, a teacher models training stage and a detector training stage. In the teacher models training stage (Sec 3.4), we train a classification and a localization models as teachers using the labeled data $\mathcal{D}_S$ and $\mathcal{D}_{T'}$. In the second training stages (Sec 3.5), the features of the positive proposals are expected to derive domain-agnostic representation from the classification-teacher model. During the testing stage

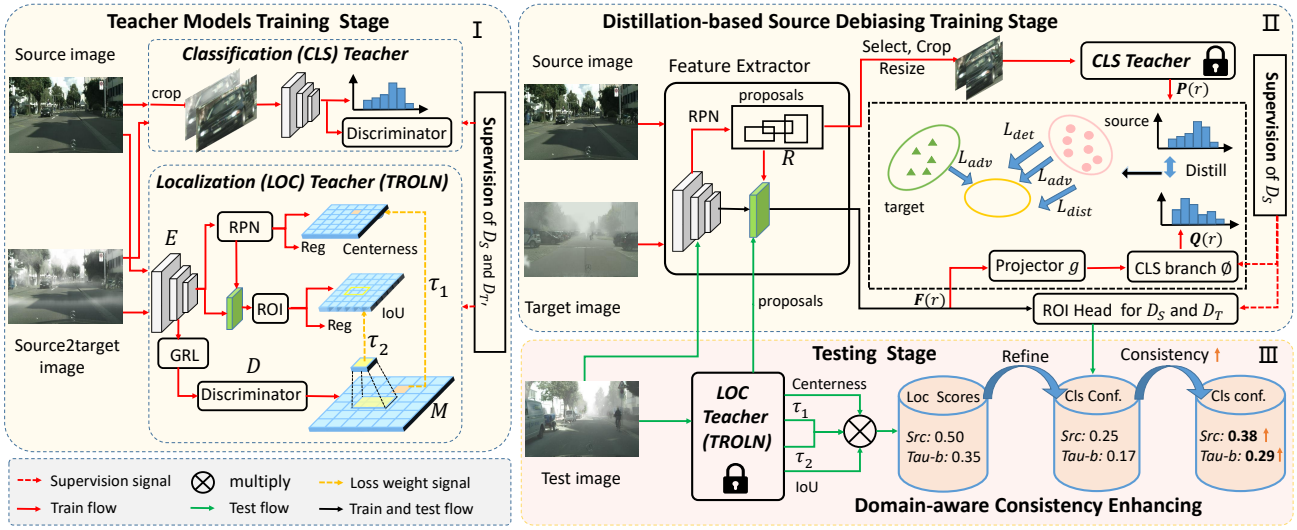

Figure 3. Overview of the proposed distillation-based source debiasing framework for DAOD. Part I shows the teacher models training stage, which includes a mix-style classifier and a Target-Relevant object localization network (TROLN) training. Part II demonstrates distillation-based source debiasing (DSD) training, in which the cross-domain detector is trained. In Part III, the Domain-aware Consistency Enhancement (DCE) strategy is introduced to refine the detector's classification scores in the testing phase, enhancing the consistency between classification and localization.

(Sec 3.6), we design the localization scores based on the output of TROLN and use it to refine the classification scores, thereby alleviating the inconsistent issue.

3.3. Detection Baseline

We use DA-Faster (Chen et al., 2018) as our base detector. DA-Faster is a two-stage cross-domain detector that consists of two major components: a standard Faster-RCNN detector and a domain adaption component that includes image-level and instance-level domain discriminators D. When the training process gradually converges, the detector tends to extract domain-invariant feature representations. Formally, the image-level adaption loss can be written as:

$$\mathcal{L}_{img} = \min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x_s \sim \mathcal{D}_S} \log D_{img}(G(x_s)) + \mathbb{E}_{x_t \sim \mathcal{D}_T} \log(1 - D_{img}(G(x_t))) \quad (1)$$

where θ_G and θ_D are the parameters of backbone G and D_{img}, x_s, x_t represent images from $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. Similarly, the instance-level adaption loss is defined as:

$$\mathcal{L}_{ins} = \min_{\theta_{Det}} \max_{\theta_D} \mathbb{E}_{f_s \sim ROI_S} \log D_{ins}(f_s) + \mathbb{E}_{f_t \sim ROI_T} \log(1 - D_{ins}(f_t)) \quad (2)$$

where θ_{Det} and θ_D are the parameters of detector and D_{ins}, f_s, f_t represent the ROI features of x_s and x_t, respectively. The training loss of DA-Faster is a summation of each individual part, which can be written as:

$$\mathcal{L}_{DA} = L_{det} + \lambda(L_{img} + L_{ins}) \quad (3)$$

where $\mathcal{L}_{det}$ is the loss of Faster-RCNN and λ is a trade-off parameter to balance the Faster-RCNN loss and domain adaption loss.

3.4. Teacher Models Training

Classification Teacher. We first construct an instance-level image dataset $\mathcal{D}$ as image corpus by extracting all class objects from the detection dataset $\mathcal{D}_S$ and $\mathcal{D}_{T'}$ according to their ground-truth bounding boxes and labels. Formally, for an input image, we first perform typical data augmentations (random cropping, color distortion, *etc.*). Then we feed the augmented image into a ResNet (He et al., 2016) classification network for supervised learning. Simultaneously, we adopt domain discriminator to align the feature distribution of S and T'.

The classifier's ability to acquire domain-agnostic knowledge can be attributed to three aspects: 1) Strong supervision signal. Since $\mathcal{D}$ contains labeled images with two different styles, classifier is enforced to fit the data distribution of different domain via the supervised loss during the training. In this process, classifier tends to acquire domain-agnostic knowledge. 2) Adversarial learning. Domain discriminator with Gradient Reverse Layer (GRL) (Ganin & Lempitsky, 2015) layer is powerful tool that can align feature distributions between two domains. It is beneficial for feature

extractor to produce domain-invariant features that cannot be discriminated by the discriminator. 3) Enriched data. Compared with sole source data, the mixed data doubles the scale and greatly enriches the training data with various data augmentations. It facilitates the classifier to learn domain-invariant representations. Our classification-teacher model is optimized in a completely independent way from the object detection. And its domain-agnostic knowledge can be transferred to object detection to suppress potential source bias.

