Revisiting Domain-Adaptive Object Detection in Adverse Weather by the Generation and Composition of High-Quality Pseudo-Labels

Rui Zhao, Huibin Yan, and Shuoyao Wang*

School of Electronics and Information Engineering,
Shenzhen University, Shenzhen, China
zhaorui2022@email.szu.edu.cn, huibinyan2020@email.szu.edu.cn,
sywang@szu.edu.cn

Abstract. Due to data collection challenges, the mean-teacher learning paradigm has emerged as a critical approach for cross-domain object detection, especially in adverse weather conditions. Despite significant progress, existing methods are still plagued by low-quality pseudo-labels in degraded images. This paper proposes a generation-composition paradigm training framework that includes the tiny-object-friendly loss, i.e., IAoU loss with a joint-filtering and student-aware strategy to improve pseudo-labels generation quality and refine the filtering scheme. Specifically, in the generation phase of pseudo-labels, we observe that bounding box regression is essential for feature alignment and develop the IAoU loss to enhance the precision of bounding box regression, further facilitating subsequent feature alignment. We also find that selecting bounding boxes based solely on classification confidence performs poorly in cross-domain noisy image scenes. Moreover, relying exclusively on predictions from the teacher model could cause the student model to collapse. Accordingly, in the composition phase, we introduce the mean-teacher model with a joint-filtering and student-aware strategy combining classification and regression thresholds from both the student and the teacher models. Our extensive experiments, conducted on synthetic and real-world adverse weather datasets, clearly demonstrate that the proposed method surpasses state-of-the-art benchmarks across all scenarios, particularly achieving a 12.4% improvement of mAP, i.e., Cityscapes to RTTS. Our code will be available at https://github.com/iu110/GCHQ/.

Keywords: cross-domain · adverse weather· pseudo-labels · tiny-object-friendly loss · joint-filtering and student-aware

1 Introduction

Over the past decade, object detection algorithms have achieved remarkable success, serving as the foundation for numerous high-level vision applications [1, 24, 31, 44, 55]. Nevertheless, the ability to identify targets in challenging weather

* Corresponding author.

conditions, a critical requirement for applications such as autonomous driving, remains constrained [35, 39]. In particular, collecting data in adverse weather involves a significant investment of human and time resources, and acquiring high-quality datasets presents a formidable challenge [19, 41]. Object detection algorithms trained under clear conditions often struggle in adverse weather due to factors like fog, rain, and night, which can severely disrupt the feature extraction, and thus introduce a significant domain gap [7].

Recently, some research efforts have been devoted to adverse-weather object detection, which can be divided into three categories. The first category involves a restoration model to preprocess images before inputting them into the detection model [19, 25]. The second way integrates image restoration tasks with object detection tasks within a unified framework [17, 41]. Both tasks share the backbone feature extraction network to facilitate inter-task interactions. Notably, restoration models in the above two categories may lose significant details during the restoration process, potentially impeding the detection work. The third approach, unsupervised domain adaptation, involves transferring pertinent feature information from the source domain, i.e., clear weather to the target domain, i.e., adverse weather, mitigating domain disparities, thus enhancing detection performance [8, 20, 29, 52, 54]. The mean-teacher model, a prevalent method for unsupervised domain adaptive object detection, involves pretraining a teacher model on source domain data, generating pseudo-labels on the target domain, and updating the teacher model via Exponential Moving Average, i.e., EMA during student model training. However, this approach relies on pseudo-labels, which are noisy. It is also easy to ignore the teacher's misguided instructions to the student model, especially during heavy rain or fog, which poses significant challenges [10, 36, 45, 49].

Drawing on the presented analysis, we propose a generation and composition training framework. In the generation phase, we design a novel regression loss that addresses issues related to non-intersecting predicted boxes and regression boxes, especially the two boxes in containment situations, improving the accuracy of bounding box regression on the target domain. Thus, facilitating subsequent feature alignment between the two domains. We also employ image rendering as well as restoration modules [11, 12, 42] to generate more clean intermediate features. In the composition phase, we incorporate a mean-teacher learning framework. Initially, we design a joint-filtering and student-aware strategy using high and low classification confidence thresholds with regression variance metrics which are predicted by the student and teacher models to filter pseudo-labels. Subsequently, we combine lower-level image restoration with super-resolution tasks to compensate for the deficiencies of pseudo-labels in the target domain.

