

Domain Adaptive Object Detection via Dual-Stream Bilevel-Cycle Optimization

Yannan Chen^{1,4}, Wei Wang¹, Wenqiang Wang¹, Ruoyu Chen²
Jiancheng Wang³, Mingbo Yang¹, Yaowei Wang⁴, and Xiaochun Cao¹

¹ School of Cyber Science and Technology, Sun Yat-sen University,
Shenzhen Campus, Guangdong 518107, China

² University of Chinese Academy of Sciences, Beijing 100049, China

³ School of Computer Science and Technology, Anhui University,
Hefei, Anhui 230601, China

⁴ PC Laboratory, Shenzhen, Guangdong 518055, China

Abstract. Cycle self-training (CST) breaks the shared classifier assumption of the standard self-training framework, which is effective for unsupervised domain adaptation and exploits unlabeled target data by training with target pseudo-labels. CST introduces a target classifier and employs an inner-outer loop updating strategy, addressing the issue of unreliable pseudo-labels and enabling pseudo-labels to generalize across domains. Despite its success in image classification, extending CST to object detection faces three main challenges. First, the upper bound of CST in object detection is constrained by three types of unreliable pseudo-labels, such as classification error alone, localization error alone, and their combination. Second, since object detection involves detecting multiple target objects, directly applying CST leads to training instability. Third, a wider numerical range of regression coordinates leads to exploding losses. To this end, we apply CST to both classification and regression and propose the Dual-Stream Bilevel-Cycle Optimization framework. Specifically, we construct CST upon Mean Teacher to prevent training instability and use extra normalization to map the regression bounding box into a standardized space, effectively addressing exploding losses. Also, we provide a theoretical derivation of the regression bound. Extensive experiments across four cross domain standard scenarios demonstrate that our framework achieves considerable results.

Keywords: Domain Adaptive Object Detection · Self-training · Mean Teacher

1 Introduction

Self-Training (ST) serves as an effective paradigm in unsupervised domain adaptation [28], which leverages a labeled source domain to annotate an unlabeled target domain, thereby reducing the high cost of manual annotation [1]. Specifically,

 Corresponding authors: Wei Wang and Xiaochun Cao.

Emails: wangwei29@mail.sysu.edu.cn and caoxiaochun@mail.sysu.edu.cn.

ST first utilizes a model trained on the source domain to generate pseudo-labels for the target domain [40]. Subsequently, it iteratively retrain the model using both the ground-truth labels from the source domain and the generated pseudo-labels from the target domain [5, 44, 50]. However, due to the distribution shift between the source and target domains, the pseudo-labels for the target domain are often unreliable, *i.e.*, they deviate from the ground-truth labels. To address this issue, Cycle Self-Training (CST) relaxes the shared classifier assumption of ST by introducing a specific target domain classifier and employing an inner-outer loop update strategy to mitigate the unreliability of pseudo-labels [21, 32].

Although CST has achieved superior performance in classification tasks [28, 48], its application in object detection remains underexplored. Through extensive experiments and theoretical analysis, we identify three primary challenges: **❶** As shown in Fig. 1 (a), similar to classification tasks [26, 44], ST in object detection also suffers from a performance upper bound. However, the unreliability of pseudo-labels in detection manifests in three forms: inaccurate classification, inaccurate localization, and inaccuracy in both classification and localization. **❷** As illustrated in Fig. 1 (b), directly applying CST to the Fully Convolutional One-Stage (FCOS) detector leads to training instability [39]. **❸** As depicted in Fig. 1 (c), Mean Teacher (MT) without normalization leads to severe bounding box shifts and model collapse [4, 37, 46]. Therefore, despite its success in image classification, extending CST to object detection is still challenging [7, 19, 41].

To break the performance upper bound caused by unreliable pseudo-labels (**Challenge ❶**), we propose a **Dual-Stream Bilevel-Cycle Optimization (DSBCO)** framework, as shown in Fig. 2. In DSBCO, we extend the CST mechanism to both classification and regression tasks. By enforcing consistency across domains, our DSBCO effectively filters out unreliable pseudo-labels. We address the training instability (**Challenge ❷**) by constructing our framework upon the stable MT paradigm. Specifically, we leverage the exponential moving average (EMA) [3] to smooth the parameter updates and stabilize the learning process in dense detection tasks. Finally, to prevent model collapse caused by the wide numerical range of regression coordinates (**Challenge ❸**), we introduce the regression normalization mechanism. This strategy projects the large regression targets into a standard distribution, which ensures stable gradient descent and effectively mitigates large-scale bounding box shifts.

We validate the effectiveness of the proposed DSBCO framework through extensive experiments across four domain adaptation scenarios. Our method exceeds all existing FCOS detector methods. Specifically, compared to the second-best Harmonious Teacher (HT) [9], DSBCO achieves effective improvements, increasing performance by 14.3% in Adverse Weather Adaptation, 3.1% in Synthetic Source Adaptation, and approximately 2% in both Distinct Camera and Diverse Context Adaptations. Detailed descriptions and evaluations of these four adaptation scenarios are provided in Section 4. Moreover, comparing with methods using other detectors such as Faster R-CNN [33], Def DETR [62], and RetinaNet [27], DSBCO exhibits competitive performance on cross-domain situa-

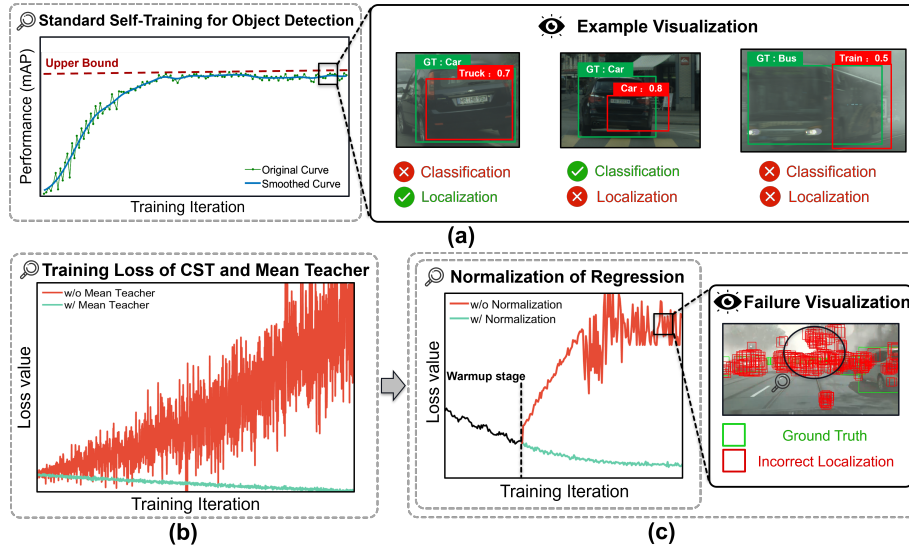


Fig. 1: Motivation of this paper illustrating the challenges of extending CST to object detection. (a) Through empirical analysis, we observe that standard ST exhibits a performance upper bound in object detection. The visualization reveals that this limitation stems from incorrect classification, inaccurate localization, and combined failure. (b) Directly applying CST leads to training instability. (c) Implementing MT without regression normalization results in severe loss and model collapse where the visualization displays large bounding box shifts.

tions, confirming its effective generalization capability. Our contributions are summarized as follows:

- We propose the DSBCO framework, which extends the CST strategy to both classification and regression tasks, effectively breaking the performance bottleneck caused by unreliable pseudo-labels.
- We construct CST upon the MT paradigm to address the training instability associated with dense predictions in multiple object scenarios.
- We introduce a extra normalization mechanism to resolve exploding losses caused by the wide numerical range of regression coordinates.
- Extensive experiments demonstrate that DSBCO achieves considerable performance on four standard cross-domain situations. Additionally, we provide the theoretical generalization bound derivation for the regression task.