TROLN Teacher. To solve the challenge of exacerbated inconsistency on the target data, we attempt to use localization indicators (IoU, centerness) from OLN (Kim et al., 2022) to calibrate the classification scores. The original OLN estimates the objectness of each region by centerness-head and IoU-head. The comprehensive loss function of OLN can be written as:

$$\begin{aligned}\mathcal{L}_{OLN} &= \mathcal{L}^{Cent}_{RPN} + \mathcal{L}^{reg}_{RPN} + \mathcal{L}^{IoU}_{RCNN} + \mathcal{L}^{reg}_{RCNN} \\ \mathcal{L}^{Cent}_{RPN} &= \frac{1}{N_{pix}} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathbb{1}^{pix}_{for} L_1(c_{x,y}, \hat{c}_{x,y}) \\ \mathcal{L}^{IoU}_{RCNN} &= \frac{1}{N_{pos}} \sum_{r=1}^{N_{pos}} \mathbb{1}^{pro}_{for} L_1(b_r, \hat{b}_r)\end{aligned} \tag{4}$$

where $\mathbb{1}^{pix}_{for}$ and $\mathbb{1}^{pro}_{for}$ denote the positive pixels and positive proposals set. $c_{x,y}$, b_r, $\hat{c}_{x,y}$, $\hat{b}_r$ are the predicted centerness, predicted IoU, groundtruth centerness and groundtruth IoU, respectively.

However, when using directly S and T' mixed data to train original OLN, the lack of guidance from the target domain results in less target-relevant images being given the same importance as more relevant ones, leading to a deterioration in knowledge learning and mining from the target domain. To effectively mine target relevant localization information in the training, Target-Relevant Object Localization Network (TROLN) has been developed to ensure that target-relevant information are encoded at the pixel and instance level. Specifically, a pixel-level domain discriminator D is placed after the feature encoder E (shown in Fig. 3 I) in the TROLN. The probability of each pixel belonging to the target domain is defined as $D(E(X)) \in \mathbb{R}^{H \times W \times 1}$ and $1 - D(E(X)) \in \mathbb{R}^{H \times W \times 1}$ represents the probability of it belonging to the source domain. The domain discriminator D is updated using binary cross-entropy loss based on the domain label d for each input image, where images from the source domain are labeled as $d = 0$ and images from target domain are labeled as $d = 1$. The discriminator loss $\mathcal{L}_{dis}$ can be expressed as:

$$\mathcal{L}_{dis} = -d \log D(E(X)) - (1 - d) \log(1 - D(E(X))) \tag{5}$$

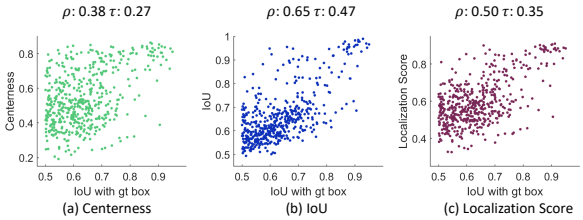

Figure 4. The correlation between localization ground-truth and centerness/IoU/localization scores of the bounding boxes on the target test dataset. The bounding boxes are filtered based on an IoU (≥ 0.5) with the corresponding ground-truth.

The large value within $D(E(X))$ indicates that the distribution of current pixel and target pixels are more similar. Based on the important cues, we denote $D(E(X))$ as target affinity score map M and adopt dynamic target-related weight to adjust the L^{Cent}_{RPN} and L^{IoU}_{RCNN}. The pixel-level and instance-level target affinity weight τ_1, τ_2 are defined as the following:

$$\begin{aligned}\tau_1 &= M(x, y) \\ \tau_2 &= \text{Average}(\text{ROIAlign}(M, p))\end{aligned} \tag{6}$$

where (x, y) represents the pixel coordinates of M and p denotes one proposal from RPN.

Subsequently, we can reweight the importance of loss items from pixel and instance level as illustrated in Fig. 3 I, and apply it to train a localization-teacher (TROLN) by reformulating the loss function in Eq. 4 as the following:

$$\begin{aligned}\mathcal{L}_{TROLN} &= \mathcal{L}^{Cent}_{RPN} + \mathcal{L}^{reg}_{RPN} + \mathcal{L}^{IoU}_{RCNN} + \mathcal{L}^{reg}_{RCNN} + \mathcal{L}_{dis} \\ \mathcal{L}^{Cent}_{RPN} &= \frac{1}{N_{pix}} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathbb{1}^{pix}_{for} (\tau_1 + 1) L_1(c_{x,y}, \hat{c}_{x,y}) \\ \mathcal{L}^{IoU}_{RCNN} &= \frac{1}{N_{pos}} \sum_{r=1}^{N_{pos}} \mathbb{1}^{pro}_{for} (\tau_2 + 1) L_1(b_r, \hat{b}_r)\end{aligned} \tag{7}$$

Based on Eq. 7, TROLN is explicitly enforced to learn from target-relevant samples, and thus prevents the interference from the information irrelevant to the target.

3.5. Distillation-based Source Debiasing

After Teacher Models Training stage, we start the training of cross-domain detector. Here, we take DA-Faster (Chen et al., 2018) as the base detector to describe the DSD framework. As shown in Fig. 3 II, in the DSD training stage, given source proposals R generated by RPN, we initially select R by assessing IoU with ground truth higher than threshold T. We crop these proposals from the source image and resize them to a fixed size using bilinear interpolation, then feed them into our classification-teacher model to obtain

classification logit $\mathbf{P}(r) \in \mathbb{R}^{K\times 1}$. Here, K represents the number of classes in the object detection task, r represents one of filtered proposals. Meanwhile, we obtain the ROI features $\mathbf{F}(r)$ for r from the RoI Align layer. Note that $\mathbf{F}(r)$ and $\mathbf{P}(r)$ are learned in the different feature space, thus we first project $\mathbf{F}(r)$ into the same feature space as $\mathbf{P}(r)$ and then obtain the classification logit of the projected feature:

$$\mathbf{Q}(r) = \phi(g(\mathbf{F}(r)) \tag{8}$$

Here, $g(\cdot)$ denotes the project function for features $\mathbf{F}(r)$, which is implemented with a 1×1 convolutional layer, while $\phi(\cdot)$ is the classification branch in the detection head. Then we minimize the L1-norm between these two logit representations to guide the learning process of the detector:

$$L_{\text{dist}} = \frac{1}{R_f K} \sum_{r=1}^{R_f} \sum_{k=1}^{K} \|\mathbf{P}_k(r) - \mathbf{Q}_k(r)\|_1 \tag{9}$$

where R_f is the number of filtered proposals. The obtained logit representation $\mathbf{Q}(r)$ can also be utilized for classification of the region proposal r. Thus, we conduct an auxiliary classification task on the logit $\mathbf{Q}(r)$.

$$\begin{aligned} p' &= \mathcal{F}_{\text{softmax}}(\mathbf{Q}(r)) \\ \mathcal{L}_{\text{cls-aux}} &= \text{CE}(y, p'), \end{aligned} \tag{10}$$

where p' is the predicted scores based on $\mathbf{Q}(r)$, y is the groundtruth label for the region proposal r. Note that the whole distillation process is only conducted on the source images.

