

# Debiased Teacher for Day-to-Night Domain Adaptive Object Detection

Yiming Cui<sup>1,2\*</sup> Liang Li<sup>2†</sup> Haibing Yin<sup>1</sup> Yuhan Gao<sup>1,3</sup> Yaoqi Sun<sup>4,3†</sup> Chenggang Yan<sup>1</sup>  
<sup>1</sup>Hangzhou Dianzi University <sup>2</sup>Institute of Computing Technology, Chinese Academy of Sciences  
<sup>3</sup>Lishui Institute of Hangzhou Dianzi University <sup>4</sup>Lishui University

{cuiyiming, yhb, yuhangao, cgyan}@hdu.edu.cn, liang.li@ict.ac.cn, sunyq2233@163.com

## Abstract

*Day-to-Night Domain Adaptive Object Detection (DN-DAOD) is a significant challenge due to the low visibility and signal-to-noise ratio at night. Although recent self-training approaches achieve promising results, they fail to address three critical biases: distribution bias, training bias, and confirmation bias. Therefore, we propose a Debiased Teacher to address the above biases from three aspects: domain transforming, representation compensating, and pseudo label calibrating. Concretely, the day-to-night domain transforming module (DN-DT) leverages physical priors to model some key day-night domain differences, thus transforming daytime images into night-like images. Then, the cross-domain representation compensating module (CDRC) selectively mixes objects from nighttime and night-like images to compensate for the model's general representation of nighttime objects. Further, to correct confirmation bias caused by learning from inaccurate pseudo labels, the pseudo label confirmation calibrating module (ConCal) is designed to obtain accurate pseudo labels for better nighttime knowledge learning. Experimental results on three benchmarks demonstrate that our method outperforms current SOTA methods by a large margin.*

## 1. Introduction

Object detection is essential in autonomous driving and intelligent surveillance. Yet, these object detectors trained on well-lit daytime images (*i.e.*, labeled source domain) suffer from poor performance on low-light nighttime images (*i.e.*, unlabeled target domain). This is because the prevalent nighttime conditions adversely affect the quality of camera measurements, such as under/overexposure, motion blur, and dazzle caused by headlamps. Day-to-Night Domain Adaptive Object Detection (DN-DAOD) is proposed to address this problem, which intends to train a detector utiliz-

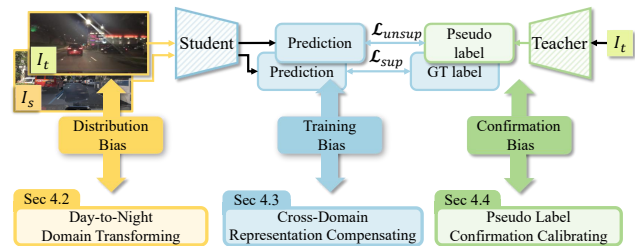


Figure 1. The vanilla self-training framework faces three biases in solving DN-DAOD tasks. Distribution bias in the input data induces a shift in the feature space, weakening the model’s ability to represent nighttime objects. Training bias drives the model’s representation to favor the source domain data and overrepresented classes. Confirmation bias causes the model to reinforce incorrect predictions and fail to self-correct. Thus, we propose the **Debiased Teacher** with three modules to mitigate the above biases.

ing labeled daytime images and unlabeled nighttime images to perform well in the unseen nighttime images.

Recently, the self-training framework achieves promising results on DN-DAOD tasks [25, 28, 67], where a teacher model produces pseudo labels from the target domain to supervise the training of a student model. 2PCNet [28] improves this framework to obtain nighttime pseudo labels combined with high and low confidence. ISP-Teacher [67] utilizes image signal processing (ISP) to model the intrinsic imaging process at night. CoS [25] employs the global-local transformation and a proxy-based target consistency to capture the spatial consistency between day and night. They leverage different techniques to mitigate the domain gap and improve the self-training framework to enhance the feature consistency from both domains, thus significantly improving detection performance.

However, the vanilla self-training framework exists the following three biases, as shown in Fig. 1. **I. Distribution bias** refers to the distribution differences between nighttime and daytime images caused by the domain gap. One type of approaches utilize image transfer to generate night style data [2, 3, 24, 36], while others employ ISP pipelines to synthesize degraded data [12, 17, 67]. Nevertheless, the former

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

†Corresponding authors.

requires training specific models and lacks control, while the latter is sensor-dependent, lacks cross-dataset generalization, and both cannot simulate complex nighttime lighting changes (*e.g.*, the flare phenomenon in the nighttime, as shown in  $I_t$  of Fig. 1). **II. Training bias** indicates the systematic tendency of the self-training framework during the training process, which primarily stems from two factors: 1) As the model initially trains in the source data, the model’s feature representation is biased toward characteristics of the source domain, 2) The class imbalance causes the model’s feature representation to favor more frequent classes, exacerbating the bias. Many works leverage adversarial learning to alleviate source domain bias by enforcing domain invariance [11, 29, 32], but it suppresses the representation of nighttime features, which is beneficial for the student [28]. **III. Confirmation bias** denotes learner struggles to correct its own mistakes when learning from inaccurate pseudo labels [1, 5]. The traditional high-confidence pseudo-labeling process neglects class imbalances, leading to low confidence for underrepresented classes and being filtered out, which amplifies the Matthew Effect<sup>1</sup>. Some methods calculate dynamic thresholds of classification branch for each class [27, 31, 63, 70], while others leverage the localization branch to assist filtering [6, 15, 25, 50]. But they all overlook the potential true positives among the discarded proposals, and relying solely on the teacher model cannot correct the error accumulation caused by confirmation bias.

Therefore, we propose a Debiased Teacher (DeT) to address the inherent three biases in the self-training framework from three aspects: domain transforming, representation compensating, and pseudo label calibrating. Firstly, to reduce the distribution bias between daytime and nighttime images, we introduce a Day-to-Night Domain Transforming module (DNDT). DNDT transforms daytime images into night-like images with similar nighttime distribution by leveraging physical priors to model some key day-night differences, including environmental illumination, night imaging noise, and optical flare effects. Secondly, to mitigate the training bias, we propose the Cross-Domain Representation Compensating module (CDRC), which employs bidirectional mixing between target and night-like objects during training to compensate for the model’s general representation of nighttime objects. Additionally, we incorporate an Inverse Class Frequency Balancing submodule (ICFB) to improve the frequency of underrepresented classes during mixing, ensuring a more balanced and robust detection performance across all classes. Thirdly, to correct confirmation bias, we design the Pseudo Label Confirmation Calibrating module (ConCal), which dynamically adjusts the class-specific thresholds based on the teacher model’s confidence scores to handle the influence of changing class

<sup>1</sup>Matthew Effect: the accuracy of well-behaved classes is further increased while that of poorly-behaved ones is decreased.

distributions. ConCal jointly distills the discarded proposals through both the teacher and student to obtain accurate pseudo labels, thereby minimizing bias accumulation. Our DeT effectively aligns the day-night distribution, compensates for the general representation of nighttime objects across each class, and transfers more reliable nighttime knowledge for the student model. Extensive experiments on three benchmarks demonstrate the extraordinary nighttime adaptation ability of our DeT.