2 Related Work

Domain Adaptive Object Detection (DAOD) aims to transfer learned visual representations from labeled source domains to unlabeled targets under challenging distribution shifts [14, 40]. To relax these shifts, adversarial

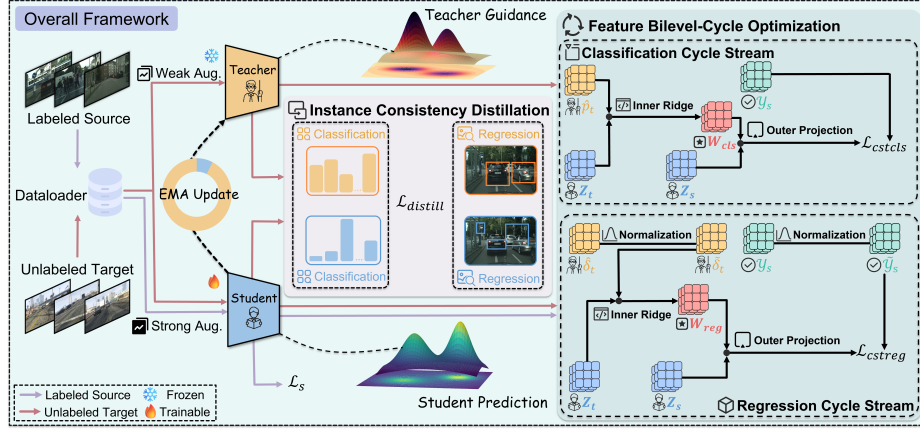


Fig. 2: Overview of the proposed DSBCO framework constructed upon the stable MT paradigm. On the left, the trainable Student network processes strongly augmented images from both domains, while the frozen Teacher, updated via EMA, uses only weakly augmented target images to generate pseudo-labels. In the middle, Instance Consistency Distillation applies self-training as a baseline. On the right, Feature Bilevel-Cycle Optimization employs Classification and Regression Cycle Streams, where the latter introduces normalization to prevent exploding losses. In both streams, the Inner Ridge and Outer Projection work together to align the domains by mapping target features back to the source, effectively extending CST to the object detection task.

feature alignment matches feature distributions to extract and explicitly align shared domain-invariant representations from both domains [10, 11, 30]. In parallel, image-to-image translation synthesizes target-style images or disentangles domain-specific frequencies to harmonize the inputs at the pixel level [32]. Beyond these global alignment strategies, graph-based methods utilize graph matching or relational learning to explicitly enforce cross-domain structural consistency, effectively aligning semantic topologies across domains [2, 44, 47]. Furthermore, transformer-based approaches leverage self-attention mechanisms to learn robust features that handle complex domain degradation [20, 46, 48, 55]. Among these diverse adaptation methods, ST effectively generates pseudo-labels on the target data to supervise the model iteratively, gradually narrowing the domain gap through reliable high-confidence predictions [1, 5, 43, 50]. Inspired by this success, our proposed method is primarily built upon the foundation of ST.

Self-Training (ST) serves as an effective paradigm in semi-supervised learning and domain adaptation [1, 28]. Early strategies typically employ a single model to generate pseudo-labels and iteratively refine the network by selecting high-confidence predictions on unlabeled target data [10, 58, 60]. To enhance the stability of this process, the MT framework emerges as a robust and effective approach by maintaining a teacher model through the EMA of student parameters [3, 4, 25, 37]. This temporal ensemble mechanism provides more consistent supervision signals and effectively handles the noise inherent in pseudo-

labels [5, 9, 15]. Recent advances in this field further improve the reliability of supervision by incorporating clustering filters to remove outliers or adopting decoupled training strategies to balance classification and localization tasks [57, 61]. However, directly applying these co-training or self-training methods to object detection often leads to severe training instability and potential loss explosion due to the wide numerical range of regression targets [41, 46, 49, 50]. To address the problem, our DSBCO framework builds upon the stable MT paradigm to ensure effective adaptation across both classification and regression tasks.

3 Methodology

3.1 Preliminaries

The DAOD task involves a labeled source domain $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$, where x_s^i represents the input image and $y_s^i = \{y_{s,i}^{cls}, y_{s,i}^{reg}\}$ contains the category labels and bounding box coordinates. Meanwhile, an unlabeled target domain $\mathcal{D}_t = \{x_t^j\}_{j=1}^{N_t}$ is available for adaptation. Both domains share the same object categories but differ in data distributions, denoted as $P(\mathcal{X}_s) \neq P(\mathcal{X}_t)$, due to the domain shift.

3.2 DSBCO Framework

In this section, we present the details of the proposed DSBCO framework. As illustrated in Fig. 2, our framework follows a left-to-right processing pipeline and mainly consists of three components: 1) Mean Teacher (MT) paradigm that provides stable pseudo-labels; 2) Instance Consistency Distillation aimed at promoting sample-level prediction consistency; and 3) Feature Bilevel-Cycle Optimization module for addressing distribution shifts at the feature level.

MT Paradigm As illustrated in Fig. 1 (b), directly applying CST to dense prediction object detectors leads to extreme training instability and even model collapse (**Challenge ②**). To address this instability problem caused by dense prediction, we adopt MT paradigm. In MT paradigm, the parameters of the Teacher model are updated via the EMA of the Student model. The update rule is formally defined as:

$$\theta_{tea}^{(k)} = \alpha \theta_{tea}^{(k-1)} + (1 - \alpha) \theta_{stu}^{(k)}, \quad (1)$$

where α represents the smoothing coefficient, and $\theta_{tea}^{(k)}$ and $\theta_{stu}^{(k)}$ denote the parameters of the Teacher and Student networks at k -th iteration. This temporal smoothing process ensures prediction consistency across iterations, thereby providing stable pseudo-label for the target domain. Furthermore, compared to the structurally complex two-stage Faster R-CNN, the anchor-free FCOS detector directly decouples the task into independent pixel-level classification and regression predictions in a fully convolutional method. This characteristic aligns well

with our proposed dual-stream optimization strategy, so we select it as the base architecture.

In terms of the data flow direction, the Teacher model only receives weakly augmented (*e.g.*, basic image resizing and horizontal flipping) target domain images $\mathcal{A}_w$ to generate reliable pseudo-labels. Meanwhile, the trainable Student model receives source domain images x_s and target domain images x_t processed by strong augmentation $\mathcal{A}_s$ (*e.g.*, introducing severe perturbations like color jittering, Gaussian blur, and random erasing).

The Student model is initially trained on the strongly augmented source images in a supervised way. The supervised loss $\mathcal{L}_{sup}$ is the sum of the classification loss and the regression loss:

$$\mathcal{L}_{sup} = \mathcal{L}_{cls1} + \mathcal{L}_{reg1}. \quad (2)$$

Specifically, the classification task employs the Focal Loss [27]:

$$\mathcal{L}_{cls1} = -(1 - p_s)^\gamma \log(p_s), \quad (3)$$

where p_s denotes the probability predicted by the Student, and γ is the focusing hyperparameter. For the regression task, to overcome the scale sensitivity of the traditional L_1 loss and the vanishing gradient problem of the standard IoU, we adopt the Generalized Intersection over Union (GIoU) loss [34]:

$$\mathcal{L}_{reg1} = 1 - \left(\frac{|B_s \cap B^*|}{|B_s \cup B^*|} - \frac{|C \setminus (B_s \cup B^*)|}{|C|} \right), \quad (4)$$

where B_s and B^* denote the predicted bounding box and the ground-truth bounding box, respectively, and C is the smallest enclosing box covering both. The GIoU provides global regression constraints and continuous gradients, which are crucial for the early training stability of anchor-free models.

Instance Consistency Distillation Building upon the stable MT paradigm, we employ an instance consistency distillation mechanism as the baseline for self-training. This mechanism aims to enforce consistency between the Student model’s predictions under strong augmentation and the Teacher model’s pseudo-labels at the instance level.