Consequently, the object detector is trained under the supervision of the three losses jointly:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{DA}} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{cls-aux}}, \tag{11}$$

where $\mathcal{L}_{\text{DA}}$ denotes the loss of DA-Faster.

3.6. Domain-aware Consistency Enhancing

To address more severe inconsistency between classification and localization, we design a novel localization representation to refine classification scores, enhancing the consistency in the testing stage. Since the objective of DAOD is to enhance the performance of the detector in target domain and TROLN has already explored the target-relevant localization information, we first investigate the relation between localization ground-truths and two localization indicators (IoU and centerness) derived from TROLN.

In the TROLN framework, each detected bounding box originates from two refinements of the corresponding anchor, *i.e.,* anchor $\rightarrow$ proposal $\rightarrow$ detected box. Here, we assume that these three share the same centerness and IoU. Similar to Fig. 2, we evaluate the trained TROLN on the *FoggyCityscapes* test datasets and visualize the correlation of two indicators (y-axis, centerness and IoU) and localization ground-truth (x-axis), as shown in Fig. 4(a) and (b). It is evident that compared to the classification score in Fig. 2(b), both IoU and centerness exhibit a higher consistency with the localization ground-truths. This indicates that these two indicators have the potential to serve as localization representation for refining the classification scores.

However, on one hand, using the highest consistency indicator (IoU) to weight the classification scores could cause an extra class confusion issue. For example, if two detected boxes simultaneously match the same ground truth box with the category "cat", where the detected box *A* predicts a "cat" confidence of 0.6 and an IoU of 0.5, and detected box *B* predicts a "tiger" confidence of 0.5 and an IoU of 0.7. In this case, employing IoU would yield a "cat" confidence of 0.3 (0.6×0.5) for detected box *A* and a "tiger" confidence of 0.35 (0.5×0.7) for detected box *B*, potentially resulting in misclassification. On the other hand, refining classification score based on centerness may not sufficiently improve the consistency. In addition, although centerness and IoU have already integrate target-relevant information, they lack the capability to adaptively adjust based on current features. Therefore, we incorporate pixel and instance-level target affinity weights τ_1, τ_2 into the localization representation and strike a balance between centerness and IoU. Here, we propose a novel localization score s as follow:

$$s = \sqrt{4 \times c \times b \times \tau_1 \times \tau_2}, \tag{12}$$

where c and b denote centerness and IoU, respectively. Due to the effect of the GRL in TROLN, τ_1 and τ_2 are close to "0.5" with the training process gradually converge, the "4" in Eq. 12 is a compensation factor. From Fig. 4(c), we can observe that the consistency of s (ρ and τ) ranges between these of c and b. The rationale behind Eq. 12 is to improve the localization scores of target-style detection boxes (with larger τ_1, τ_2) and suppress source-style ones (with smaller τ_1, τ_2) during the testing stage. This operation is to adaptively retain more target-style detected boxes.

Then we use s to refine classification scores, termed as the Domain-aware Consistency Enhancing strategy. Concretely, in the testing stage, given an image I, we feed it to the TROLN to obtain a proposals set $\mathcal{R} = \{(box_i, s_i)\}_{i=1}^{N_p}$, where box_i and s_i represent the spatial coordinates and the localization score of the i-th proposal, N_p denotes the total number of proposals. Simultaneously, we feed I into the trained detector, replacing the detector's proposals with $\mathcal{R}$. As a result, we obtain the ROI head output $\mathcal{T} = \{(reg_i, cls_i)\}_{i=1}^{N_p}$, where reg_i and cls_i denote the regression results and classification scores respectively. After experimenting with various forms such as squaring and other transformations, we ultimately adopt Eq. 13 to refine cls_i,

Table 1. Results from ***Cityscapes→FoggyCityscapes*** based on different base detectors with various backbones.

Method	Backbone	person	rider	car	truck	bus	train	mcycle	bcycle	mAP
SWDA (Saito et al., 2019)	VGG16	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
TIA (Zhao & Wang, 2022)		34.8	46.3	49.7	31.1	52.1	48.6	37.7	38.1	42.3
TDD (He et al., 2022)		39.6	47.5	55.7	33.8	47.6	42.1	37.0	41.4	43.1
MGA (Zhou et al., 2022)		45.7	47.5	60.6	31.0	52.9	44.5	29.0	38.0	43.6
PT (Chen et al., 2022)		40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
SCAN (Li et al., 2022a)		41.7	43.9	57.3	28.7	48.6	48.7	31.0	37.3	42.1
SIGMA++ (Li et al., 2023)		46.4	45.1	61.0	32.1	52.2	44.6	34.8	39.9	44.5
HT (Deng et al., 2023)		52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
OADA (Yoo et al., 2022)		47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4
SIGMA (Li et al., 2022b)		43.9	52.7	56.8	26.2	46.2	12.4	34.8	43.0	43.5
DA-Faster(Chen et al., 2018)		43.9	52.9	56.8	26.2	46.2	12.4	34.9	43.0	39.6
DA-Faster(Chen et al., 2018) + DSD-DA		46.5	54.1	61.9	28.3	49.5	26.7	40.0	46.3	**44.2**
AT (Li et al., 2022c)		45.3	55.7	63.6	36.8	64.9	34.9	42.1	51.3	49.3
AT (Li et al., 2022c) + DSD-DA		49.1	59.3	66.2	35.8	60.0	47.1	45.2	54.9	**52.2**
CMT (Cao et al., 2023)		45.9	55.7	63.7	39.6	66.0	38.8	41.4	51.2	50.3
CMT (Cao et al., 2023) + DSD-DA		49.0	59.6	65.3	35.7	61.0	46.5	43.9	57.3	**52.3**
GPA (Xu et al., 2020b)	ResNet50	32.9	46.7	54.1	24.7	45.7	41.1	32.4	38.7	39.5
CRDA (Xu et al., 2020a)		39.9	38.1	57.3	28.7	50.7	37.2	30.2	34.2	39.5
DIDN (Lin et al., 2021)		38.3	44.4	51.8	28.7	53.3	34.7	32.4	40.4	40.5
DSS (Wang et al., 2021)		42.9	51.2	53.6	33.6	49.2	18.9	36.2	41.8	40.9
DA-Faster(Chen et al., 2018)		36.4	47.2	53.7	29.3	48.8	34.4	33.8	38.5	40.2
DA-Faster(Chen et al., 2018) + DSD-DA		43.7	49.1	60.7	30.8	55.7	43.4	33.7	44.6	**45.2**
CADA (Hsu et al., 2020)	ResNet101	41.5	43.6	57.1	29.4	44.9	39.7	29.0	36.1	40.2
D-adapt (Jiang et al., 2021)		42.8	48.4	56.8	31.5	42.8	37.4	35.2	42.4	42.2
DA-Faster(Chen et al., 2018)		37.2	45.1	54.5	30.9	48.9	43.3	29.3	39.5	41.1
DA-Faster(Chen et al., 2018) + DSD-DA		43.9	50.7	61.6	31.8	52.2	47.1	32.1	46.1	**45.7**

Table 2. Results on ***Kitti→Cityscapes*** with VGG-16. SO represents the source only results and GAIN indicates the adaption gains compared with the source only model.