**Our key contributions are:** (1) We propose a Debiased Teacher to address the three biases in the self-training framework for DN-DAOD: distribution bias, training bias, and confirmation bias. (2) We introduce DNDT as a physical prior to model the day-night distribution bias. Then, we propose CDRC to enhance the model’s feature representation for nighttime objects, thus preventing the model from being biased toward the source domain and overrepresented classes. Furthermore, we design ConCal with class-specific threshold and joint distillation to generate accurate pseudo labels that correct the confirmation bias. (3) Extensive experiments demonstrate that our method outperforms all existing SOTA on three benchmarks.

## 2. Related Work

The rapid progress in deep learning has significantly advanced computer vision [16, 20, 44, 45, 49, 52–54, 59, 66, 71, 72] and multimedia understanding [9, 10, 33, 35, 42, 55–58, 60, 62, 65, 68, 69], particularly in Domain Adaptive Object Detection (DAOD). Next, we systematically review representative methods from the following perspectives:

**Pseudo-labeling Process.** Many methods propose specific pseudo-labeling processes to improve the quality of pseudo labels. GbR [31] leverages the gradients from each class to balance the class distribution. ACRST [63] introduces a multi-label classifier to generate additional coarse-grained pseudo labels. RCI [27] adjusts a uniform threshold based on the background and foreground confidence scores. Although the above methods aim to filter proposals from a classified perspective, they ignore the false positives introduced by inaccurate localization. Therefore, PT [6] leverages uncertainty-guided consistency training to promote classification and localization. However, the above methods only perform subtraction, *i.e.*, filtering out a portion of proposals, but this does not guarantee that the discarded proposals do not contain true positives.

**Feature Representation Enhancing** Enhancing the feature representation ability without changing the network structure is studied by many researchers. Several methods leverage masking techniques to improve the model’s context representation ability [11, 23]. SOCCER [11] introduces complementary masks consistency learning to enhance models’ context representation. Other methods employ the mixing strategy to improve the cross-domain representation,

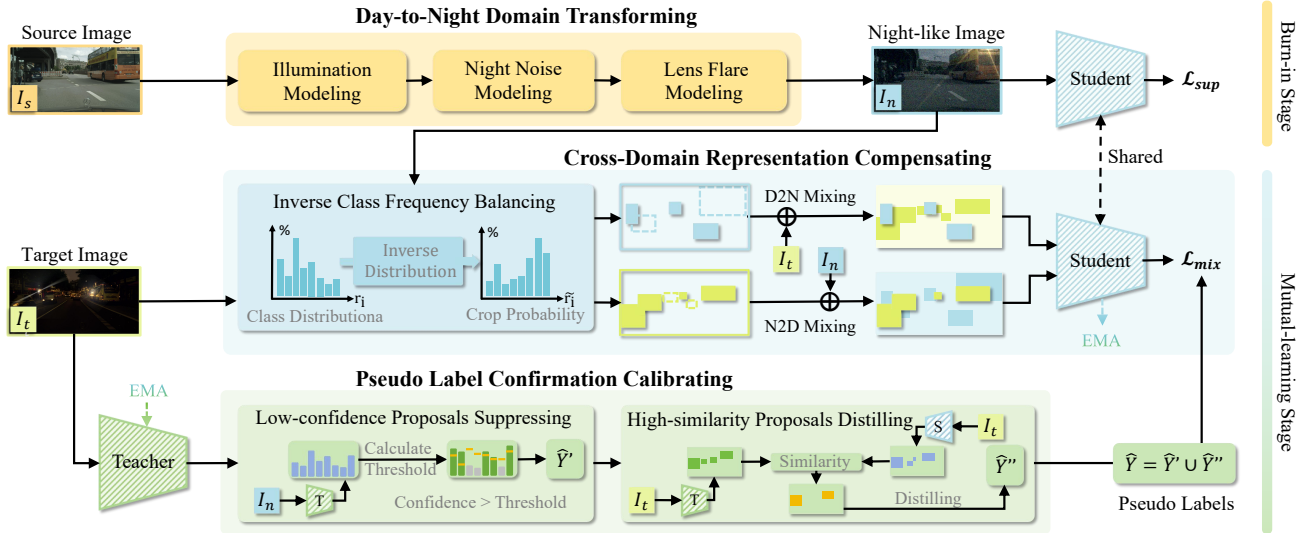


Figure 2. **Overview of our proposed Debiased Teacher (DeT)**, which comprises three modules: 1) Day-to-Night Domain Transforming (DNDT) utilizes physical priors to model some key day-night differences. 2) Cross-Domain Representation Compensating (CDRC) bidirectionally and selectively mixes both domain data to compensate for the general nighttime feature representation and alleviate the model’s systemic tendency of the source domain and overrepresented classes. 3) Pseudo Label Confirmation Calibrating (ConCal) leverages class-specific thresholds and teacher-student joint prediction to obtain accurate pseudo labels, thereby correcting confirmation bias.

which can be primarily categorized into two types: soft mixing [18, 21, 29, 64] and hard mixing [30, 37, 43, 51]. The former generates mixed images through pixel-wise weighted averaging of two images, while the latter extracts patches from other images and pastes them to the current image. CAT [29] uses an Inter-Class Relation module to identify majority and minority classes and employs Mix-up to blend minority classes with highly similar majority classes. Due to the huge domain gap between day and night, forcibly applying masking or cross-domain mixing can impair the representation of nighttime features and cause class ambiguity issues while generating interpolated labels.

**Day-to-Night Domain Adaptive Object Detection.** As night driving is integral to daily life, researchers increasingly focus on DN-DAOD tasks [25, 28, 67]. 2PCNet [28] proposes the NightAug with a series of data augmentations to simulate the nighttime characteristics. ISP-Teacher [67] introduces the ISP pipeline and predicts the ISP-related parameters as a self-supervised learning task. CoS[25] designs the GLT module to further categorize night augmentation into image-level and box-level. Different from them, our DNDT module leverages parameter-free physical priors for night modeling and it not only reduces the day-night distribution bias but also serves as a prerequisite for the execution of the subsequent two modules.

### 3. Preliminaries

#### 3.1. Problem Formulation

Given a labeled daytime source domain  $\mathcal{D}_s = (I_s, B_s, C_s)$  and an unlabeled nighttime target domain  $\mathcal{D}_t = (I_t)$ , where

$I_s$  and  $I_t$  present  $N_s$  source images and  $N_t$  target images,  $B_s = \{b_s^i\}_{i=1}^{N_s}$  denotes the bounding box annotations and  $C_s = \{c_s^i\}_{i=1}^{N_s}$  denotes corresponding class labels. We aim to train a domain adaptive detector with  $\mathcal{D}_s$  and  $\mathcal{D}_t$  to perform effectively within the unseen nighttime target domain.

#### 3.2. Self-training Framework

We use the self-training framework comprising student and teacher models that share identical architectures and network parameters. For a fair comparison with other methods [11, 25, 28, 67], both student and teacher models are Faster RCNN [40] object detectors.