Following the two steps above, we can create a high-performance object detection model that works well in adverse weather conditions. Here is a summary of our contributions:

- We propose a generation and composition two-stage framework comprising an optimization feature alignment generation model for pseudo-labels and a composition model with a joint-filtering and student-aware strategy.

- More Features for Feature Alignment: To cope with the challenges associated with transferring from high-quality images to low-quality images for feature extraction, we need more feature information to align features effectively. i) We develop the IAoU loss to enhance bounding box regression accuracy to contain more useful information for feature alignment to narrow the domain gap. ii) We integrate restoration-enabled alignment with image-to-image translation to provide clean intermediate features to bridge the two domains.
- Composition of High-Quality Pseudo-Labels: In the teacher-student model, We require high-quality pseudo-labels to steer student model training. i) To filter out the low-quality pseudo-labels, we design a joint-filtering and student-aware strategy by preventing classification-localization inconsistency and misleading instances. ii) To obtain more high-quality pseudo labels in noisy target domains, we augment the target domain with restored and super-resolution images and investigate the uncertain pseudo labels.
- Extensive experiments demonstrate that our model outperforms previous state-of-the-art methods on synthetic datasets, i.e., Rain Cityscapes, Foggy Cityscapes, and real-world datasets, i.e., RTTS, BDD100K in adverse weather conditions, particularly achieving a 12.4% mAP improvement on the RTTS dataset.

2 Related work

2.1 Object detection in adverse weather

To cope with adverse weather conditions, [19, 25] involve pre-processing, such as removing rain [40, 50, 56] or fog [51, 53], before feeding the images into the detection model. Accordingly, some studies have developed multi-task models for restoration and detection, utilizing feature information extracted from a unified backbone network for both tasks [17, 41]. While the above approaches can enhance the overall image quality, the restoration model used in these approaches may result in the loss of crucial information necessary for detection. Additionally, there are domain adaptation methods that treat adverse weather object detection as a domain transfer problem. The adversarial feature learning-based models, such as DA-YOLO and MS-DAYOLO, introduced the Domain Adaptation Network (DAN) to learn domain-invariant features across multiple scales [15,57]. In addition, R-YOLO adopted a novel image-to-image translation method and designed a pixel-global level feature calibration module to narrow the domain gap [39]. The mean-teacher-based model, such as SSDA-YOLO [62], incorporated a teacher-student knowledge distillation framework. Ref [62] introduced a distillation loss to correct cross-domain differences and a novel consistency loss to mitigate biases in cross-domain targets.

2.2 Unsupervised Domain Adaptation for Object Detection

Unsupervised domain adaptation has rapidly developed in recent years, with three main approaches: feature adversarial learning [8, 14, 33, 35, 37, 47, 52, 59],

image-to-image translation [3, 16, 18], and mean-teacher learning [5, 10, 23, 23, 30, 36, 38, 60]. Feature adversarial learning methods exploit the gradient reversal layer and domain classifier to obtain the domain-invariant feature between the two domains. The image-to-image translation approaches obtain intermediate features by the generative model to reduce the two-domain discrepancy. To the best of our knowledge, adversarial feature learning is constrained by various factors, including the integrity of intermediate features and bounding box regression accuracy. Similarly, image-to-image translation is hindered by the poor performance of the transfer approach in low-quality image scenarios.

Currently, the most dominant approach is the mean-teacher learning framework. UMT [10] integrated the image-to-image translation with the mean teacher to alleviate the bias in the student model and designed the dynamically adjust strategy to select the most suitable pseudo-labels, AT [23] applied weak-strong augmentation to reduce the domain gap at image-level and adversarial training to align the feature, PT [5] proposed the uncertainty-guided self-training strategy to improve the pseudo-labels, CMT [30] developed an object-level contrast learning strategy to enhance the representation of pseudo-labels. While these approaches have made great progress from clean weather to synthetic adverse weather, they did not account for regression branching for the pseudo-labels filtering and only considered the teacher model. When facing substantial domain variations, merely improving pseudo-labels in the framework mining process is insufficient.

3 Proposed Method

This paper introduces a generation-composition training framework for cross-domain object detection in adverse weather conditions. In this section, we will go over each phase in detail. Firstly, we will explore how to achieve high-quality feature alignment through image-to-image translation. Secondly, we will elaborate on the motivation and implementation details of the loss optimization. lastly, we will discuss the composition learning model, which utilizes a joint-filtering and student-aware strategy.