Method	Car	SO/GAIN
CADA (Hsu et al., 2020)	43.2	34.4/ 8.8
MEGA (Vs et al., 2021)	43.0	30.2/ 12.8
SSAL (Munir et al., 2021)	45.6	34.9/ 10.7
KTNet (Tian et al., 2021)	45.6	34.4/ 11.2
SIGMA (Li et al., 2022b)	45.8	34.4/ 11.4
DA-Faster (Chen et al., 2018)	43.4	34.5/ 8.9
DA-Faster (Chen et al., 2018)+ DSD-DA	**46.9**	34.5/ 12.4
AT (Li et al., 2022c)	47.7	34.5/ 13.2
AT (Li et al., 2022c) + DSD-DA	**49.3**	34.5/ 14.8

and obtain the adjusted classification score $cls_i^{'}$:

$$cls_i^{'} = \mathcal{F}_{\text{softmax}} \sqrt[4]{cls_i \times s_i} \tag{13}$$

The refined output $\mathcal{T}^{'} = \{(reg_i, cls_i^{'})\}_{i=1}^{N_p}$ are used to participate NMS and evaluate the performance by following DA-Faster (Chen et al., 2018) protocol.

4. Experiments

4.1. Datasets and Implementation Details

We conduct our experiments on four datasets, including (1) ***Cityscapes*** (Cordts et al., 2016) contains authentic urban street scenes captured under normal weather conditions, encompassing 2,975 training images and 500 validation images with pixel-level annotations. (2) ***FoggyCityscapes*** (Sakaridis et al., 2018) is a derivative dataset that simulates dense foggy conditions based on Cityscapes, maintaining the same train/validation split and annotations. (3)

Table 3. Results from ***SIM10k→Cityscapes***.

Method	Car	SO/GAIN
SWDA (Inoue et al., 2018)	40.1	34.3/ 5.8
MAF (He & Zhang, 2019)	41.1	34.3/ 6.8
HTCN (Chen et al., 2020)	42.5	34.4/ 8.1
CFFA (Zheng et al., 2020)	43.8	34.3/ 9.5
ATF (He & Zhang, 2020)	42.8	34.3/ 8.5
MeGA-CDA (Vs et al., 2021)	44.8	34.3/ 10.5
UMT (Deng et al., 2021)	43.1	34.3/ 8.8
DA-Faster (Chen et al., 2018)	43.6	34.6/ 9.0
DA-Faster (Chen et al., 2018) + DSD-DA	**47.8**	34.6/ 13.2
AT (Li et al., 2022c)	51.4	34.6/ 16.8
AT (Li et al., 2022c) + DSD-DA	**52.5**	34.6/ 17.9

KITTI (Geiger et al., 2012) is one popular dataset for autonomous driving including 7,481 labeled images for training. (4) ***SIM10k*** (Johnson-Roberson et al., 2016) is a synthetic dataset containing 10,000 images rendered from the video game Grand Theft Auto V (GTA5).

We report AP_{50} of each class for object detection following (Chen et al., 2018) for all experimental setting as follows: (1) ***Cityscapes→FoggyCityscapes***. It aims to perform adaptation across different weather conditions. (2) ***Kitti→Cityscapes***. It is cross camera adaption, where the source and target domain data are captured with different camera setups. (3) ***SIM10k→Cityscapes***. To adapt the synthetic scenes to the real one, we utilize the entire ***SIM10k*** dataset as the source domain and the training set of ***Cityscapes*** as the target domain. Following (Li et al., 2023), we only report the performance on car for the last two scenarios.

We evaluate the proposed DSD-DA method (DSD + DCE)

Table 4. Detection performance on the source and target domain using different backbones. AP_s/AP_t: Detection performance (AP) on the source/target.

Method	Backbone	$AP_s\uparrow$	$AP_t\uparrow$
Source Only	VGG16	49.02	20.18
Baseline		48.91	39.56
Baseline+DSD		**50.09**	**42.00**
Source Only	ResNet50	50.12	23.92
Baseline		50.21	40.90
Baseline+DSD		**51.48**	**43.05**

on DA-Faster (Chen et al., 2018), AT (Li et al., 2022c) and CMT (Cao et al., 2023). In the DSD training stage, we resize all the cropped images to 224×224, and set IoU threshold $T = 0.8$. DA-Faster was trained with SGD optimizer with a 0.001 learning rate, 2 batch size, momentum of 0.9, and weight decay of 0.0005 for 70k iterations on 1 Nvidia GPU 2080Ti.

4.2. Main Results

Cityscapes→FoggyCityscapes. We present the comparison with VGG16, ResNet50 and ResNet101 backbones in Table 1. When base detector is DA-Faster, our method achieves 44.2%, 45.2%, and 45.7% mAP, respectively, improving mAP by 4.6%, 5.0% and 4.6% compared to (Chen et al., 2018). Simultaneously, our method has achieved consistent improvements on the state of the art such as AT (Li et al., 2022c) and CMT (Cao et al., 2023). This fully demonstrates the effectiveness of our approach and its compatibility with the different backbone networks.

Kitti→Cityscapes. In Table 2, we illustrate the performance comparison on the cross-camera task. The proposed method reaches an AP_{50} of 46.9% and 49.3% with a gain of +12.4% and +14.8% over the source only model with different base detector, respectively.

SIM10k→Cityscapes. Table 3 shows that our method consistently improves performance across different base detectors. This further illustrates that the proposed approach has strong generalization capabilities, effectively adapting from synthetic to real setting.

4.3. Analysis of Source Bias

The "source bias" refers to the model's tendency to favor the source data, even when it is trained on both the source and target data. This occurs because the source data provides stronger supervision signals, leading to an imbalance in the training process. Empirical experiments have confirmed the existence of "source bias" issue.

1) The presence of source bias is evident in the performance gap between the source and target data. For example, in Table 4, the *Baseline* detector shows significantly better performance on the source data (48.91 vs. 39.56), highlighting a clear source bias.

Table 5. Detection performance on the different test set with various fog density.

Test Set	Baseline (AP_{50})	Baseline +DSD (AP_{50})
Foggy Cityscapes (0.02, Target)	40.90	43.05
Foggy Cityscapes (0.01)	44.04	45.82
Foggy Cityscapes (0.005)	46.83	47.56
Cityscapes (Source)	50.21	51.48

Table 6. Detection performance in different similarity intervals at the object-level.