In the burn-in stage, the student model acquires the daytime knowledge from the source domain by supervised training:

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

where  $C'_s$  and  $B'_s$  denote classes and bounding boxes in  $I_s$  predicted by the student model and  $Y_s = (B_s, C_s)$  denotes the ground truth.  $\mathcal{L}_{cls}$  and  $\mathcal{L}_{reg}$  denote the classification loss and regression loss of Faster RCNN.

In the mutual-learning stage, by forcing consistency constraints between the predictions of student and teacher models, the student model can learn the knowledge of the nighttime target domain from the pseudo labels  $\hat{Y}$ :

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

where  $C'_t$  and  $B'_t$  denote classes and bounding boxes in  $I_t$  predicted by the student model,  $\hat{C}$  and  $\hat{B}$  denote classes and bounding boxes in teacher model generated pseudo labels.

The teacher model is updated by Exponential Moving Average (EMA) from the weights of the student model without gradient accumulation:

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

where  $\theta_t$  and  $\theta_s$  denote the model parameters of teacher and student respectively, and  $\varphi$  is an update hyper-parameter.

## 4. Methodology

### 4.1. Overview

As illustrated in Fig. 2, we propose the Debiased Teacher (DeT) with three modules to address the above biases. Concretely, in Sec. 4.2, we first introduce the distribution bias, and then we propose DNDT to transform night-like images by modeling the day-night distribution bias. In Sec. 4.3, we propose CDRC by bidirectionally mixing both domain objects to mitigate the systematic tendency of the self-training framework during the training. Finally, in Sec. 4.4, we design ConCal with two steps to obtain high-quality pseudo labels, thus correcting the confirmation bias.

### 4.2. Day-to-Night Domain Transforming

We define the distribution bias as the difference between daytime and nighttime distributions:

$$\mathcal{P}_{night} = \mathcal{P}_{day} + \mathcal{P}_{bias}, \quad (4)$$

where  $\mathcal{P}_{day}$  and  $\mathcal{P}_{night}$  represent the daytime and nighttime distributions,  $\mathcal{P}_{bias}$  denotes the distribution bias between them. We design the Day-to-Night Domain Transforming module (DNDT) with the following modeling to explicitly model distribution bias  $\mathcal{P}_{bias}$ , thus transforming the daytime images into night-like images.

**Illumination Modeling.** Illumination modeling is used to simulate low-light situations. In low-light environments, the number of photons received by the camera sensor is reduced, resulting in a decrease in the overall brightness of the image. Therefore, we reduce the brightness of source images to realistically simulate this situation. Concretely, we introduce a darkness coefficient  $d$ , drawn from a truncated Gaussian distribution with mean 0.1 and variance 0.08, to simulate illumination changes within the range of (0.01, 1.0) [12, 67]:

$$I_{dark} = I_s \cdot d, \quad d \sim \mathcal{N}(\mu_n = 0.1, \sigma_n^2 = 0.08). \quad (5)$$

**Night Noise Modeling.** In nighttime scenarios, the number of photons received by the sensor is reduced, and the signal-to-noise ratio is reduced. To improve the brightness of the image, the camera increases the sensitivity (ISO), which introduces more electronic noise. Inspired by [12, 46, 47, 67], we classify night imaging noise into two sources: shot noise and read noise. Then invert the RGB image into a RAW

format. Shot noise, due to photon collection uncertainty, is modeled as a Poisson distribution  $\mathcal{P}$ :

$$(C + N_{shot}) \sim \mathcal{P}(C), \quad (6)$$

where  $C$  denotes the number of photons. Read noise, attributed to camera electronics, is approximated by a Gaussian distribution with mean 0 and variance  $\sigma_{read}^2$ :

$$N_{read} \sim \mathcal{N}(0, \sigma_{read}^2). \quad (7)$$

Therefore, we get the noise image by:

$$I_{noise} = I_{dark} + N_{shot} + N_{read}. \quad (8)$$

This process effectively models the noise characteristics of nighttime camera imaging. Subsequently, we convert the RAW format back to RGB to facilitate further processing.

**Lens Flare Modeling.** Due to the diffraction/wave optics effects [34], when a point light passes through imperfect apertures or scratched lenses, it produces glare, streaks, and shimmer instead of a clear light dot [13, 26], which we call flare. Flare is a common visual phenomenon in night driving, but existing approaches ignore the influence of flare on nighttime detection. We construct a flare bank based on Flare7K dataset [13], which contains 7,000 simulated light source samples. We perform Random Affine Transformation on the flare and scale it to the standard size of  $I_{noise}$  to increase the diversity of flares. Then, we apply random brightness coefficient  $d'$  to simulate varying light source intensities, and then perform element-wise addition with  $I_{noise}$  to obtain the final night-like image  $I_n$ :

$$I_n = Clip(\mathcal{T}(I_{flare}) \cdot d' + I_{noise}), \quad (9)$$

where  $I_{flare}$  denotes the original flare image random selected from flare bank,  $\mathcal{T}$  denotes Random Affine Transformation,  $Clip(\cdot)$  denotes clipping the addition to the range of [0, 1],  $d'$  is same to  $d$  but in the range of (0.1, 2.0).

After transforming, the obtained night-like images  $I_n$  can simulate the nighttime distribution  $\mathcal{P}_{night}$ , thereby reducing the distribution bias at the input. In Tab. 6 and Tab. 7 of Sec. 5.4, we provide a detailed analysis of the benefits of night-like images to CDRC and ConCal modules.

### 4.3. Cross-Domain Representation Compensating

To mitigate the training bias, we propose a Cross-Domain Representation Compensating module (CDRC), which comprises the bidirectional mixing strategy and Inverse Class Frequency Balancing (ICFB) submodule. As the student model is initially trained on the source domain, there exists a source bias, which is one source of the training bias. Bidirectional mixing with target and night-like images forces the student to see similarly distributed objects for each iteration, compensating the general representation

of night objects and thus preventing the model’s source domain bias. Another source is the phenomenon of class imbalance that exists in datasets, which continues during training (As shown in Fig. 4 left). Therefore, we propose ICFB to balance the number of trained classes during training.

**Bidirectional Mixing.** In each iteration of the CDRC, the mixed data is generated in both D2N and N2D directions. For D2N mixing, the objects in a night-like image  $I_n$  are cut based on the ground truth  $Y_s$  and pasted onto a target domain image  $I_t$  according to the original position, resulting in a mixed target image  $I_{D \rightarrow N}^M$ . Its corresponding image label  $Y_{D \rightarrow N}^M$  is created by adding the ground truth  $Y_s$  to the generated pseudo labels  $\hat{Y}$ . Similarly, we utilize the pseudo label to obtain the N2D mixing data  $I_{N \rightarrow D}^M$  and  $Y_{N \rightarrow D}^M$ . Bidirectional mixing eliminates model source bias, avoids the impact of redundant background, and enhances the ability of intensive detection. We provide rich ablation studies in Tab. 6 to verify the effectiveness of this strategy.