Given that the Teacher model may produce unreliable predictions containing substantial background noise on the target domain, we first generate a binary mask $\mathbb{M}$ using the Teacher model’s predicted classification probability $\hat{p}_t$ and localization quality $\hat{c}_t$ (*i.e.*, centerness) to retain only high-quality samples:

$$\mathbb{M} = \mathbb{I}((\hat{p}_t > \tau_{cls}) \wedge (\hat{c}_t > \tau_{reg})), \quad (5)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function (evaluating to 1 if the condition holds, retaining the sample) [45], and τ_{cls} and τ_{reg} are the predefined thresholds for classification probability and localization confidence, respectively. The instance consistency loss $\mathcal{L}_{distill}$ is computed within the valid regions defined by $\mathbb{M}$:

$$\mathcal{L}_{distill} = \frac{1}{\sum \mathbb{M}} \sum \mathbb{M} \cdot (\mathcal{L}_{cls2} + \mathcal{L}_{reg2}). \quad (6)$$

For classification alignment loss $\mathcal{L}_{cls2}$, since the Teacher model provides continuous soft labels, we directly minimize the discrepancy by computing Quality Focal Loss between the Student model’s prediction p_s and the Teacher model’s soft label $\hat{p}_t$ [24]. The formula is as follows:

$$\mathcal{L}_{cls2} = -|\hat{p}_t - p_s|^\beta [\hat{p}_t \log(p_s) + (1 - \hat{p}_t) \log(1 - p_s)], \quad (7)$$

where β is a modulating factor used to suppress the contribution of easy samples where the Student model’s prediction is already highly consistent with the Teacher model’s soft label [54]. For regression consistency loss $\mathcal{L}_{reg2}$, we similarly compute the GIoU loss between the Student model’s predicted boxes and the Teacher model’s pseudo-label boxes.

Feature Bilevel-Cycle Optimization As shown in Fig. 1 (a), standard self-training exhibits a performance upper bound in object detection due to its heavy reliance on unreliable pseudo-labels (**Challenge ❶**). To overcome this, we propose Feature Bilevel-Cycle Optimization strategy. This method formulates domain adaptation as a bilevel optimization problem. Unlike the original image level classification method CST, we leverage the decoupling characteristics of FCOS to design a dual-stream strategy containing classification and regression tasks, achieving optimization in the feature space.

Classification Cycle Stream. This branch is responsible for aligning semantic category features. In the inner loop (Inner Ridge), we abandon the unstable sample kernel matrix inversion method used in traditional CST and instead construct a covariance matrix along the feature dimension. We aim to find an optimal linear projection matrix $\mathbf{W}_{cls} \in \mathbb{R}^{d \times C}$ that can reconstruct the Teacher’s pseudo-label vectors $\hat{p}_t$ from the Student’s target domain features $\mathbf{Z}_t \in \mathbb{R}^{N_t \times d}$. This is formulated as a ridge regression problem with a closed-form solution:

$$\mathbf{W}_{cls} = ((\mathbf{Z}_t)^\top \mathbf{Z}_t + \lambda I)^{-1} (\mathbf{Z}_t)^\top \hat{p}_t, \quad (8)$$

where λ is the regularization coefficient. In the outer loop (Outer Projection), we directly apply this explicit mapping head $\mathbf{W}_{cls}$ learned on target domain to the source domain features $\mathbf{Z}_s \in \mathbb{R}^{N_s \times d}$, and enforce the projected results to align with source ground-truth labels. The classification cycle loss is defined as:

$$\mathcal{L}_{cstcls} = \|\mathbf{Z}_s \mathbf{W}_{cls} - y_s\|_F^2, \quad (9)$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm of a matrix [42]. By minimizing this loss, we encourage the network to learn domain-invariant semantic features.

Regression Cycle Stream. As shown in Fig. 1 (c), due to the unbounded nature and wide numerical range of bounding box regression offsets, directly applying cycle consistency to the regression task leads to severe exploding losses (**Challenge ❸**). To address this issue, we introduce a regression normalization

mechanism in this stream. In the inner loop, we first compute the mean μ_t and standard deviation σ_t of the target domain regression offsets $\hat{\delta}_t$ generated by the Teacher model, and normalize them into a standard normal distribution:

$$\tilde{\delta}_t = \frac{\hat{\delta}_t - \mu_t}{\sigma_t + \epsilon}, \quad (10)$$

where ϵ is a small constant to prevent division by zero. Subsequently, we utilize these normalized targets to solve for the optimal regression matrix $\mathbf{W}_{reg} \in \mathbb{R}^{d \times 4}$:

$$\mathbf{W}_{reg} = ((\mathbf{Z}_t)^\top \mathbf{Z}_t + \lambda I)^{-1} (\mathbf{Z}_t)^\top \tilde{\delta}_t. \quad (11)$$

During the outer loop projection, we normalize the source domain's ground-truth labels y_s using the same target domain statistics (μ_t, σ_t) to obtain $\tilde{y}_s$. Finally, the regression cycle loss $\mathcal{L}_{cstreg}$ is computed as:

$$\mathcal{L}_{cstreg} = \|\mathbf{Z}_s \mathbf{W}_{reg} - \tilde{y}_s\|_F^2. \quad (12)$$

Through this normalization and reconstruction mechanism, we map the large regression values into a stable space, effectively preventing gradient divergence. This effectively extends CST to the localization task in object detection.

Total Optimization Objective The overall optimization objective $\mathcal{L}_{total}$ combines the supervised loss from the source domain, the instance level consistency distillation loss, and the feature level cycle alignment loss. The complete loss function is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_{unsup} \mathcal{L}_{distill} + \lambda_{cycle} (\mathcal{L}_{cstcls} + \eta \mathcal{L}_{cstreg}), \quad (13)$$

where λ_{unsup} and λ_{cycle} are employed to balance the contributions of the unsupervised signals. Meanwhile, η is a hyperparameter specifically designed to adjust the relative importance of the regression cycle alignment. The complete training process is detailed in Algorithm 1.

3.3 Theoretical Regression Bound

Extending generalization bounds to continuous bounding box regression is non-trivial due to the infinite output space and optimization instability. We therefore analyze DSBCO by bounding the target regression error, characterizing the failure of standard ST, and showing how cycle optimization tightens the bound.

Theorem 1 (Generalization Upper Bound). *Under the Lipschitz continuity assumption for continuous box regression, the expected target error $\mathcal{E}_{\mathcal{Q}}(f_s)$ is bounded by the sum of cycle reconstruction error and stability terms:*

$$\mathcal{E}_{\mathcal{Q}}(f_s) \leq \hat{\mathcal{E}}_{\mathcal{P}}(f_t) + \hat{c}_{\mathcal{Q}}(f_s, f_t) + \mathcal{R}(f_t, \delta) + \mathcal{O}(\mathfrak{R}_n(\mathcal{F})). \quad (14)$$

Algorithm 1 DSBCO Algorithm

Input: Labeled source domain $\mathcal{D}_s = \{(x_s, y_s)\}$, Unlabeled target domain $\mathcal{D}_t = \{x_t\}$,
 Max Iterations T_{max} , Burn-in T_{burn} , learning rate l_r
Output: Optimized Student Parameters θ_{stu}^* .