Similarity Score	0-0.6	0.6-0.75	0.75-0.85	0.85-1.0
Baseline (AP_{50})	5.7	13.6	37.9	65.0
Baseline +DSD (AP_{50})	11.3	15.5	41.6	67.7

2) Reduced bias is linked to better performance. In our experiments with Cityscapes→Foggy Cityscapes, the latter includes three levels of fog density (0.02, 0.01, 0.005), with lower values indicating thinner fog. As shown in Table 5, a closer resemblance to Cityscapes results in increasingly similar performance. (In the paper, fog level 0.02 is default as target domain). Addressing bias is possible, our proposed method effectively improves the performance of the source and target datasets.

3) We also conduct experiments on the object level. Firstly, Baseline and our detector are applied to the target images, followed by an assessment of the style similarity score of the detected boxes to the source-style. The style similarity score is measured using the trained instance-level domain discriminator. Then we evaluate the performance of the detectors across different similarity intervals. Table 6 results demonstrate that much higher performance is achieved when the detected objects are more similar to the source-style.

4.4. Ablation Study

In this section, we conduct ablation studies to validate our contributions. All experiments are conducted on the ***FoggyCityscapes*** validation set with the DA-Faster as base detector.

Effectiveness of individual component. We first investigate the impact of DSD and DCE on detection performance. As shown in Table 7, both DSD and DCE can improve the performance of the baseline under different backbone configurations. Finally, with all these components, we observe a respective enhancement in mAP of the baseline by 4.65%, 5.07%, and 4.64% when employing ResNet50, ResNet101, and VGG16 as backbones. This demonstrates the effectiveness and necessity of DSD and DCE.

The enhanced generalization of the DSD. In addition, to verify the improvement in detector's generalization brought by DSD, we evaluate the performance of the three methods on the source and target domain. As shown in Table 4, compared with the *Source Only*, *Baseline* and *Baseline+DSD*

Table 7. Ablation study on the proposed DSD and DCE.

Module		mAP		
DSD	DCE	VGG16	ResNet50	ReNet101
		39.56	40.15	41.09
✓		42.00	43.05	42.63
	✓	42.14	43.23	43.11
✓	✓	**44.21**	**45.22**	**45.73**

Table 8. Effects of distillation data for assessing DSD module in our method. S means using source data for distillation. T_{obj} and T_{cls} represent filtering strategies based on objectness and classification, respectively, and $T_{obj\&cls}$ represents a strategy that filters both objectness and classification simultaneously. Here, the filtering threshold is set to 0.8, retaining samples with scores higher than 0.8.

S	T_{obj}	T_{cls}	$T_{obj\&cls}$	AP_{50}
				40.15
✓				**43.05**
✓	✓			38.87
✓		✓		37.64
✓			✓	38.43

improve the AP_t by large margins, which demonstrates the positive impact of feature alignment. Furthermore, in comparison to *Source Only*, *Baseline* demonstrates a substantial enhancement exclusively in AP_t, whereas *Baseline+DSD* exhibits significant improvements in both AP_s and AP_t. This may be because our DSD framework distills more domain-agnostic knowledge to the detector, improving the generalization of detector on both source and target.

Source or Target. In the DSD framework, distillation loss is utilized to guide the detector to learn domain-agnostic knowledge. We conduct an ablation study on choices of distillation data. Since the annotations of target data are not available, as shown in Table 8, we adopt three filtering strategies based on objectness, classification scores and objectness&classification to select high-quality samples from RPN for distillation, respectively. The results indicate that when target data is involved in the distillation process, the model's performance decreases. This suggests that regardless of how we filter target samples, the selected samples inevitably contain a significant amount of noise, leading to a performance drop. In the end, we choose to distill only source samples in the DSD framework as the final solution, achieving the highest performance.

Choice of localization representation. For brevity, we refer to centerness and IoU as c and b. Here, we attempt to conduct ablation experiments on the localization-teacher and the localization representation. As shown in Table 9, when we train original OLN, using either c, b, or $\sqrt{bc}$ as localization representation to refine classification scores can improve the model's performance to different extents. This indicates that enhancing the consistency between classification and localization can effectively improve the detector's performance. Furthermore, when training with TROLN (ours), using the localization score ($\sqrt{4bc\tau_1\tau_2}$) as the localization representation to calibrate classification scores, the model achieves the highest performance improvement (+2.17% compared with baseline). This further validates the necessity and effectiveness of the TROLN training strategy and the DCE testing strategy. Moreover, we evaluate the variety of consistency in different classes before and after classification scores refinement on the test dataset. As shown in Fig. 5, it can be observed that the consistency has been improved to varying degrees across almost all categories, further demonstrating the effectiveness of the DCE mechanism.

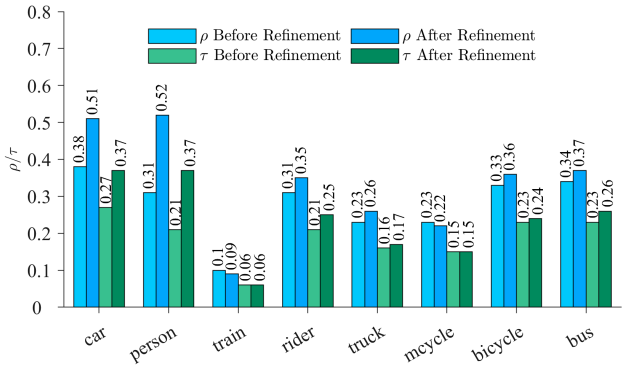

Figure 5. The variety of consistency across different classes before and after classification scores refinement on the ***FoggyCityscapes*** test datasets.

Table 9. The effect of localization-teacher and localization representation.

Training Stage	DCE Testing Stage				Representation	AP_{50}
	c	b	τ_1	τ_2		
						43.05
OLN	✓				c	43.95
		✓			b	43.98
	✓	✓			$\sqrt{bc}$	44.12
TROLN	✓	✓			$\sqrt{bc}$	44.76
	✓	✓	✓	✓	$\sqrt{4bc\tau_1\tau_2}$	**45.22**

5. Conclusion

To address source bias issue in domain adaptive object detection, we propose a distillation-based source debiasing framework. We train an instance-level classification-teacher model to guide the detector to acquire more domain-agnostic knowledge, improving the generalization on both domains. We also design a novel localization representation to refine classification scores, further improving the performance of the detector. Finally, our method achieved considerable improvement on several benchmark datasets under different base detectors for domain adaptation, demonstrating the effectiveness.

Acknowledgements

This work was supported by Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, and the National Natural Science Foundation of China under Grant 62176017.

Impact Statement

This paper presents work whose goal is to advance the field of Deep Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 19, 2006.