**Inverse Class Frequency Balancing.** In ICFB, the objects are selectively mixed based on the inverse distribution of different classes by the inverse distribution function:

$$\tilde{r}_i = \frac{(1 - r_i)^\alpha}{\sum_{j=1}^k (1 - r_j)^\alpha}, \quad (10)$$

where  $\alpha$  denotes the balance factor,  $\tilde{r}_i$  denotes the probability that an object of class  $i$  will be crop mixed.  $r_i$  represents the current proportion of class  $i$  in the total number of classes and  $\sum_{i=1}^k r_i = 1$ . By increasing the value of  $\alpha$ , we amplify the advantage of underrepresented classes, ensuring a more balanced representation through the ICFB module (As validation in Fig. 4 and Tab. 5). In ICFB, we count the overall class distribution across both domains to balance the number of mixed classes. The mix is not performed when the IoU between the mixed object and the original object is greater than 0.9 or occludes the original object.

**Training Loss.** Mixed data are used to train the student model with a bidirectional domain mix training loss:

$$\begin{aligned} \mathcal{L}_{mix} = & \lambda_{D \rightarrow N} \cdot \mathcal{L}_{unsup}(I_{D \rightarrow N}^M, B_{D \rightarrow N}^M, C_{D \rightarrow N}^M) \\ & + \lambda_{N \rightarrow D} \cdot \mathcal{L}_{unsup}(I_{N \rightarrow D}^M, B_{N \rightarrow D}^M, C_{N \rightarrow D}^M), \end{aligned} \quad (11)$$

where  $B_{D \Rightarrow N}^M, C_{D \Rightarrow N}^M$  denote the mixed bounding box annotations and corresponding class labels from  $Y_{D \Rightarrow N}^M$ .  $\lambda_{D \Rightarrow N}$  are the weights for balancing losses.

#### 4.4. Pseudo Label Confirmation Calibrating

While traditional high-confidence pseudo-labeling serves as reliable supervision for the student model based on a fixed high-confidence threshold, they are biased toward overly confident proposals or classes. The fixed threshold causes the number of proposals to increase unboundedly and accumulates massive errors. More importantly, classification

confidence can not ensure the positioning accuracy of proposals and can not calibrate the confirmation bias between teacher and student models. To tackle the above problems, we propose a Pseudo Label Confirmation Calibrating module (ConCal) with two steps:

**Low-confidence Proposals Suppressing.** In the first step, we mitigate bias propagation in the pseudo-labeling process by filtering out low-confidence proposals. Concretely, we set a base threshold for each class, then dynamically update the thresholds by leveraging ground truth to statistic the average confidence of correctly predicted classes by the teacher model in night-like images:

$$\delta = \delta_{base} + \beta \cdot \text{softmax}(C_{1:k}^{avg}), \delta \in [\delta_{lower}, \delta_{upper}], \quad (12)$$

where  $\delta_{base}$  denotes base thresholds for all classes,  $C_{1:k}^{avg}$  denotes the average confidence of each class, amplitude factor  $\beta$  is used to control the amplitude of the dynamic threshold, and  $\delta_{lower}, \delta_{upper}$  denote the lower limit and upper limit of  $\delta$ .  $\text{softmax}(\cdot)$  enhances the relative importance of different class average confidence scores, which helps mitigate the impact of class imbalances during the pseudo-labeling process. By utilizing the ground truth of the source domain we can more accurately calculate the confidence of the teacher model for each class. More details and discussion of the rationality and effectiveness can be found in Tab. 7. Then, we use dynamic threshold  $\delta$  to suppress low-confidence proposals:

$$\hat{Y}' = \{(\hat{B}'_i, \hat{C}'_i) | \max(p_i) > \delta_{\hat{C}'_i}\}, \quad (13)$$

where  $\hat{C}'_i = \arg \max(p_i)$  denotes the class corresponding to the maximum class score  $p_i$ .  $\hat{B}'_i$  and  $\hat{C}'_i$  denote the bounding boxes and classes that reach the dynamic threshold  $\delta_{\hat{C}'_i}$ .

**High-similarity Proposals Distilling.** In the second step, to ensure that the framework can correct its confirmation bias, we introduce the student model utilizing the joint predictions to distill potential true positives among the discarded proposals. We distill the high-similarity proposals predicted by the teacher and student models and add them to the pseudo labels, even if their confidence does not reach the dynamic threshold  $\delta$ . The criterion for high-similarity is determined by the following three requirements when the proposal does not reach the corresponding class threshold:

- 1) The value in the IoU matrix exceeds a specified threshold  $\tau$  is considered matched proposals;
- 2) The confidence of this teacher model’s proposal is greater than the minimum confidence requirement  $\delta_{min}$ ;
- 3) The predicted classes of this proposal by both the teacher and student models are consistent.

$$\hat{Y}'' = \left\{ (\hat{B}_i^T, \hat{C}_i^T) \left| \begin{array}{l} \text{IoU}(\hat{B}_i^T, B_i^S) > \tau, \\ \delta_{\hat{C}_i^T} > \max(p_i^T) > \delta_{min}, \\ \hat{C}_i^T = C_i^S \end{array} \right. \right\}, \quad (14)$$

Table 1. Results of BDD100k dataset (daytime → nighttime).

Method	person	rider	car	truck	bus	mcycle	bicycle	t-light	t-sign	mAP
Source [39]	50.0	28.9	66.6	47.8	47.5	32.8	39.5	41.0	56.5	41.1
Oracle [39]	52.1	35.0	73.6	53.5	54.8	36.0	41.8	52.2	63.3	46.2
SADA [8]	46.2	25.1	64.2	35.8	33.9	18.4	25.6	45.7	60.0	39.4
MIC [23]	46.4	21.8	64.8	35.6	35.8	17.3	28.7	47.4	60.9	39.8
2PCNet <sup>2</sup> [28]	54.4	30.8	73.1	53.8	55.2	37.5	44.5	49.5	65.2	46.4
2PCNet [28]	50.2	30.0	68.6	50.3	50.0	31.1	41.0	43.7	63.7	47.6
CoS <sup>2</sup> [25]	50.2	<b>55.6</b>	<b>73.9</b>	41.3	54.7	43.5	<b>57.1</b>	<b>50.0</b>	67.8	49.4
CoS [25]	50.8	32.3	69.1	50.3	47.9	35.9	39.6	45.0	64.3	48.4
SOCCER [11]	48.5	27.9	70.4	43.0	39.8	26.4	32.7	43.0	64.2	44.0
ISP-Teacher <sup>2</sup> [67]	<b>57.8</b>	39.4	72.9	54.6	<b>55.9</b>	43.8	48.1	49.6	66.3	48.8
ISP-Teacher [67]	52.5	33.2	67.7	<b>55.3</b>	49.8	33.7	40.1	43.3	67.2	49.2
Ours	52.8	35.7	63.2	54.8	50.2	<b>45.7</b>	46.0	45.8	<b>68.0</b>	<b>50.9</b>

where  $(\hat{B}_i^T, \hat{C}_i^T)$  and  $(B_i^S, C_i^S)$  are predictions from teacher and student models. Finally, we average the coordinates of the matched high-similarity proposals and distill them into pseudo labels to regenerate final pseudo labels  $\hat{Y}$ :

$$\hat{Y} = \hat{Y}' \cup \left\{ \left( (\hat{B}_i^T + B_i^S) / 2, \hat{C}_i^T \right) \mid (\hat{B}_i^T, \hat{C}_i^T) \in \hat{Y}' \right\}. \quad (15)$$

#### 4.5. Overall Optimization Objective

The overall objective of DeT is:

$$\mathcal{L} = \lambda_{sup} \cdot \mathcal{L}_{sup} + \lambda_{mix} \cdot \mathcal{L}_{mix}, \quad (16)$$

where  $\mathcal{L}_{sup}$  is employed in the burn-in stages and  $\mathcal{L}_{mix}$  is employed in the mutual-learning stage,  $\lambda_{[\cdot]}$  are the weights for balancing objectives. We provide the detailed pseudocode of our training process in *Appendix E*.