- 1: $\theta_{stu} \leftarrow \text{RandomInit}(\cdot)$, $\theta_{tea} \leftarrow \theta_{stu}$. // **Initialization**
- 2: **for** $t = 1$ **to** T_{max} **do**
- 3: $\{(x_s, y_s)\} \sim \mathcal{D}_s$ & $\{x_t\} \sim \mathcal{D}_t$ // **Sample Batch**
- 4: /* Instance Consistency Distillation */
- 5: $\hat{p}_t, \hat{\delta}_t \leftarrow F_{\theta_{tea}}(\mathcal{A}_w(x_t))$
- 6: $\mathbb{M} = \mathbb{I}((\hat{p}_t > \tau_{cls}) \wedge (\hat{\delta}_t > \tau_{reg}))$
- 7: $\mathcal{L}_{sup} \leftarrow (F_{\theta_{stu}}(\mathcal{A}_s(x_s)), y_s)$
- 8: $\mathcal{L}_{distill} \leftarrow (\hat{p}_t, \hat{\delta}_t, F_{\theta_{stu}}(\mathcal{A}_s(x_t)), \mathbb{M})$
- 9: $\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_{unsup} \mathcal{L}_{distill}$
- 10: /* Distribution Bilevel-Cycle Optimization */
- 11: **if** $t > T_{burn}$ **then**
- 12: $\mathbf{Z}_s \leftarrow \Phi_{\theta_{stu}}(x_s)$ & $\mathbf{Z}_t \leftarrow \Phi_{\theta_{stu}}(x_t)$ // **Extract Features**
- 13: // **Classification Cycle Stream**
- 14: $\mathbf{W}_{cls} = ((\mathbf{Z}_t)^\top \mathbf{Z}_t + \lambda \mathbf{I})^{-1} (\mathbf{Z}_t)^\top \hat{p}_t$
- 15: $\mathcal{L}_{cstcls} = \|\mathbf{Z}_s \mathbf{W}_{cls} - y_s\|_F^2$
- 16: // **Regression Cycle Stream**
- 17: $\tilde{\delta}_t = (\hat{\delta}_t - \mu_t) / (\sigma_t + \epsilon)$ // **Normalization**
- 18: $\mathbf{W}_{reg} = ((\mathbf{Z}_t)^\top \mathbf{Z}_t + \lambda \mathbf{I})^{-1} (\mathbf{Z}_t)^\top \tilde{\delta}_t$
- 19: $\hat{y}_s = (y_s - \mu_t) / (\sigma_t + \epsilon)$ // **Normalization**
- 20: $\mathcal{L}_{cstreg} = \|\mathbf{Z}_s \mathbf{W}_{reg} - \hat{y}_s\|_F^2$
- 21: $\mathcal{L}_{total} = \mathcal{L}_{total} + \lambda_{cycle} (\mathcal{L}_{cstcls} + \eta \mathcal{L}_{cstreg})$
- 22: **end if**
- 23: /* Parameter Optimization */
- 24: $\theta_{stu} = \theta_{stu} - l_r \nabla \mathcal{L}_{total}$
- 25: $\theta_{tea} = \alpha \theta_{tea} + (1 - \alpha) \theta_{stu}$
- 26: **end for**
- 27: **return** θ_{stu}^*

Here, $\mathcal{E}_{\mathcal{Q}}(f_s) = \mathbb{E}_{(x,y) \sim \mathcal{Q}}[\mathcal{L}_{reg}(f_s(x), y)]$ is the target regression error, $\hat{\mathcal{E}}_{\mathcal{P}}(f_t) = \frac{1}{n_s} \sum_i \mathcal{L}_{reg}(f_t(x_s^i), y_s^i)$ is the empirical source cycle reconstruction error, $\hat{\mathcal{C}}_{\mathcal{Q}}(f_s, f_t) = \frac{1}{n_t} \sum_j d_{box}(f_s(x_t^j), f_t(x_t^j))$ measures the empirical student-teacher discrepancy on target samples, $\mathcal{R}(f_t, \delta)$ denotes the probability of large box shifts under local perturbations, and $\mathfrak{R}_n(\mathcal{F})$ is the Rademacher complexity of the detector class.

Theorem 2 (Failure of Standard ST). Under a hard scenario where target pseudo-labels are dominated by background correlations, standard ST converges to a spurious solution $f_{spurious}$ with an irreducible target regression error:

$$\mathcal{E}_{\mathcal{Q}}(f_{ST}) \geq \epsilon > 0. \quad (15)$$

Theorem 3 (Convergence of DSBCO). Minimizing the cycle consistency term $\hat{\mathcal{E}}_{\mathcal{P}}(f_t)$ acts as a source-domain verification mechanism. A spurious solu-

tion incurs high source reconstruction error, whereas the causal object-semantic solution preserves localization semantics and approaches zero cycle loss.

Theorem 4 (Comprehensive Generalization Bound for DSBCO). *By combining the regression generalization bound (Theorem 1) and the robustness analysis, we prove that optimizing the DSBCO objective $\mathcal{L}_{DSBCO}$ directly minimizes the dominant empirical terms in the target error upper bound:*

$$\mathcal{E}_{\mathcal{Q}}(f_s) \leq \mathcal{L}_{DSBCO} + \mathcal{O}(\mathfrak{R}_n(\mathcal{F})) + \text{const.} \quad (16)$$

Theoretical Analysis. Complete proofs are provided in **Section A** of the supplementary material. **Theorem 1** formally shows that target localization error is controlled by source cycle reconstruction, student-teacher discrepancy, and regression non-robustness. **Theorem 2** further explains why standard ST can be trapped by background-biased pseudo-labels. Furthermore, **Theorem 3** and **Theorem 4** show that DSBCO tightens the bound through complementary mechanisms. Specifically, **Theorem 3** uses cycle losses to reject spurious mappings and stabilize source-target regression correspondence, while **Theorem 4** employs instance consistency distillation to reduce $\hat{c}_{\mathcal{Q}}$ and dual-stream thresholding to effectively suppress $\mathcal{R}(f_t, \delta)$ against noisy pseudo-label perturbations.

4 Experiments

In this section, we present a comprehensive evaluation of the proposed DSBCO framework. We evaluate DSBCO on four different domain adaptation scenarios to test the model’s robustness against weather changes, scene variations, camera differences, and synthetic-to-real shifts.

4.1 Experimental Setup

Cross Domain Scenarios and Datasets. We evaluate our framework on four standard benchmarks covering diverse domain shifts. **Adverse Weather Adaptation** transfers from **Cityscapes** [8], which captures clear street scenes from 50 German cities, to **Foggy Cityscapes** [35], synthesized with realistic fog density $\beta = 0.02$, to test robustness against visibility degradation. **Diverse Context Adaptation** adapts Cityscapes to the daytime driving subset of **BDD100K** [51], collected across various locations with complex lighting conditions, addressing variations in geography and scene layout. **Distinct Camera Adaptation** utilizes **KITTI** [12], characterized by a wide aspect ratio and unique sensor setup, as the source for Cityscapes to handle significant cross-camera geometric mismatches from differing camera intrinsics. Finally, **Synthetic Source Adaptation** leverages **Sim10K** [18], rendered from the game GTA V, to adapt to the target **Cityscapes** domain, bridging the synthetic-to-real texture gap. Details are provided in Supplementary **Section B**.