Bolya, D., Foley, S., Hays, J., and Hoffman, J. Tide: A general toolbox for identifying object detection errors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 558–573. Springer, 2020.

Cao, S., Joshi, D., Gui, L.-Y., and Wang, Y.-X. Contrastive mean teacher for domain adaptive object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23839–23848, 2023.

Chen, C., Zheng, Z., Ding, X., Huang, Y., and Dou, Q. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878, 2020.

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

Chen, Y., Li, W., Sakaridis, C., Dai, D., and Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3339–3348, 2018.

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

Deng, J., Li, W., Chen, Y., and Duan, L. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4091–4101, 2021.

Deng, J., Xu, D., Li, W., and Duan, L. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23829–23838, 2023.

Feng, C., Zhong, Y., Gao, Y., Scott, M. R., and Huang, W. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3490–3499. IEEE Computer Society, 2021.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189. PMLR, 2015.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. IEEE, 2012.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, M., Wang, Y., Wu, J., Wang, Y., Li, H., Li, B., Gan, W., Wu, W., and Qiao, Y. Cross domain object detection by target-perceived dual branch distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9580, 2022.

He, Z. and Zhang, L. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6668–6677, 2019.

He, Z. and Zhang, L. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 309–324. Springer, 2020.

Hsu, C.-C., Tsai, Y.-H., Lin, Y.-Y., and Yang, M.-H. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 733–748. Springer, 2020.

Inoue, N., Furuta, R., Yamasaki, T., and Aizawa, K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5001–5009, 2018.

Jiang, B., Luo, R., Mao, J., Xiao, T., and Jiang, Y. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 784–799, 2018.

Jiang, J., Chen, B., Wang, J., and Long, M. Decoupled adaptation for cross-domain object detection. *arXiv preprint arXiv:2110.02578*, 2021.

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S. N., Rosaen, K., and Vasudevan, R. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016.

Kim, D., Lin, T.-Y., Angelova, A., Kweon, I. S., and Kuo, W. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2): 5453–5460, 2022.

Li, W., Liu, X., Yao, X., and Yuan, Y. Scan: Cross domain object detection with semantic conditioned adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1421–1428, 2022a.

Li, W., Liu, X., and Yuan, Y. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5291–5300, 2022b.

Li, W., Liu, X., and Yuan, Y. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., and Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020.

Li, Y.-J., Dai, X., Ma, C.-Y., Liu, Y.-C., Chen, K., Wu, B., He, Z., Kitani, K., and Vajda, P. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7581–7590, 2022c.

Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., and Cai, J. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8771–8780, 2021.

Munir, M. A., Khan, M. H., Sarfraz, M., and Ali, M. Ssal: Synergizing between self-training and adversarial learning for domain adaptive object detection. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pp. 22770–22782, 2021.

Pu, Y., Liang, W., Hao, Y., Yuan, Y., Yang, Y., Zhang, C., Hu, H., and Huang, G. Rank-detr for high quality object detection. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

Ramamonjison, R., Banitalebi-Dehkordi, A., Kang, X., Bai, X., and Zhang, Y. Simrod: A simple adaptation method for robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3570–3579, 2021.

Redmon, J. and Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, 2015.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252, 2015.

Saito, K., Ushiku, Y., Harada, T., and Saenko, K. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965, 2019.

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

Tian, K., Zhang, C., Wang, Y., Xiang, S., and Pan, C. Knowledge mining and transferring for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9133–9142, 2021.

Tian, Z., Shen, C., Chen, H., and He, T. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019.

Tychsen-Smith, L. and Petersson, L. Improving object localization with fitness nms and bounded iou loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6877–6885, 2018.

Vs, V., Gupta, V., Oza, P., Sindagi, V. A., and Patel, V. M. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4516–4526, 2021.

Wang, Y., Zhang, R., Zhang, S., Li, M., Xia, Y., Zhang, X., and Liu, S. Domain-specific suppression for adaptive

object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9603–9612, 2021.

Wu, S., Li, X., and Wang, X. Iou-aware single-stage object detector for accurate localization. *Image and Vision Computing*, 97:103911, 2020.

Xu, C.-D., Zhao, X.-R., Jin, X., and Wei, X.-S. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11724–11733, 2020a.

Xu, M., Wang, H., Ni, B., Tian, Q., and Zhang, W. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12355–12364, 2020b.

Xu, Y., Sun, Y., Yang, Z., Miao, J., and Yang, Y. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14329–14339, 2022.

Yoo, J., Chung, I., and Kwak, N. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *European Conference on Computer Vision*, pp. 691–708. Springer, 2022.

Zhang, H., Wang, Y., Dayoub, F., and Sunderhauf, N. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8514–8523, 2021a.

Zhang, M., Song, G., Liu, Y., and Li, H. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6601–6610, 2023.

Zhang, Y., Wang, Z., and Mao, Y. Rpn prototype alignment for domain adaptive object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12425–12434, 2021b.

Zhao, L. and Wang, L. Task-specific inconsistency alignment for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14217–14226, 2022.

Zheng, Y., Huang, D., Liu, S., and Wang, Y. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13766–13775, 2020.

Zhou, W., Du, D., Zhang, L., Luo, T., and Wu, Y. Multi-granularity alignment domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9581–9590, 2022.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

A. More Implementation Details

A.1. Classification-Teacher

The classification-teacher model employs a ResNet101 (He et al., 2016) architecture pre-trained on ImageNet (Russakovsky et al., 2015) as its backbone, with an input size of 224×224. Data augmentation strategies encompass random horizontal flipping, color distortion, Gaussian blur, and solarization. The AdamW optimizer is employed for optimizing the classification-teacher model, initialized with a learning rate of 0.0001 for 12 epochs. We decay the learning rate by ratio 0.1 at epoch 9 and 11 and the total batch size is set to 64.

A.2. Localization-Teacher (TROLN)

TROLN. We reconstruct TROLN via adding a pixel-level global discriminator base on OLN (Kim et al., 2022). As shown in Table 10, we present the detailed architecture of this discriminator which consists of a Gradient Reversal Layer (GRL) and 4 conv layers. TROLN is trained with SGD optimizer with a 0.005 learning rate, 2 batch size for 12 epochs and we decay the learning rate by ratio 0.1 at epoch 6 and 7.

Table 10. Architectures of the adversarial alignment modules.