## 5. Experiments

### 5.1. Evaluation and Datasets

We use the mean Average Precision (mAP) at the IoU threshold of 0.5 to compare with state-of-the-art methods.

**BDD100k:** BDD100k (B) [61] is a widely used autonomous driving dataset. We split the BDD100k dataset into two parts using the labels “day” and “night”. The source domain contains 36,728 daytime images, the target domain contains 27,971 nighttime images, and we use an additional 3,929 nighttime images for validation.

**SHIFT:** SHIFT (S) [41] is a simulated autonomous driving dataset that contains scenes in various environments. We use images with the “day” and “night” labels as source and target domains, respectively, which contain 19,452 daytime images, 8,497 nighttime images for training, and 1,200 nighttime images for validation.

**TDND:** TDND (T) [38] not only covers severe weather such as heavy rain and snow but also retains complicated illumination conditions. We take 1,916 daytime images as the source domain, 7,663 nighttime images as the target domain, and we use 2,523 nighttime images for validation.

<sup>2</sup>The original reports of these methods are trained on 36,728 daytime images and 32,998 nighttime images, as well as validated on 4,707 images.

Table 2. Results of SHIFT dataset (daytime → nighttime).

Method	Venues	person	car	truck	bus	mcycle	bicycle	mAP
Source [39]	NeurIPS’15	40.4	44.5	49.9	53.7	14.3	46.7	41.6
Oracle [39]	NeurIPS’15	49.7	51.5	56.0	53.6	19.2	52.4	47.0
DA FR [7]	CVPR’18	43.0	48.8	47.8	52.1	19.9	55.8	43.7
UMT [14]	CVPR’21	7.7	47.5	18.4	46.8	16.6	49.2	31.1
SADA [8]	IJCV’21	48.5	49.0	51.6	61.9	18.1	52.0	46.8
AT [32]	CVPR’22	25.8	33.0	54.7	49.5	20.7	52.3	38.9
MIC [23]	CVPR’23	52.1	51.3	52.3	63.3	19.9	54.6	48.9
2PCNet [28]	CVPR’23	51.4	54.6	54.8	56.6	23.9	54.2	49.1
CoS [25]	ICME’24	50.8	56.0	57.2	<b>64.5</b>	22.2	55.5	51.0
SOCCER [11]	MM’24	52.1	55.6	<b>59.3</b>	61.0	25.4	55.6	51.5
ISP-Teacher [67]	AAAI’24	51.6	<b>59.1</b>	58.7	62.3	24.1	<b>58.3</b>	52.4
Ours	-	<b>52.3</b>	57.7	58.4	62.3	<b>27.1</b>	56.9	<b>52.8</b>

Table 3. Results of TDND dataset (daytime → nighttime).

Method	Venues	car	person	bus	minibus	truck	t-sign	mAP
Source [39]	NeurIPS’15	56.3	30.3	32.3	12.8	20.6	28.8	30.2
Oracle [39]	NeurIPS’15	71.9	42.1	39.8	17.4	26.9	30.9	38.2
SADA [8]	IJCV’21	72.4	35.7	36.9	14.7	14.9	30.1	34.1
MRT [70]	ICCV’23	71.4	45.5	29.4	13.6	14.8	24.0	33.1
2PCNet [28]	CVPR’23	77.4	39.3	52.5	8.4	10.5	25.9	35.6
MIC [23]	CVPR’23	79.6	28.0	42.4	17.2	22.9	27.5	36.3
ISP-Teacher [67]	AAAI’24	72.3	44.5	37.5	15.4	27.0	26.2	37.2
SOCCER [11]	MM’24	73.9	17.4	51.4	16.2	38.7	26.5	37.4
CoS [25]	ICME’24	74.8	<b>49.3</b>	53.1	17.6	27.2	31.2	42.2
Ours	-	<b>79.8</b>	48.3	<b>57.6</b>	<b>20.9</b>	<b>40.8</b>	<b>32.8</b>	<b>46.7</b>

### 5.2. Implementation Details

We adopt the Faster RCNN [39] with the ResNet-50 backbone [22] as our detection model. All images are scaled by resizing their shorter side to 600 pixels. For the CDRC module, we set the balance factor  $\alpha = 4$ . For the ConCal module, we set the amplitude factor  $\beta = 0.8$ , IoU matching threshold  $\tau = 0.6$ , base thresholds  $\delta_{base} = 0.8$ , lower and upper limit of the threshold  $\delta_{lower} = 0.8$ ,  $\delta_{upper} = 0.95$ , the minimum threshold  $\delta_{min} = 0.5$ . All experiments are implemented using Detectron2 [48] and conducted on 3 RTX4090 GPUs. We provide detailed parameter settings (iterations, batch size, loss weights, etc.) in *Appendix F*.

### 5.3. Comparison with SOTA

We compare DeT with other SOTA. “Source” refers to the base Faster RCNN trained on daytime data, while “Oracle” represents the base Faster RCNN trained on nighttime data. **Comparison on BDD100k.** In Tab. 1, our DeT achieves 50.9% mAP, the best performance of all self-training-based methods, and shows exceptional day-to-night adaptation in real driving scenarios. Benefiting from the CDRC module, we increase the frequency of underrepresented classes. As a result, our method achieves more balanced performance across classes, with an overall performance standard deviation of 9.41%, compared to 9.97% for CoS, 9.99% for ISP-Teacher, 12.33% for 2PCNet, and 14.30% for SOCCER.

**Comparison on SHIFT.** Tab. 2 demonstrates the cross-domain adaptation ability of our DeT in simulated driving

scenarios, validating that the DNDT module compensates for day-night distribution difference not only in real-world scenarios but also in simulated scenarios. DeT maintains a leading position in recognizing dense and easily confused objects (e.g., “person” and “mcycle”) while achieving the best performance at 52.8% mAP.

**Comparison on TDND.** Tab. 3 shows the excellent adaptation results of our DeT in severe weather driving scenarios. DeT achieves 46.7% mAP and outperforms all the SOTA methods. As the effect of confirmation bias, MIC and ISP-Teacher underperform in severe nighttime scenarios. In contrast, our method leverages ConCal to obtain accurate nighttime pseudo labels, providing effective guidance.