Evaluation Metrics. Following standard protocols in cross-domain object detection, we employ the mean Average Precision (mAP) as the primary evaluation

Table 1: Quantitative results on **Adverse Weather Adaptation** (Cityscapes $\rightarrow$ Foggy Cityscapes). The mean Average Precision (%) is evaluated on the target domain. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	Detector	Person	Rider	Car	Truck	Bus	Train	Motor	Bicycle	mAP $\uparrow$
D_adapt [17]	ICLR '22	Faster R-CNN	44.9	54.2	61.7	25.6	36.3	24.7	37.3	46.1	41.3
TDD [15]	CVPR '22	Faster R-CNN	39.6	47.5	55.7	33.8	47.6	42.1	37.0	41.4	43.1
AT [25]	CVPR '22	Faster R-CNN	45.5	55.1	64.2	35.0	56.3	54.3	38.5	51.9	50.9
PT [6]	ICML '22	Faster R-CNN	40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
CMT [4]	CVPR '23	Faster R-CNN	45.9	55.7	63.7	<u>39.6</u>	<u>66.0</u>	38.8	41.4	51.2	50.3
AT + DSD-DA [11]	ICML '24	Faster R-CNN	49.1	59.3	66.2	35.8	60.0	47.1	45.2	54.9	52.2
CMT + DSD-DA [11]	ICML '24	Faster R-CNN	49.0	59.6	65.3	35.7	61.0	46.5	43.9	57.3	52.3
AT + REACT [20]	TIP '24	Faster R-CNN	51.4	57.9	67.4	37.7	58.4	52.8	44.6	54.6	53.1
AQT [16]	IJCAI '22	Def DETR	49.3	52.3	64.4	27.7	53.7	46.5	36.0	46.4	47.1
O ² Net [13]	ACM MM '22	Def DETR	48.7	51.5	63.6	31.1	47.6	47.8	38.0	45.9	46.8
MTTrans [53]	ECCV '22	Def DETR	47.7	49.9	65.2	25.8	45.9	33.9	32.6	43.5	46.8
DA-DETR [56]	CVPR '23	Def DETR	49.9	50.0	63.1	24.0	45.8	37.5	31.6	46.3	43.5
BLADT [14]	ICCV '23	Def DETR	52.2	58.9	69.2	31.7	55.0	45.1	42.6	51.3	50.8
MRT [59]	ICCV '23	Def DETR	52.8	51.7	68.7	35.9	58.1	54.5	41.0	47.1	51.2
MTM [46]	AAAI '24	Def DETR	51.0	53.4	67.2	37.2	54.4	41.6	38.4	47.7	48.9
ACCT [55]	ICASSP '24	Def DETR	53.6	58.9	69.4	31.1	53.5	33.7	42.6	54.4	49.6
KTNet [38]	Neural Networks '23	RetinaNet	43.7	48.9	56.9	34.9	51.8	<u>55.3</u>	32.8	42.8	45.9
MGCAMT [5]	TIP '25	RetinaNet	<u>60.2</u>	<u>66.6</u>	76.5	33.2	60.1	43.2	<u>49.8</u>	<u>57.9</u>	<u>55.9</u>
Source Only	-	FCOS	38.5	41.8	46.0	28.6	42.1	11.0	31.1	42.2	35.1
SIGMA [22]	CVPR '22	FCOS	44.0	43.9	60.3	31.6	50.4	51.5	31.7	40.6	44.2
OADA [49]	ECCV '22	FCOS	47.8	46.5	62.9	32.1	48.5	50.9	34.3	39.8	45.4
SIGMA++ [23]	TPAMI '23	FCOS	46.4	45.1	61.0	32.1	52.2	44.6	34.8	39.9	44.5
CIGAR [29]	CVPR '23	FCOS	46.1	47.3	62.1	27.8	56.6	44.3	33.7	41.3	44.9
HT [9]	CVPR '23	FCOS	52.1	55.8	67.5	32.7	55.9	49.1	40.1	50.3	50.4
LSCDIGM [47]	J SUPERCOMPUT '25	FCOS	44.7	43.1	60.9	28.9	48.3	48.0	36.7	37.7	43.5
DSTS [10]	Information '25	FCOS	51.4	54.3	68.4	37.8	54.6	40.2	42.4	51.8	50.1
DSBCO (Ours)	-	FCOS	61.1	66.8	<u>73.0</u>	58.1	72.3	62.2	58.8	65.6	64.7

metric with an Intersection-over-Union (IoU) threshold of 0.5. For experiments involving the Cityscapes, Foggy Cityscapes, and BDD100K datasets, we report the mAP across all shared semantic categories. In the Synthetic Source Adaptation and Distinct Camera Adaptation scenarios, we report the Average Precision (AP) for the common category Car. This rigorous evaluation protocol ensures a fair and direct comparison with prior methodologies.

Implementation Details and ST Baseline. We implement FCOS with a VGG-16 backbone [36], optimized via SGD with a learning rate of 0.004. Training consists of 30k pre-training iterations on the source data $\mathcal{D}_s$, followed by 90k adaptation iterations on a single RTX 3090 GPU. For the ST baseline, the pre-trained detector generates pseudo-labels for target data $\mathcal{D}_t$, filtering predictions with confidence scores below 0.45. The model is then iteratively retrained on both the labeled source and pseudo-labeled target data. Further implementation details are provided in Supplementary **Section C**.

4.2 Main Results

Quantitative evaluations across four cross domain scenarios demonstrate that DSBCO effectively transcends the generalization error bound. In Adverse Weather Adaptation (Table 1), DSBCO effectively achieves the best mAP of 64.7%, outperforming the second-best overall method (MGCAMT) [5] by 8.8%, and exceeding the FCOS-based second-best (HT) [9] by 14.3%. Significant gains in rigid structures across like Bus (72.3%) confirm the regression cycle stream mit-

Table 2: Quantitative results on **Diverse Context Adaptation** (Cityscapes $\rightarrow$ BDD100K). The mean Average Precision (%) is evaluated on the target domain. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	Detector	Person	Rider	Car	Truck	Bus	Motor	Bicycle	mAP $\uparrow$
TDD [15]	CVPR '22	Faster R-CNN	39.6	38.9	53.9	24.1	25.5	24.5	28.8	33.6
PT [6]	ICML '22	Faster R-CNN	40.5	39.9	52.7	25.8	<u>33.8</u>	23.0	28.8	34.9
AT + REACT [20]	TIP '24	Faster R-CNN	-	-	-	-	-	-	-	35.8
AQT [16]	IJCAI '22	Def DETR	38.2	33.0	58.4	17.3	18.4	16.9	23.5	29.4
O ² Net [13]	ACM MM '22	Def DETR	40.4	31.2	58.6	20.4	25.0	14.9	22.7	30.5
MTTrans [53]	ECCV '22	Def DETR	44.1	30.1	61.5	25.1	26.9	23.0	17.7	32.6
BIADT [14]	ICCV '23	Def DETR	42.1	34.0	60.9	17.4	19.5	18.2	25.7	33.6
MRT [59]	ICCV '23	Def DETR	48.4	30.9	63.7	24.7	25.5	20.2	22.6	33.7
MTM [46]	AAAI '24	Def DETR	53.7	35.1	<u>68.8</u>	23.0	28.8	23.8	28.0	37.3
ACCT [55]	ICASSP '24	Def DETR	51.8	41.4	61.8	26.0	23.4	36.9	31.7	39.0
MGCAMT [5]	TIP '25	RetinaNet	61.2	44.9	75.3	32.5	30.0	<u>31.9</u>	<u>37.8</u>	44.8
Source Only	-	FCOS	36.9	22.4	49.7	16.1	16.3	13.0	22.1	25.2
SIGMA [22]	CVPR '22	FCOS	46.9	29.6	64.1	20.2	23.6	17.9	26.3	32.7
HT [9]	CVPR '23	FCOS	53.4	40.4	63.5	27.4	30.6	28.2	38.0	40.2
SIGMA++ [23]	TPAMI '23	FCOS	47.5	30.4	65.6	21.1	26.3	27.1	17.8	33.7
DSBCO (Ours)	-	FCOS	54.8	38.2	67.9	<u>32.3</u>	35.1	29.3	35.7	<u>41.9</u>

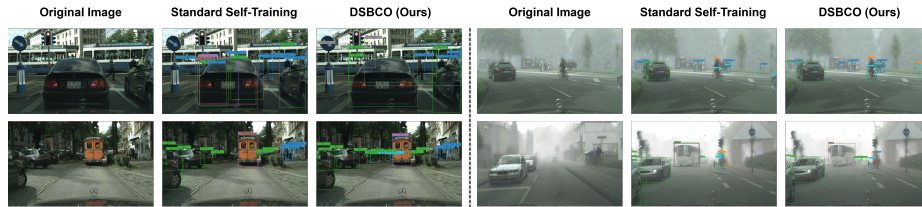


Fig. 3: Qualitative comparative detection results. DSBCO shows effective and robust improvement in resolving severe bounding box shifts and misclassifications.

igates visibility-induced localization drift. In Diverse Context Adaptation (Table 2), our method establishes the best FCOS-based performance with 41.9% mAP, improving approximately 2% over HT under challenging scenarios. For Distinct Camera Adaptation (Table 3), DSBCO attains a leading AP of 62.2%, surpassing HT by approximately 2% in challenging settings. Finally, in Synthetic Source Adaptation (Table 4), DSBCO reaches the highest AP of 68.6%, showing improvement over MGCAMT, while outperforming HT by 3.1%. Overall, these results confirm that the bilevel optimization strategy successfully aligns distributions to prevent mode collapse results.