3.1 Generation phase

The overview of our generation framework is shown in Fig. 1. Firstly, considering the difference in image level between the two domains, we adopt a rendering model trained on the paired images (I_s, I_t) to get the target-like fake source images I_s^f, where I_s is the source domain images and I_t is the target domain. Then, restoring the target domain images to generate source-like fake target images I_t^f. After that, IAOU loss is proposed to optimize the object regression. Then, we input all the translated images into the detectors with the Global Feature Alignment module to further reduce the differences between the two domains.

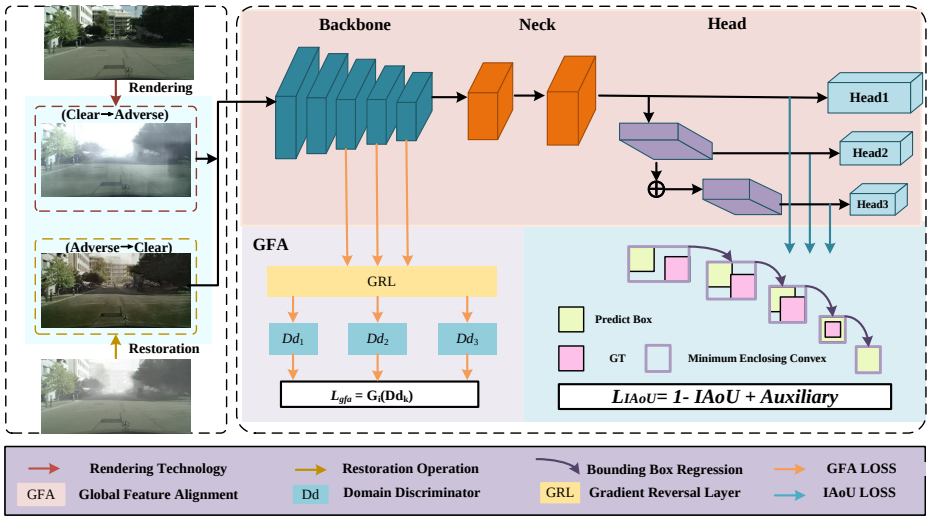

Fig. 1: The framework of the generation phase.

Image to image translation. To bridge the gap between the two domains at the image level, this paper aims to generate source domain images in the style of the target domain i.e., from clear to adverse, and target domain images in the kind of the source domain i.e., from adverse to clear [3, 28, 32].

Given the instability of generative models based on GANs, this paper utilizes rendering techniques to create intermediate images [11, 39]. The figure in supplementary material compares the methods used in this paper with GAN-based methods such as Cut and CycleGAN [30, 64]. The rendering approach achieves better results in translating images from normal to adverse, preserving most of the intermediate features with minimal noise. However, we observe *a significant amount of noise when transforming from a noisy environment i.e., adverse weather to a normal setting i.e., clear weather*. Accordingly, we employ the restoration models to preserve the essential features when translating images from adverse to normal. The restored images, as shown in the supplementary material, exhibit less susceptibility to noise compared to the generated images and contain a more comprehensive set of source domain features.

IAoU loss. The loss is crucial for improving the accuracy of bounding box regression, which is vital to global feature alignment. However, in the case of small targets, bounding box regression is more sensitive to position than medium and large targets. During the training phase, this may lead to scenarios where the ground truth boxes and prediction boxes do not intersect, resulting in the Intersection over Union i.e., IoU being zero and thus zero gradient. To address this issue, the original YOLOv5 incorporates the CIoU loss [61], which includes