**Additionally,** we conduct further quantitative experiments using other metrics, as detailed in Appendix. D.

#### 5.4. Ablation Studies and Component Analysis

**Ablation Studies.** We conduct different experiments of each component in Tab. 4. Comparison between row 1 (baseline) and row 4 shows that DNDT compensates for the distribution bias and improves an average performance of 3.3% mAP on three benchmarks. Rows 2-4 shows each modeling’s effect. We find that noise and flare modeling significantly enhance performance in real night scenes (e.g., BDD100k and TDND). CDRC (row 6) reduces the training bias, obtaining an average gain of 3.2% mAP than only employing DNDT (row 4). However solely applying bidirectional mixing without incorporating ICFB may exacerbate class imbalance, leading to performance degradation (row 4 to 5). In rows 7-8, when integrating LPS and HPD, our average performance improves 1.0% mAP than row 6, highlighting ConCal generates more accurate labels.

##### Parameter Sensitivity.

**(1) Balance factor:** In Tab. 5 top, we observe that the balance factor  $\alpha$  set too low fails to balance classes, while set too high reduces the effectiveness of the CDRC module. Thus, we set  $\alpha = 4$  for better balance.

**(2) Amplitude factor:** In Tab. 5 middle, we explore the effect of amplitude factor  $\beta$  in ConCal module for adjusting the dynamic threshold. We find that when  $\beta = 0.8$ , the calculated dynamic threshold effectively maximizes suppression of low-confidence proposals across classes.

Table 4. Ablation study of each component on three benchmarks.

row	DNDT			CDRC		ConCal		B	S	T
	IM	NNM	LFM	Bi-Mix	ICFB	LPS	HPD			
1								45.8	47.6	34.5
2	✓							46.6	49.1	35.8
3	✓	✓						48.2	50.0	37.1
4	✓	✓	✓					48.9	50.2	38.7
5	✓	✓	✓	✓				48.1	49.5	37.6
6	✓	✓	✓	✓	✓			50.3	52.3	44.8
7	✓	✓	✓	✓	✓	✓		50.7	52.5	45.6
8	✓	✓	✓	✓	✓	✓	✓	<b>50.9</b>	<b>52.8</b>	<b>46.7</b>

Table 5. Parameter sensitivity experiment of  $\alpha$  in Eq. (10),  $\beta$  in Eq. (12), and  $\tau$  in Eq. (14). When adjusting one parameter, the other two in the blue background remain unchanged.

$\alpha$	1	2	3	4	5
S (mAP)	50.8	51.2	52.4	<b>52.8</b>	51.2
T (mAP)	45.3	45.2	45.6	<b>46.7</b>	44.8
$\beta$	0.2	0.4	0.6	0.8	1.0
S (mAP)	51.2	51.5	52.1	<b>52.8</b>	52.5
T (mAP)	44.1	44.4	44.3	<b>46.7</b>	45.4
$\tau$	0.2	0.4	0.6	0.8	1.0
S (mAP)	49.6	51.3	<b>52.8</b>	52.6	52.6
T (mAP)	43.8	44.9	<b>46.7</b>	46.5	46.1

**(3) IoU matching threshold:** In Tab. 5 bottom, we find that setting the IoU matching threshold  $\tau$  too low distills noisy proposals, leading to inaccurate pseudo labels, while setting  $\tau$  too high imposes stringent filter criteria without providing additional benefits. And the lower the  $\tau$ , the more it affects performance. We set  $\tau = 0.6$  as the optimal trade-off.

**The Effect of Different Mixing Strategies.** We explore different mixing strategies in Tab. 6. By default, we only change the discussed module, while the rest remain unchanged. Row 1 presents the results without mixing strategy i.e., without CDRC. In comparison with row 1, unidirectional mixing with ICFB (row 2,3) also enhances the feature representation and improves average performance of 1.8% mAP. In rows 4-6, we also adopt bidirectional strategy to explore some mature mixing solutions. Because the solutions ignore the class distribution, causing the severe class imbalance (similar in Tab. 4 row 4 to 5). For row 6, we set the mixing ratio to 0.5 and the class label is mixed when objects occlude. However, soft mixing introduces class ambiguity that affects the model’s feature representation. By comparing row 7 and row 8, we find that using source domain images for mixing results in a 2.8/5.0% mAP performance drop. This is because the large distribution bias causes the model to learn the biased features that are poorly adapted.

**The Effect of Different Sources for Calculating Confidence.** As shown in Tab. 7, we analyze four sources for calculating average confidence and find that directly utilizing the target images to calculate resulted in only 50.9% mAP, even lower than using the source images. This is because, in the early training, the model cannot accurately identify the correct classes, making it difficult to estimate the av-

Table 6. Ablation study of different mixing strategies.

row	Mixing Strategy	$I_{mixed}$	B (mAP)	T (mAP)
1	w/o CDRC	-	49.5	43.4
2	D2N Mixing	$I_n$	50.5 (+1.0)	46.3 (+2.9)
3	N2D Mixing	$I_n$	50.2 (+0.7)	45.9 (+2.5)
4	Mosaic [4]	$I_n$	47.6 (-1.9)	37.7 (-5.7)
5	Copy-Paste [19]	$I_n$	48.9 (-0.6)	41.2 (-2.2)
6	Mix-up [64]	$I_n$	49.2 (-0.3)	44.9 (+1.5)
7	CDRC	$I_s$	48.1 (-1.4)	41.7 (-1.7)
8	CDRC	$I_n$	<b>50.9 (+1.4)</b>	<b>46.7 (+3.3)</b>

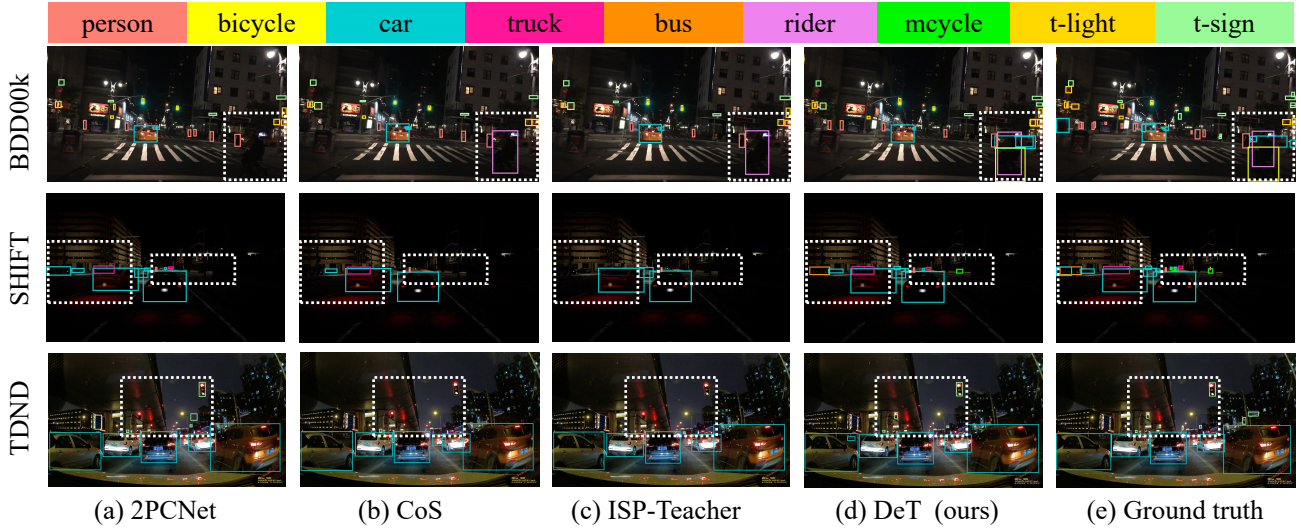


Figure 3. Qualitative results comparison on BDD100k (top), SHIFT (middle), and TDND (bottom) for SOTA: (a) 2PCNet [28], (b) CoS [25], (c) ISP-Teacher [67], (d) Ours, and (e) Ground truth.