Global Discriminator
Gradient Reversal Layer (GRL)
Conv $256 \times 3 \times 3$, stride $1 \rightarrow$ LeakyReLU slope 0.2
Conv $128 \times 3 \times 3$, stride $1 \rightarrow$ LeakyReLU slope 0.2
Conv $128 \times 3 \times 3$, stride $1 \rightarrow$ LeakyReLU slope 0.2
Conv $1 \times 3 \times 3$, stride 1

A.3. DSD Training Based on AT and CMT

Due to original AT (Li et al., 2022c) or CMT (Cao et al., 2023) uses two-stage training, here we train AT or CMT for 20k iteration in the first stage and 10k iterations in the second stage. Moreover, our DSD module is only added to the detector in the second stage. Other hyper-parameters are the same as in the original implementation of AT and CMT. Our implementation is based on Detectron2 and the publicly available code by AT and CMT. Each experiment is conducted on 4 NVIDIA 3090 GPUs.

B. Further Ablation Studies

To further analyze the effect of the data on the DSD and TROLN, we conduct extensive ablation studies in this section. Here, all experiments are done with DA-Faster (Chen et al., 2018) with ResNet50 backbone on *FoggyCityscapes* test set.

B.1. IoU Threshold in DSD Framework

We empirically choose IoU threshold T to analyze how T affects the detector's performance in the DSD framework. As shown in Table 11, we test a range of T. On one hand, when T is smaller, the filtered positive samples have significant differences from the ground truth, introducing extra noise into the classification-teacher model and causing a drop in model performance. On the other hand, a larger T results in a significant reduction in the number of samples, weakening the effect of distillation. Eventually, we set $T = 0.8$ with the best performance.

Table 11. Effects of T for estimating the DSD module.

T	0.65	0.7	0.75	0.8	0.85	0.9
mAP	40.12	40.64	41.86	**43.05**	42.56	42.12

B.2. Choices of Training Data for TROLN

In order to investigate the effect of training data on TROLN, we conduct an ablation study on choices of training data. As shown in Table 12, we train the original OLN (Kim et al., 2022) using source data (S), source2target data (T') and $S\&T'$,

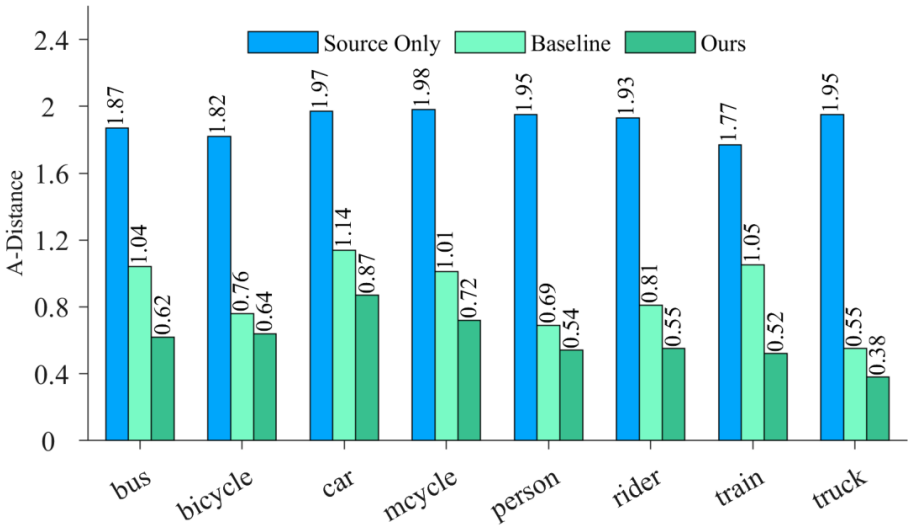

Figure 6. Feature distribution discrepancy of foregrounds.

respectively. It's worth noting that training OLN solely on T' data can not effectively mine target related information. This may be because there is still a domain shift between generation data T' and the real target data. Moreover, training OLN based on both S and T' effectively enhances baseline performance. This indicates that mixed-style data is beneficial for the localization network to generalize to target data. Finally, when incorporating the domain affinity weights τ_1 and τ_2 into both the training of TROLN and the design of localization quality representation, the model achieves the highest performance (45.22%).

Table 12. Effects of Training data in TROLN on detector's performance.

Method	Training Data		DCE Testing Stage				Metric	AP_{50}
	S	T'	c	b	τ_1	τ_2		
								43.05
OLN	✓		✓	✓			$\sqrt{bc}$	35.76
		✓	✓	✓			$\sqrt{bc}$	41.96
	✓	✓	✓	✓			$\sqrt{bc}$	44.12
TROLN	✓	✓	✓	✓			$\sqrt{bc}$	44.76
	✓	✓	✓	✓	✓	✓	$\sqrt{4bc\tau_1\tau_2}$	**45.22**

C. Further Analysis

C.1. Distribution Discrepancy of Foregrounds

The theoretical result in (Ben-David et al., 2006) indicates that $\mathcal{A}$-distance can serve as a metric for quantifying domain discrepancy. In practice, we calculate the Proxy $\mathcal{A}$-distance as an approximation, which is defined as $d_{\mathcal{A}} = 2(1-\epsilon)$. Here, ϵ represents the generalization error of a binary classifier (implemented with two fully connected layers in our experiments) that tries to distinguish which domain the ROI features come from. Fig. 6 displays the distances for each category on the ***Cityscapes-to-FoggyCityscapes*** task with the foreground features extracted from the models of *Source Only* (Ren et al., 2015), *Baseline* (Chen et al., 2018) and *Ours*. Compared with the *Source Only* model, *Baseline* and *Ours* reduce the distances in all the categories by large margins, which demonstrates the necessity of feature alignment. Furthermore, since we utilize a classification-teacher to distill domain-agnostic knowledge to the detector, we achieve a smaller $\mathcal{A}$-distance compared to *Baseline*.

After training, for each category, we randomly sample the same number of ROI features from the source and target domain for T-SNE visualization, as shown in Fig. 7. It can be observed that those similar categories (truck, bus and train) can be separated clearly by our method, which benefits the following detection.

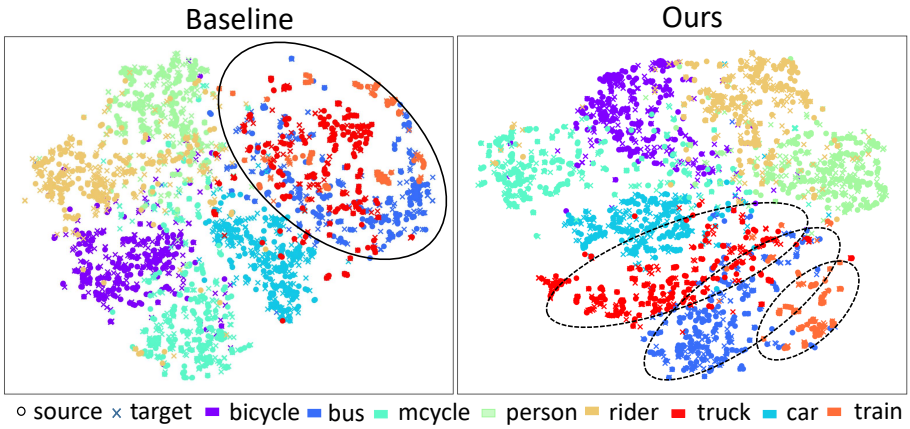

Figure 7. T-SNE visualization of features produced by the baseline model and our method.