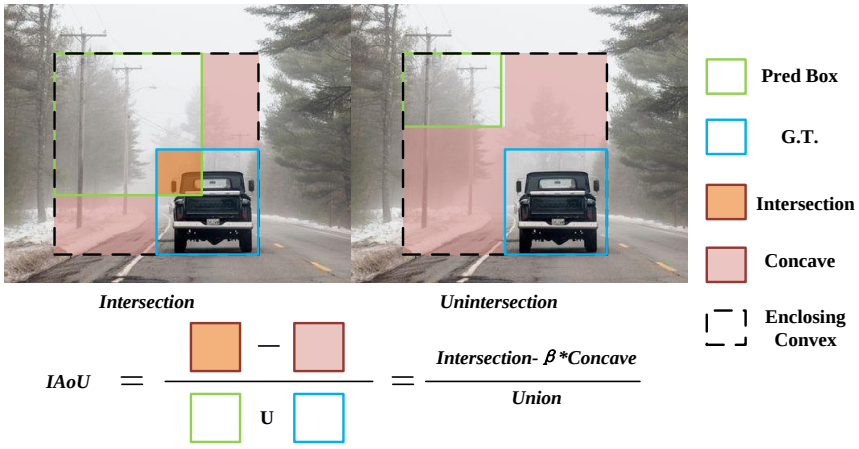

Fig. 2: The schematic diagram of the IAoU.

a penalty term υ when the IoU is zero:

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^*}{h^*} - \arctan\frac{\hat{w}}{\hat{h}}\right)^2, \tag{1}$$

where w^* and h^* are the weight and height of the ground truth, and $\hat{w}$ and $\hat{h}$ are the weight and height of the prediction boxes, respectively. Unfortunately, the penalty term creates difficulties in simultaneously increasing or decreasing the width and height of the prediction box, which hinders the convergence of the loss function.

In this paper, we introduce a novel loss function called IAoU loss, in which the similarity between two bounding boxes is computed using the Intersection Auxiliary over Enclosing Convex i.e., IAoU, which is inspired by the Intersection over Convex i.e., IoC and MPDIoU [34, 58]. The supplementary material will provide detailed comparisons of the two losses mentioned above. As shown in Fig. 2, the IAoU is defined as:

$$IAoU = \frac{I - \beta(E - U)}{U}, \tag{2}$$

where β is the balancing coefficient, E is the minimum enclosing convex of the predicted box, and the ground truth i.e., GT, I and U is the intersection and union of the two boxes, respectively.

Compared to IoU, the numerator of this equation includes an additional term called 'concave', which is subtracted from the intersection to form the numerator. 'Concave' represents the portion obtained by subtracting the union of the two bounding boxes from the minimum enclosing bounding box of the predicted box and the ground truth. Notably, to balance the intersection and

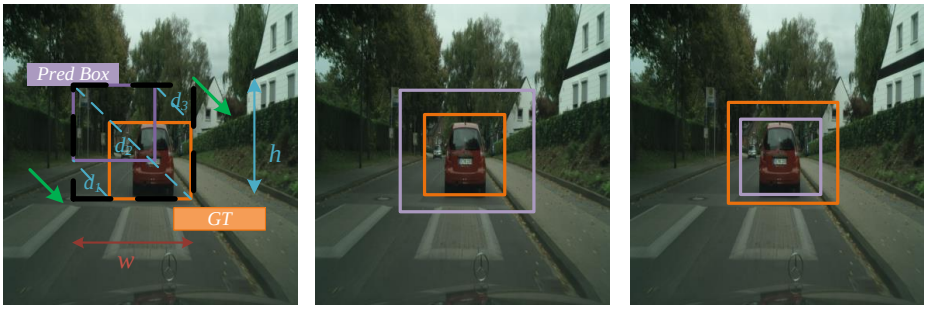

Fig. 3: The auxiliary term. (a) is the schematic of the proposed auxiliary term, (b) shows the situation where the prediction box surrounds the GT, and (c) depicts the reversal.

concave components, a balancing coefficient β is introduced. *In cases where the two bounding boxes do not intersect, IAoU is still related to the distance between the prediction and the GT thus continuously providing training gradients.* As the two bounding boxes gradually approach each other during training, the reduction in IAoU is smoother compared to the Intersection over Union (IoU). Additionally, we have introduced a supplementary training term, namely, the ratio of the distance between the top-right and bottom-left corners of two bounding boxes to the diagonal of the minimum enclosing convex, as defined in Equation 3:

$$\begin{aligned} L_{IAoU} &= 1 - IAoU + \frac{d_1^2 + d_2^2}{w^2 + h^2}, \\ d_1^2 &= \left(x_1^* - \hat{x_1}\right)^2 + \left(\mathrm{y}_1^* - \hat{y_1}\right)^2, \\ d_2^2 &= \left(x_2^* - \hat{x_2}\right)^2 + \left(\mathrm{y}_2^* - \hat{y_2}\right)^2, \end{aligned} \tag{3}$$

where $\hat{x}$ and x^* represent the predicted bounding box and the ground truth, $(x_1,\ y_1)$ and $(x_2,\ y_2)$ denote the coordinates of their respective top-right and bottom-left corners, and w and h correspond to the width and height of the minimum enclosing convex.