Table 7. The experiment of the statistical sources for class average confidence in the ConCal module on the SHIFT dataset.

Sources	mAP	mAP <sub>s</sub>	mAP <sub>m</sub>	mAP <sub>l</sub>
Target Images $I_t$	50.9	11.7	56.2	77.9
Source Images $I_s$	51.2 (+0.3)	12.0 (+0.3)	57.3 (+1.1)	78.6 (+0.7)
Night-like Images $I_n$	52.6 (+1.7)	12.7 (+1.0)	<b>60.0 (+3.8)</b>	80.0 (+2.1)
Night-like Images $I_n$ w/ GT	<b>52.8 (+1.9)</b>	<b>13.0 (+1.3)</b>	59.9 (+3.7)	<b>80.2 (+2.3)</b>

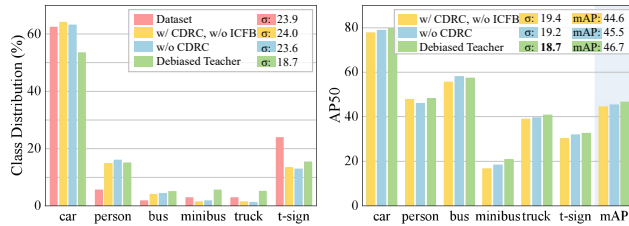


Figure 4. Left: We conduct four group experiments to exhibit the class distribution from the TDND dataset and three training settings. Right: We also exhibit the performance of the model for each class from three settings.  $\sigma$  denotes the standard deviation.

erage confidence for each class fairly. Compared with  $I_s$ , night-like images generated by DNDT provide more objective conditions for calculating each class’ average confidence. Finally, we utilize the ground truth to calculate the confidence of matched objects obtains 52.8% mAP.

**The Effect of ICFB for Balancing Classes.** As shown in Fig. 4, we observe that the variant with CDRC but without ICFB intensifies class imbalance (the  $\sigma$  of class distribution increases by 0.4%, and the  $\sigma$  of performance increases by 0.2%, even damaging detection performance by 0.9% mAP). While with ICFB, the DeT achieves the smallest  $\sigma$  in both class distribution and class performance.

**Qualitative Results.** As shown in Fig. 3, DeT demonstrates

accurate localization and classification of most objects, especially for underrepresented classes (e.g., “rider” and “bicycle” in row 1(d), and “bus” in row 2(d) and distant objects under low-light conditions (e.g., objects at the end of the road in row 2(d) and “traffic signs” in the sky in row 3(d)). In contrast, 2PCNet mistakenly detects something as “traffic signs” (row 3 (a)), while CoS and ISP-Teacher fail to detect objects at a distance both on the street and sky (row 2, 3 (b, c)). Meanwhile, we exhibit the visualization of mixed images in *Appendix. B*.

## 6. Conclusions

In this paper, we propose a Debiased Teacher (DeT), which systematically addresses the three biases inherent in the self-training framework for DN-DAOD task. Firstly, DNDT reduces the distribution bias by modeling for the key day-night differences to obtain the night-like images. Secondly, CDRC mitigates the training bias by selectively mixing objects from two domains to alleviate class imbalance and source domain bias. Thirdly, ConCal corrects the confirmation bias by generating high-quality pseudo labels for better nighttime knowledge learning. We conduct comprehensive experiments on three benchmarks to validate the cross-domain effectiveness of our DeT.

## Acknowledgement

This work was supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang Province(2023C01046, 2023C01038, 2024C01023), the National Nature Science Foundation of China (62322211, U21B2024, 62336008), the National Key R&D Program of China under Grant (2023YFB4502803), and Zhejiang Provincial Natural Science Foundation of China (LDT23F01014F01).

## References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 2
- [2] Vinicius F Arruda, Thiago M Paixao, Rodrigo F Berriel, Alberto F De Souza, Claudine Badue, Nicu Sebe, and Thiago Oliveira-Santos. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 1
- [3] Jing Chong Beh, Kam Woh Ng, Jie Long Kew, Che-Tsung Lin, Chee Seng Chan, Shang-Hong Lai, and Christopher Zach. Cyeda: Cycle-object edge consistency domain adaptation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2986–2990. IEEE, 2022. 1
- [4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 7
- [5] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debaised self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35:32424–32437, 2022. 2
- [6] Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yuet-ing Zhuang, Di Xie, et al. Learning domain adaptive object detection with probabilistic teacher. *arXiv preprint arXiv:2206.06293*, 2022. 2
- [7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 6
- [8] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021. 6
- [9] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2023. 2
- [10] Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Hengel, Jian Yang, and Qingming Huang. Emodubber: Towards high quality and emotion controllable movie dubbing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15863–15873, 2025. 2
- [11] Yiming Cui, Liang Li, Jiehua Zhang, Chenggang Yan, Hongkui Wang, Shuai Wang, Jin Heng, and Wu Li. Stochastic context consistency reasoning for domain adaptive object detection. In *ACM Multimedia 2024*, 2024. 2, 3, 6
- [12] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2553–2562, 2021. 1, 4
- [13] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7k: A phenomenological nighttime flare removal dataset. *Advances in Neural Information Processing Systems*, 35:3926–3937, 2022. 4
- [14] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 6
- [15] Jinhong Deng, Dongli Xu, Wen Li, and Lixin Duan. Harmonious teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23829–23838, 2023. 2
- [16] Peihua Deng, Jiehua Zhang, Xichun Sheng, Chenggang Yan, Yaoqi Sun, Ying Fu, and Liang Li. Multi-granularity class prototype topology distillation for class-incremental source-free unsupervised domain adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30566–30576, 2025. 2
- [17] Zhipeng Du, Miaoqing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12666–12676, 2024. 1
- [18] Li Gao, Jing Zhang, Lefei Zhang, and Dacheng Tao. Dsp: Dual soft-paste for unsupervised domain adaptive semantic segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 2825–2833, 2021. 3
- [19] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 7
- [20] Boyang Guo, Liang Li, Jiehua Zhang, Yaoqi Sun, Chenggang Yan, and Xichun Sheng. Prompt learning with knowledge regularization for pre-trained vision-language models. *IEEE Transactions on Multimedia*, 2025. Accepted for publication. 2
- [21] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3714–3722, 2019. 3
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [23] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 2, 6
- [24] Tzuhsuan Huang, Chen-Che Huang, Chung-Hao Ku, and Jun-Cheng Chen. Blenda: Domain adaptive object detection through diffusion-based blending. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4075–4079, 2024. 1