Table 13. TIDE error analysis. The ΔAP^{box}@0.5 metric is defined as how much AP_{50} can be added to the detector if an oracle fixes a certain error type in TIDE (Bolya et al., 2020).

Module		ΔAP^{box}@0.5					
DSD	DCE	*cls*↓	*loc*↓	*both*↓	*dup*↓	*bg*↓	*miss*↓
		5.45	12.46	1.41	0.01	**1.32**	15.62
✓		4.94	9.95	**1.40**	**0.00**	1.47	17.07
✓	✓	**3.78**	**9.06**	1.58	0.05	1.97	**13.79**

C.2. Error Analysis of Detection Results

To further validate the effect of the proposed framework for cross-domain object detection, we analyze the detection errors of the models of *Baseline*, *Baseline+DSD* and *Baseline+DSD+DCE (Ours)* via the TIDE toolbox (Bolya et al., 2020) on the ***Cityscapes-to-FoggyCityscapes*** task. As shown in Table 13, we follow TIDE to categorize the detection errors into six types: *cls*: localized correctly, but classified incorrectly; *loc*: classified correctly, but localized incorrectly; *both*: classified incorrectly and localized incorrectly; *dup*: two or more detected boxes match with the same ground-truth box; *bg*: classifying the background as foreground mistakenly; *miss*: foreground objects are not detected by the detector. (See (Bolya et al., 2020) for more details and discussion.)

We observe that both *Baseline+DSD* and *Ours* make fewer classification and localization errors than *Baseline*. It indicates that the DSD module effectively distills domain-agnostic features from the classification-teacher to the detector, guiding the detector to extract the superior feature representation of the foreground. Besides, *Ours* performs the best in terms of the *miss* error category among the three models. This further illustrates that employing DCE to refine the classification scores can effectively improve the consistency between classification and localization. The DCE enables detected boxes with better localization to also have higher classification scores, thereby reducing *miss* error.

D. Qualitative Results

We present more qualitative comparisons among (a) *Source Only*, (b) *Baseline* (Chen et al., 2018), (c) *Ours*, and (d) *Ground-truth* in Fig. 8. Our method can eliminate some missing errors and avoid some wrong classification cases compared with the *Baseline*, which verifies the effectiveness of proposed DSD module and DCE strategy.

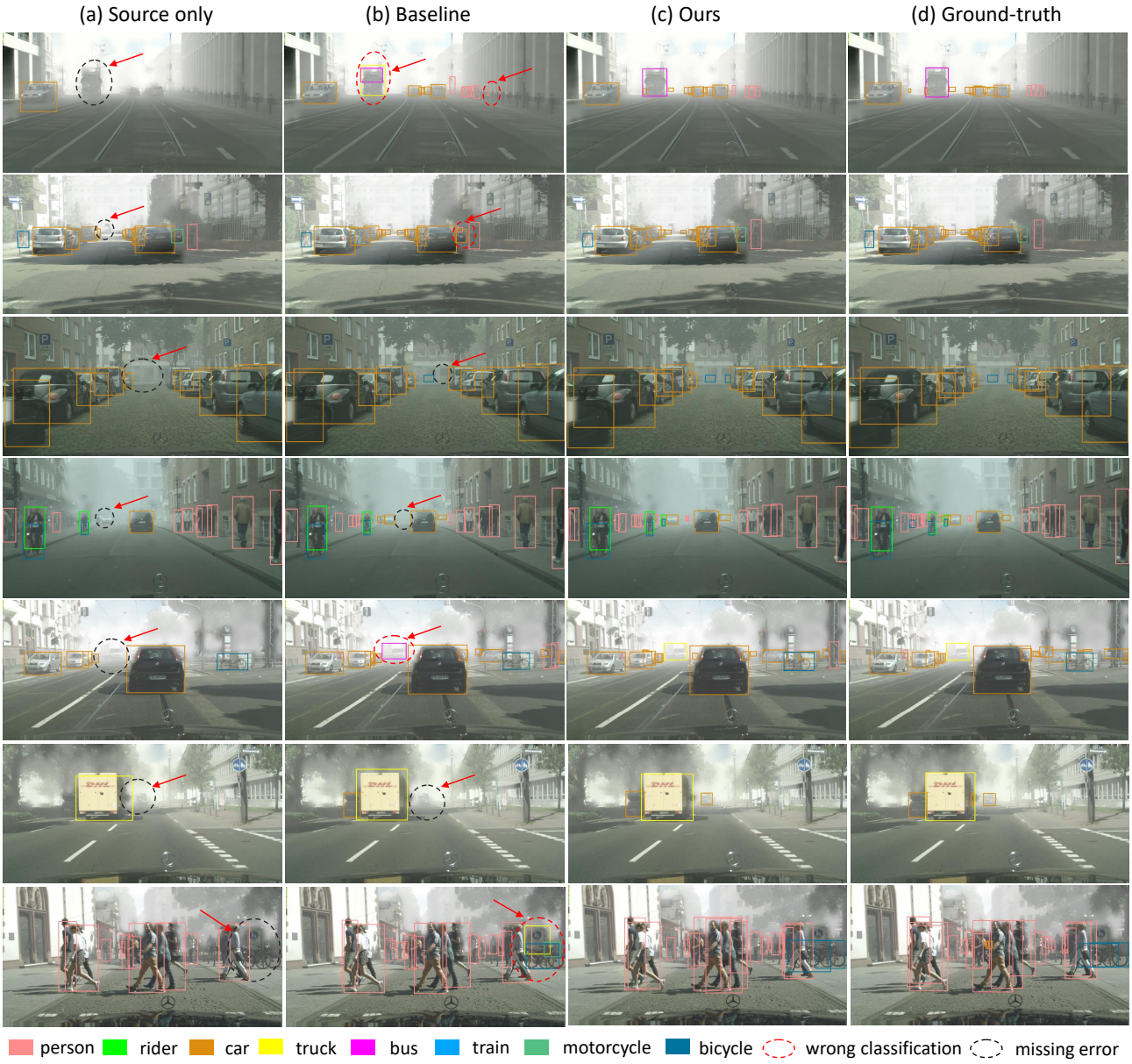

Figure 8. Qualitative results on the ***Cityscapes-to-FoggyCityscapes*** adaptation scenario of (a) the *Source Only* model, (b) *Baseline* (Chen et al., 2018), (c) *Ours*, and (d) *Ground-truth*. (Zooming in for best view.)