4.3 Computational Efficiency

Standard bilevel optimization is often too slow and unstable for object detection, as numerous repeatedly constructed large Hessian matrices from dense bounding-box samples can ultimately cause training collapse. Unlike the original CST approach [28], DSBCO improves efficiency and scalability by using compact feature-based covariance matrix estimation instead of sample-based matrix

Table 3: Quantitative results on **Distinct Camera Adaptation** (KITTI $\rightarrow$ Cityscapes). Average Precision (%) is evaluated on the target domain. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	Detector	AP of Car $\uparrow$
TDD [15]	CVPR '22	Faster R-CNN	47.4
PT [6]	ICML '22	Faster R-CNN	60.2
AT + DSD-DA [11]	ICML '24	Faster R-CNN	49.3
AT + REACT [20]	TIP '24	Faster R-CNN	59.5
DA-DETR [56]	CVPR '23	Def DETR	48.9
MGCAMT [5]	TIP '25	RetinaNet	62.2
Source Only	-	FCOS	34.4
SIGMA [22]	CVPR '22	FCOS	45.8
OADA [49]	ECCV '22	FCOS	47.8
CIGAR [29]	CVPR '23	FCOS	48.5
HT [9]	CVPR '23	FCOS	<u>60.3</u>
LSCDIGM [47]	J SUPERCOMPUT '25	FCOS	50.0
DSTS [10]	Information '25	FCOS	50.9
DSBCO (Ours)	-	FCOS	62.2

Table 5: Comparison of computational complexity and training time cost. Naive Bilevel refers to standard bilevel optimization, requiring M internal gradient steps (typically $M \gg 1$, *e.g.*, $M \approx 50$) per outer update. DSBCO bypasses these iterative inner loops for high efficiency. Here, N is the number of samples, and T is the unit training time of Standard ST.

Optimization Strategy	Complexity	Time	Cost
Standard ST	$\mathcal{O}(N)$	T	
Naive Bilevel Optimization	$\mathcal{O}(N \cdot M)$	$\approx M \cdot T$	
DSBCO (Ours)	$\mathcal{O}(N)$	$\approx 1.2 \cdot T$	

Table 4: Quantitative results on **Synthetic Source Adaptation** (Sim10K $\rightarrow$ Cityscapes). Average Precision (%) is evaluated on the target domain. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	Detector	AP of Car $\uparrow$
D_adapt [17]	ICLR '22	Faster R-CNN	50.3
TDD [15]	CVPR '22	Faster R-CNN	53.4
PT [6]	ICML '22	Faster R-CNN	55.1
AT + DSD-DA [11]	ICML '24	Faster R-CNN	52.5
MT + DAS [52]	NeurIPS '24	Faster R-CNN	55.4
AT + REACT [20]	TIP '24	Faster R-CNN	58.6
O ² Net [13]	ACM MM '22	Def DETR	54.1
DA-DETR [56]	CVPR '23	Def DETR	54.7
BIADT [14]	ICCV '23	Def DETR	56.6
MRT [59]	ICCV '23	Def DETR	62.0
MTM [46]	AAAI '24	Def DETR	58.1
MGCAMT [5]	TIP '25	RetinaNet	67.5
Source Only	-	FCOS	39.8
SIGMA [22]	CVPR '22	FCOS	53.7
OADA [49]	ECCV '22	FCOS	59.2
CIGAR [29]	CVPR '23	FCOS	58.5
HT [9]	CVPR '23	FCOS	65.5
LSCDIGM [47]	J SUPERCOMPUT '25	FCOS	56.2
DSTS [10]	Information '25	FCOS	60.4
DSBCO (Ours)	-	FCOS	68.6

Table 6: Ablation study on **Adverse Weather Adaptation** (Cityscapes $\rightarrow$ Foggy Cityscapes). This table provides a comprehensive and incremental evaluation to analyze the individual contribution of Instance Consistency ($\mathcal{L}_{distill}$), the Classification Cycle ($\mathcal{L}_{cstcls}$), and the Regression Cycle ($\mathcal{L}_{cstreg}$). The mean Average Precision (%) (mAP) is evaluated on the target domain.

Method	Baseline	$\mathcal{L}_{distill}$	$\mathcal{L}_{cstcls}$	$\mathcal{L}_{cstreg}$	mAP $\uparrow$
Standard ST	✓				41.2
w/ Instance Consistency	✓	✓			48.3
w/ Classification Cycle	✓		✓		54.5
w/ Regression Cycle	✓			✓	50.9
DSBCO (Ours)	✓	✓	✓	✓	64.7

inversion. This avoids substantial high memory costs and instability from large-matrix inversion. By using a linear solver and regression normalization, DSBCO skips heavy Hessian computations. As shown in Table 5, DSBCO maintains training speeds similar to standard ST while ensuring robust convergence.

4.4 Ablation Study

Table 6 evaluates DSBCO components on the Cityscapes $\rightarrow$ Foggy Cityscapes adaptation. The Standard ST baseline yields 41.2% mAP. Adding Instance Consistency ($\mathcal{L}_{distill}$) provides an effective and stable improvement, boosting performance to 48.3%. Building upon this, individually incorporating the Classification ($\mathcal{L}_{cstcls}$) or Regression ($\mathcal{L}_{cstreg}$) Cycle Stream elevates accuracy to 54.5% and

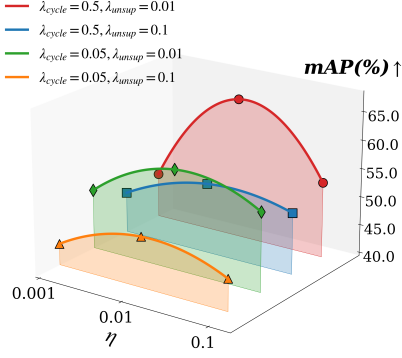


Fig. 4: Hyperparameter sensitivity analysis of our proposed DSBCO.

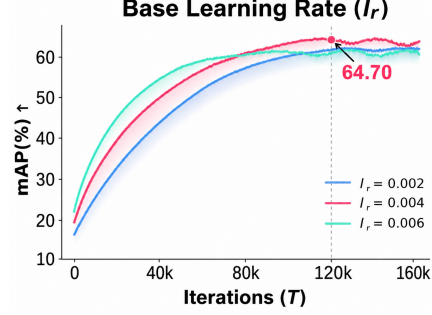


Fig. 5: Hyperparameter sensitivity analysis of the base learning rate (l_r).

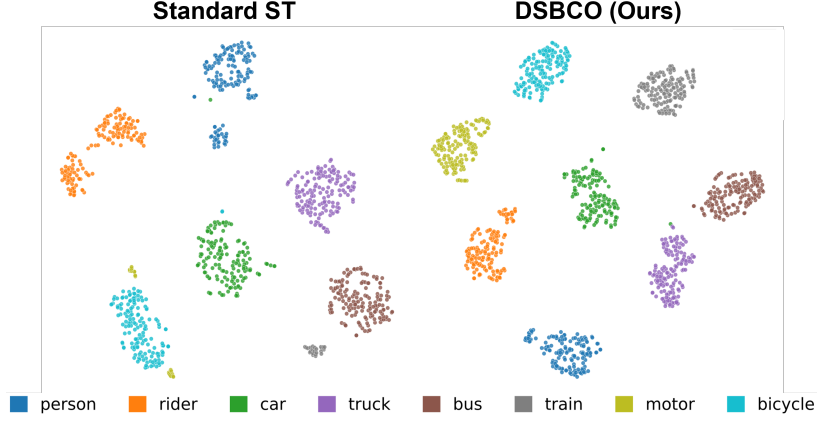


Fig. 6: The t-SNE visualization of feature representations. Compared with Standard ST, DSBCO produces tighter clusters and clearer boundaries between different classes.