- [25] Yuan Jicheng, Le-Tuan Anh, Hauswirth Manfred, and Le-Phuoc Danh. Cooperative students: Navigating unsupervised domain adaptation in nighttime object detection. *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024. 1, 2, 3, 6, 8
- [26] Masanori Kakimoto, Kaoru Matsuoka, Tomoyuki Nishita, Takeshi Naemura, and Hiroshi Harashima. Glare generation based on wave optics. In *12th Pacific Conference on Computer Graphics and Applications, 2004. PG 2004. Proceedings.*, pages 133–140. IEEE, 2004. 4
- [27] Purbayan Kar, Vishal Chudasama, Naoyuki Onoe, and Pankaj Wasnik. Revisiting class imbalance for end-to-end semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4570–4579, 2023. 2
- [28] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11484–11493, 2023. 1, 2, 3, 6, 8
- [29] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. Cat: Exploiting inter-class dynamics for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16541–16550, 2024. 2, 3
- [30] Daehan Kim, Minseok Seo, Kwanyong Park, Inkyu Shin, Sanghyun Woo, In So Kweon, and Dong-Geol Choi. Bidirectional domain mixup for domain adaptive semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1114–1123, 2023. 3
- [31] Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Yingying Li, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Gradient-based sampling for class imbalanced semi-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16390–16400, 2023. 2
- [32] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 2, 6
- [33] Zeyi Li, Pan Wang, and Zixuan Wang. Flowganomaly: Flow-based anomaly network intrusion detection with adversarial learning. *Chinese Journal of Electronics*, 33(1):58–71, 2024. 2
- [34] Ariel Lipson, Stephen G Lipson, and Henry Lipson. *Optical physics*. Cambridge University Press, 2010. 4
- [35] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3003–3018, 2022. 2
- [36] Amitangshu Mukherjee, Ameya Joshi, Anuj Sharma, Chinmay Hegde, and Soumik Sarkar. Generative semantic domain adaptation for perception in autonomous driving. *Journal of big data analytics in transportation*, 4(2):103–117, 2022. 1
- [37] Yuzuru Nakamura, Yasunori Ishii, Yuki Maruyama, and Takayoshi Yamashita. Few-shot adaptive object detection with cross-domain cutmix. In *Proceedings of the Asian Conference on Computer Vision*, pages 1350–1367, 2022. 3
- [38] Chang Nie, Muhammad Ali Qadar, Shaodong Zhou, Hui Zhang, Yang Shi, Jinwu Gao, and Zhifeng Sun. Transnational image object detection datasets from nighttime driving. *Signal, Image and Video Processing*, 17(4):1123–1131, 2023. 6
- [39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6
- [40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 3
- [41] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 6
- [42] Ye Tian, Ying Fu, and Jun Zhang. Transformer-based under-sampled single-pixel imaging. *Chinese Journal of Electronics*, 32(5):1151–1159, 2023. 2
- [43] Wilhelm Truhedden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1379–1389, 2021. 3
- [44] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4926–4943, 2024. 2
- [45] Yunbin Tu, Liang Li, Li Su, and Qingming Huang. Query-centric audio-visual cognition network for moment retrieval, segmentation and step-captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7464–7472, 2025. 2
- [46] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020. 4
- [47] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2021. 4
- [48] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6

- [49] Hang Xu, Xinyuan Liu, Haonan Xu, Yike Ma, Zunjie Zhu, Chenggang Yan, and Feng Dai. Rethinking boundary discontinuity problem for oriented object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17406–17415, 2024. 2
- [50] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3060–3069, 2021. 2
- [51] Guo Yachan, Xiao Yi, Xue Danna, Jose Luis Gomez Zurita, and Antonio M López. Synth-to-real unsupervised domain adaptation for instance segmentation. *arXiv preprint arXiv:2405.09682*, 2024. 3
- [52] Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1445–1451, 2020. 2
- [53] Chenggang Yan, Zhisheng Li, Yongbing Zhang, Yutao Liu, Xiangyang Ji, and Yongdong Zhang. Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(4):1–17, 2020.
- [54] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video technology*, 32(1):43–51, 2021. 2
- [55] Chenggang Yan, Tong Teng, Yutao Liu, Yongbing Zhang, Haoqian Wang, and Xiangyang Ji. Precise no-reference image quality evaluation based on distortion identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s):1–21, 2021. 2
- [56] Chenggang Yan, Lixuan Meng, Liang Li, Jiehua Zhang, Zhan Wang, Jian Yin, Jiyong Zhang, Yaoqi Sun, and Bolun Zheng. Age-invariant face recognition by multi-feature fusion and decomposition with self-attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–18, 2022.
- [57] Chenggang Yan, Yaoqi Sun, Hao Zhong, Chenwei Zhu, Zunjie Zhu, Bolun Zheng, and Xiaofei Zhou. Review of omni-media content quality evaluation. *J. Signal Process.*, 38(6):1111–1143, 2022.
- [58] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. 2
- [59] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised cross-media hashing learning via knowledge graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. 2
- [60] Jiong Yin, Liang Li, Jiehua Zhang, Chenggang Yan, Lei Zhang, and Zunjie Zhu. Reducing intrinsic and extrinsic data biases for moment localization with natural language. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4584–4594, 2023. 2
- [61] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 6
- [62] Beichen Zhang, Liang Li, Shuhui Wang, Shaofei Cai, Zheng-Jun Zha, Qi Tian, and Qingming Huang. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [63] Fangyuan Zhang, Tianxiang Pan, and Bin Wang. Semi-supervised object detection with adaptive class-rebalancing self-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3252–3261, 2022. 2
- [64] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3, 7
- [65] Tao Zhang, Ying Fu, and Jun Zhang. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1):303–312, 2024. 2
- [66] Tao Zhang, Ying Fu, and Jun Zhang. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1):303–312, 2024. 2
- [67] Yin Zhang, Yongqiang Zhang, Zian Zhang, Man Zhang, Rui Tian, and Mingli Ding. Isp-teacher: Image signal process with disentanglement regularization for unsupervised domain adaptive dark object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7387–7395, 2024. 1, 3, 4, 6, 8
- [68] Zhedong Zhang, Liang Li, Gaoxiang Cong, Haibing Yin, Yuhao Gao, Chenggang Yan, Anton van den Hengel, and Yuankai Qi. From speaker to dubber: movie dubbing with prosody and duration consistency learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7523–7532, 2024. 2
- [69] Zhedong Zhang, Liang Li, Chenggang Yan, Chunshan Liu, Anton van den Hengel, and Yuankai Qi. Prosody-enhanced acoustic pre-training and acoustic-disentangled prosody adapting for movie dubbing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 172–182, 2025. 2
- [70] Zijing Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, and Yang Liu. Masked retraining teacher-student framework for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19039–19049, 2023. 2, 6
- [71] Zhiqian Zhao, Liang Li, Jiehua Zhang, Yaoqi Sun, Xichun Sheng, Haibing Yin, and Shaowei Jiang. Heterogeneous prompt-guided entity inferring and distilling for scene-text aware cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10537–10545, 2025. 2
- [72] Jinkai Zheng, Xinchun Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20228–20237, 2022. 2