50.9%, respectively. Integrating all components jointly achieves the best 64.7% mAP. As shown in Fig. 4, this optimal overall performance is reached using $\lambda_{cycle} = 0.5$, $\lambda_{unsup} = 0.01$, and $\eta = 0.01$. Sensitivity results in Fig. 5 further show that DSBCO obtains the best performance at $l_r = 0.004$. The complete hyperparameter analysis is provided in Supplementary **Section C**.

4.5 Visualization Analysis

Qualitative results are shown in Fig. 3. DSBCO reduces false positives and improves detection across diverse scenes. Compared with the baseline, which often misses occluded objects or produces box misalignment, our framework yields more precise predictions with higher confidence. This indicates that DSBCO

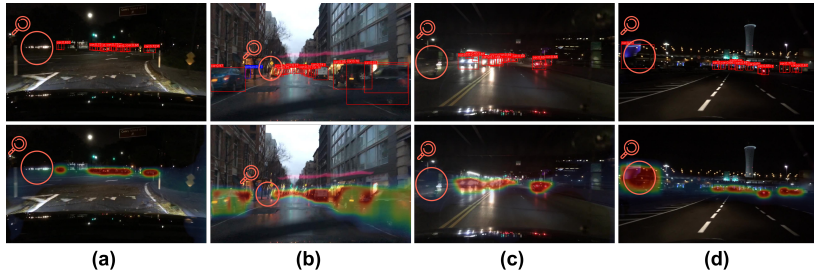


Fig. 7: Case analysis of DSBCO under the Diverse Context Adaptation scenario.

stabilizes training and enhances model robustness. Furthermore, Fig. 6 visualizes feature embeddings via t-SNE [31]. Compared with Standard ST, DSBCO yields compact clusters for Person and Rider, with clearer separations between Bicycle and Motor, and Train and Bus in embedding space.

Finally, Fig. 7 presents failure cases and heatmaps under Diverse Context Adaptation. Case (a) shows false negatives for distant cars at night under low visibility, Case (b) shows missed nearby persons under daytime motion blur, Case (c) shows missed nearby cars during fast nighttime driving with severe blur, and Case (d) shows a night-time failure where Train is incorrectly classified as Car, indicating that these four failure cases require further investigation.

5 Conclusion and Limitations

We propose DSBCO to overcome generalization barriers in domain adaptation by reformulating it as a bilevel optimization problem. DSBCO utilizes regression normalization to effectively enforce cycle consistency and resolve regression instability. Empirical results confirm that DSBCO consistently outperforms competitive methods across diverse adaptation scenarios. However, several limitations still exist. First, extending this dual-stream optimization to modern architectures like Transformers requires further study. Second, the fixed thresholding mechanism may struggle against complex background noise in challenging scenarios, prompting us to explore more adaptive strategies in future work.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62306343, U2541229, and 62441619), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515011322), and the Shenzhen Science and Technology Program (Grant Nos. KQTD20221101093559018 and SYSRD20250529113401002).

References

1. Amini, M.R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., Maximov, Y.: Self-training: A survey. *Neurocomputing* **616**, 128904 (2025)
2. Bi, J., Dornaika, F., Charafeddine, J.: A re-node self-training approach for deep graph-based semi-supervised classification on multi-view image data. *Information Fusion* **127**, 103887 (2026)
3. Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C., Tu, Z., Soatto, S.: Exponential moving average normalization for self-supervised and semi-supervised learning. In: *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*. pp. 194–203. IEEE, Nashville, TN, USA (2021)
4. Cao, S.C., Joshi, D., Gui, L.Y., Wang, Y.X.: Contrastive mean teacher for domain adaptive object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*. pp. 23839–23848. IEEE, Vancouver, BC, Canada (2023)
5. Chen, J., Liu, L., Deng, W., Liu, Z., Liu, Y., Wei, Y., Liu, Y.: Refining pseudo labeling via multi-granularity confidence alignment for unsupervised cross domain object detection. *IEEE Transactions on Image Processing* **34**, 279–294 (2025)
6. Chen, M.L., Chen, W.J., Yang, S.C., Song, J., Wang, X.C., Zhang, L., Yan, Y.F., Qi, D.L., Zhuang, Y.T., Xie, D., Pu, S.L.: Learning domain adaptive object detection with probabilistic teacher. In: *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*. pp. 3040–3055. PMLR, Baltimore, MD, USA (2022)
7. Chen, R., Zhang, H., Li, J., Liu, L., Huang, Z., Cao, X.: Generalized semantic contrastive learning via embedding side information for few-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(8), 6496–6514 (2025)
8. Cordts, M., Enzweiler, M., Omran, M., Benenson, R., Ramos, S., Franke, U., Roth, S., Rehfeld, T., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. pp. 3213–3223. IEEE, Las Vegas, NV, USA (2016)
9. Deng, J., Xu, D., Li, W., Duan, L.: Harmonious teacher for cross-domain object detection. In: *Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)*. pp. 23829–23838. IEEE, Vancouver, BC, Canada (2023)
10. Dong, S., Deng, K., Zou, K.: Reconstructing domain-specific features for unsupervised domain-adaptive object detection. *Information* **16**(6), 439 (2025)
11. Feng, Y.C., Li, S.W., Gao, Y.J., Huang, Z.Y., Zhang, Y.N., Liu, Q.J., Wang, Y.H.: Dsd-da: Distillation-based source debiasing for domain adaptive object detection. In: *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. pp. 1–16. PMLR, Vienna, Austria (2024)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*. pp. 3354–3361. IEEE, Providence, RI, USA (2012)
13. Gong, K.X., Li, S., Li, S.G., Zhang, R., Liu, C.H., Chen, Q.: Improving transferability for domain adaptive detection transformers. In: *Proceedings of the 30th ACM International Conference on Multimedia (MM 2022)*. pp. 1543–1551. ACM, Lisboa, Portugal (2022)

14. He, L., Wang, W., Chen, A., Sun, M., Kuo, C.H., Todorovic, S.: Bidirectional alignment for domain adaptive detection with transformers. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV 2023). pp. 18729–18739. IEEE, Paris, France (2023)
15. He, M.Z., Wang, Y.L., Wu, J.X., Wang, Y.R., Li, H.Q., Li, B., Gan, W.H., Wu, W., 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 (CVPR 2022). pp. 9560–9570. IEEE, New Orleans, LA, USA (2022)
16. Huang, W.J., Lu, Y.L., Lin, S.Y., Xie, Y.S., Lin, Y.Y.: Aqt: Adversarial query transformers for domain adaptive object detection. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022). pp. 972–979. IJCAI, Vienna, Austria (2022)
17. Jiang, J.G., Chen, B.X., Wang, J.M., Long, M.S.: Decoupled adaptation for cross-domain object detection. In: Proceedings of the 10th International Conference on Learning Representations (ICLR 2022). pp. 1–16. OpenReview.net, Virtual (2022)
18. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In: Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA 2017). pp. 746–753. IEEE, Singapore (2017)
19. Krishna, O., Ohashi, H., Sinha, S.: Mila: Memory-based instance-level adaptation for cross-domain object detection. In: Proceedings of the 34th British Machine Vision Conference (BMVC 2023). pp. 1–13. BMVA Press, London, UK (2023)
20. Li, H., Zhang, R., Yao, H., Zhang, X., Hao, Y., Song, X., Li, L.: React: Remainder adaptive compensation for domain adaptive object detection. *IEEE Transactions on Image Processing* **33**, 3735–3748 (2024)
21. Li, J., Ren, F., Liang, M., Yi, J., Cao, F.: Cross domain object detection with multi-level domain feature refinement. *Computing* **108**(2), 1–20 (2026)
22. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). pp. 5281–5290. IEEE, New Orleans, LA, USA (2022)
23. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(7), 9022–9040 (2023)
24. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2020). pp. 21002–21012. Curran Associates, Inc., Virtual (2020)
25. Li, Y.J., Dai, X.L., Ma, C.Y., Liu, Y.C., Wu, B., He, Z., Kitani, K., Vajda, P., Chen, K.: Cross-domain adaptive teacher for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). pp. 7581–7590. IEEE, New Orleans, LA, USA (2022)
26. Li, Z., Yang, Y., Wang, L., Xu, M., Xin, Z., Wei, J., Wang, Y.: Cycle self-training with joint adversarial for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 5531117 (2024)
27. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017). pp. 2999–3007. IEEE, Venice, Italy (2017)

28. Liu, H., Wang, J., Long, M.: Cycle self-training for domain adaptation. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). pp. 1–31. Curran Associates, Inc., Virtual (2021)
29. Liu, Y., Wang, J., Huang, C., Wang, Y., Xu, Y.: Cigar: Cross-modality graph reasoning for domain adaptive object detection. In: Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). pp. 23776–23786. IEEE, Vancouver, BC, Canada (2023)
30. Liu, Y., Zhang, Z., Su, Y., Hao, C., Liu, X., Tian, J.: Unsupervised domain adaptive object detection via semantic consistency and compactness learning. *IEEE Transactions on Image Processing* **35**, 2276–2291 (2026)
31. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008)
32. Qu, J., Ren, J., Dong, W., Xiao, S., Li, Y.: Cycle-based frequency disentanglement diffusion model with self-training for cross-domain hyperspectral-rgb change detection. *IEEE Transactions on Image Processing* **34**, 8130–8144 (2025)
33. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NeurIPS 2015). pp. 91–99. MIT Press, Montreal, Canada (2015)
34. Rezaatofghi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). pp. 658–666. IEEE, Long Beach, CA, USA (2019)
35. Sakaridis, C., Dai, D., Gool, L.V.: Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* **126**(3), 973–992 (2018)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). pp. 1–14. CBLIS, San Diego, CA, United states (2015)
37. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017). pp. 1195–1204. Curran Associates, Inc., Long Beach, CA, USA (2017)
38. Tian, K., Zhang, C., Wang, Y., Xiang, S.: Domain adaptive object detection with model-agnostic knowledge transferring. *Neural Networks* **161**, 213–227 (2023)
39. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019). pp. 9626–9635. IEEE, Seoul, South Korea (2019)
40. Tingting, Z., Youjun, W.: A survey on domain adaptive object detection. *The Computer Journal* **69**(2), 357–375 (2026)
41. Wang, K., Shen, Y., Lauer, M.: Adversarial defense teacher for cross-domain object detection under poor visibility conditions. In: arXiv:2403.15786v1. pp. 1–20 (2024)
42. Wang, T., Zhang, X., Yuan, L., Feng, J.: Few-shot adaptive faster r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). pp. 7166–7175. IEEE, Long Beach, CA, USA (2019)
43. Wang, W., Li, H., Wang, C., Huang, C., Ding, Z., Nie, F., Cao, X.: Deep label propagation with nuclear norm maximization for visual domain adaptation. *IEEE Transactions on Image Processing* **34**, 1246–1258 (2025)
44. Wang, W., Wang, M., Huang, C., Wang, C., Mu, J., Nie, F., Cao, X.: Optimal graph learning-based label propagation for cross-domain image classification. *IEEE Transactions on Image Processing* **34**, 1529–1544 (2025)

45. Wang, X., Yang, X., Zhang, S., Li, Y., Feng, L., Fang, S., Lyu, C., Chen, K., Zhang, W.: Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). pp. 3240–3249. IEEE, Vancouver, Canada (2023)
46. Weng, W.X., Yuan, C.: Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2024). pp. 5912–5920. Association for the Advancement of Artificial Intelligence, Vancouver, Canada (2024)
47. Xu, Z., Qiu, Y., Zhang, J.: Learnable semantic completion and distance-induced graph matching for domain adaptive object detection. *The Journal of Supercomputing* **81**(5), 716–741 (2025)
48. Yao, H., Zhao, S., Lu, S., Chen, H., Li, Y., Liu, G., Xing, T., Yan, C., Tao, J., Ding, G.: Source-free object detection with detection transformer. *IEEE Transactions on Image Processing* **34**, 5948–5963 (2025)
49. Yoo, J., Chung, I., Kwak, N.: Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In: Proceedings of the 17th European Conference on Computer Vision (ECCV 2022). pp. 697–708. Springer, Tel Aviv, Israel (2022)
50. Yoon, I., Kwon, H., Kim, J., Park, J., Jang, H., Sohn, K.: Enhancing source-free domain adaptive object detection with low-confidence pseudo label distillation. In: Proceedings of the 18th European Conference on Computer Vision (ECCV 2024). pp. 337–353. Springer, Milan, Italy (2024)
51. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020). pp. 2633–2642. IEEE, Seattle, USA (2020)
52. Yu, H., Deng, J., Li, W., Duan, L.: Towards unsupervised model selection for domain adaptive object detection. In: Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024). pp. 1–13. Curran Associates, Inc., New Orleans, LA, USA (2024)
53. Yu, J., Liu, J., Wei, X., Zhou, H., Nakata, Y., Gudovskiy, D.A., Okuno, T., Li, J., Keutzer, K., Zhang, S.: MTTrans: Cross-domain object detection with mean teacher transformer. In: European Conference on Computer Vision (ECCV 2022). pp. 629–645. Springer, Tel Aviv, Israel (2022)
54. Zadeh, S.G., Schmid, M.: Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(9), 3126–3137 (2021)
55. Zeng, Z., Ding, Y., Lu, H.: Enhancing cross-domain detection: Adaptive class-aware contrastive transformer. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024). pp. 6670–6674. IEEE (2024)
56. Zhang, J.Y., Huang, J.X., Luo, Z.P., Zhang, G.J., Zhang, X.Q., Lu, S.J.: Da-detr: Domain adaptive detection transformer with information fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). pp. 23787–23798. IEEE, Vancouver, BC, Canada (2023)
57. Zhang, S., Zhang, X., Xu, Y., Wang, K.: Few-shot object detection on remote sensing images based on decoupled training, contrastive learning, and self-training. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **19**, 3983–3997 (2026)

58. Zhang, Y., Ma, D., Wu, X.: Cluster-filter-based pseudo-label refinement for source-free domain adaptation fundus image segmentation. *Pattern Recognition* **175**, 112996 (2026)
59. Zhao, Z.J., Wei, S.T., Chen, Q.C., Li, D.H., Yang, Y.F., Peng, Y.X., Liu, Y.: Masked retraining teacher-student framework for domain adaptive object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023)*. pp. 18993–19003. IEEE, Paris, France (2023)
60. Zheng, L., Zhu, H., Wang, X., Li, X., Li, T., Yan, Y.: Pseudo-labeling based unsupervised domain adaptation for llm-based asr. *IEEE Transactions on Audio, Speech and Language Processing* **34**, 808–819 (2026)
61. Zhou, H., Jiang, F., Lu, H.: Ssda-yolo: Semi-supervised domain adaptive yolo for cross-domain object detection. *Computer Vision and Image Understanding* **229**, 103649 (2023)
62. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. pp. 1–16. ICLR, Virtual (2021)