As depicted in Fig. 3, compared to the complex auxiliary training terms used in prior work, the auxiliary term introduced in this paper is robust to outliers and mitigates substantial fluctuations. Furthermore, when the predicted bounding box contains the ground truth bounding box, it remains flexible instead of becoming a fixed value. In situations where the two bounding boxes do not intersect, the auxiliary term exerts influence and promotes *faster convergence of the predicted bounding box towards the ground truth.* Conversely, when the two bounding boxes intersect, the auxiliary term is used to *fine-tune the coordinates of the predicted bounding box.*

Global feature alignment. To further mitigate high-level differences, this paper introduces a global feature alignment module using gradient reverse lay-

ers i.e., GRL and domain classifiers [39, 63]. In summary, the strategy involves maximizing the domain classification loss to obtain domain-invariant features to reduce the dissimilarities between the two domains. As shown in Equation 4:

$$L_{dc} = -\sum_{k,x,y} \left[C_k \ln b_k^{(x,y)} + (1 - C_k) \ln \left(1 - b_k^{(x,y)} \right) \right], \tag{4}$$

where is a binary cross-entropy form and $C_k \in \{0, 1\}$ is the domain label. Here, $C_k = 0$ and $C_k = 1$ denote the labels of the source and target domains, respectively, and b_k is the probability of belonging to a certain category at the feature map (x, y) of the k-th image.

3.2 Composition phase

As shown in Fig. 4, the student model is trained on weakly augmented source domain images, strongly augmented target domain images, and restored target domain images. The inclusion of weakly-strongly augmented images as inputs serves two purposes: reducing the differences between the source and target domains and mitigating the impact of adverse weather conditions.

It needs to be emphasized that, apart from previous work, to generate more high-quality pseudo-labels for the target domain, we propose the addition of restoration and super-resolution techniques as new strong augmentations, taking into account the interference of fog or rain factors and the presence of small targets. These methods are implemented in conjunction with existing strong augmentation techniques, such as left-right flipping, mosaic, Gaussian blurring, and cutout. Weak augmentations are still mosaic and mixed up.

Pseudo-Labels filtering. The quality of pseudo-labels significantly influences the performance of cross-domain learning models [10, 36, 46]. In the generation phase, our pre-training model aims to improve the quality of pseudo-labels. In the composition phase, an effective strategy for filtering pseudo-labels is crucial, as it constitutes one of the fundamental challenges in semi-supervised learning models. Existing methods employ a single confidence threshold from the teacher model for pseudo-label filtering, but it is insufficient for adequately filtering pseudo-labels.

In this paper, we introduce a more elaborated pseudo-label filtering strategy i.e., PLF. Given the inconsistency between classification and regression tasks, achieving a high classification score does not necessarily ensure accurate regression results, and vice versa. Moreover, only relying on the teacher's classification confidence and regression uncertainty predictions to select pseudo-labels that guide the student model cannot prevent misleading instances. Accordingly, we propose a joint-filtering and student-aware approach that incorporates an extra filtering step for regression bounding boxes based on Gaussian variance thresholds. This additional step complements the classification threshold, and both thresholds are predicted by the student and teacher models. Firstly, our approach involves dividing the classification confidence threshold into two levels:

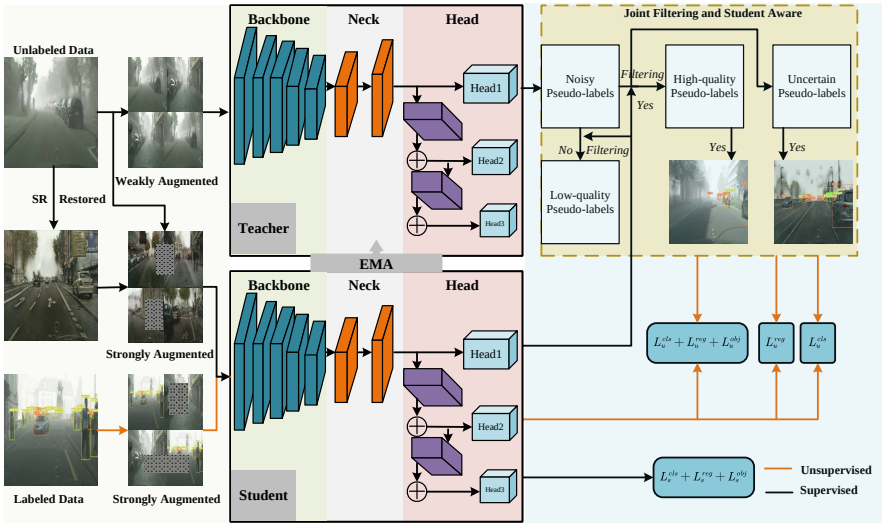

Fig. 4: The end-to-end pipeline of the composition training phase. Restored and SR are the restoration and super-resolution techniques.

a high threshold i.e., θ_1, and a low threshold i.e., θ_2 [46]. Pseudo-labels with scores above the high threshold are retained as high-quality classification labels, they meet the conditions of the first step and need to further determine whether they are high-quality pseudo-labels. While those below the low threshold are discarded as low-quality labels. For labels with scores between the two thresholds, secondary filtering is applied because of their uncertainty. Assessing the accuracy of regression bounding boxes directly is challenging, but the confidence derived from the uncertainty of regression bounding boxes can still enhance the quality of pseudo-labels [9, 27].

Secondly, to assess the uncertainty of regression models, we employ a Gaussian modeling approach. In addition to the output coordinate information, variances for each coordinate of the regression box i.e., B_{bbx} are generated, as indicated in Equation 5:

$$B_{bbx} = [\mu_x, \mu_y, \mu_w, \mu_h, \sigma_x, \sigma_y, \sigma_w, \sigma_h], \tag{5}$$

where μ_x, μ_y, μ_w, and μ_h denote the mean of the center-x, center-y, weight, and height of the prediction bounding box, respectively. Notations $\sigma_x, \sigma_y, \sigma_w, \sigma_h$ represent the variances of the center-x, center-y, weight, and height, which indicates the uncertainty associated with the regression box. For ease of further processing, we normalize the output variances to a range of 0 to 1. The overall confidence in Gaussian regression for the bounding boxes is calculated using Equation 6:

$$Conf_{bbx} = 1 - \frac{\sigma_x + \sigma_y + \sigma_w + \sigma_h}{4}, \tag{6}$$

where a higher confidence value implies lower accuracy of the bounding box. To maintain a balance between the labeled source domain and the unlabeled target domain in the composition learning process, we introduce an unsupervised loss, denoted as L_u. The overall loss function is defined in Equation 7:

$$L_2 = L_s + \alpha L_u, \tag{7}$$

where L_s represents the loss during training on labeled data, while L_u signifies the loss during training on unlabeled data. The balancing coefficient,α, is set to 1.5 in this paper. L_s aligns with Equation 10, and we refrain from modifying the L_{reg} due to the uncertainties in the composition learning. All three sub-losses: L_u^{cls}, L_u^{obj}, and L_u^{reg} correspond to unsupervised classification loss, unsupervised confidence loss, and unsupervised regression loss, respectively. ω are the weighting factors, and they are consistent with the original model, as expressed in Equation 8:

$$L_u = \omega_1 L_u^{cls} + \omega_2 L_u^{obj} + \omega_3 L_u^{reg}. \tag{8}$$

To increase the number of pseudo-labels, we retained uncertain ones rather than discarding them. We treat them as soft labels. To calculate the total classification threshold, referred to as $Conf$, which is determined by multiplying the confidence threshold obj_{conf} with the class confidence cls_{conf}, as shown in Equation 9:

$$Conf = obj_{conf} \cdot cls_{conf}. \tag{9}$$

For a pseudo-label to be classified as high-quality, it must meet three criteria: the classification scores from the teacher model must exceed the high threshold θ_1 and the scores predicted by the student model, and the confidence in the Gaussian regression of the bounding box from the teacher model must be less than θ_3, as shown in Equation 10, the $Conf_s$ is the confidence from the student model and $Conf_t$ is the teacher model:

$$L_u = \begin{cases} L_u^{cls} + L_u^{obj} + L_u^{req}, & \text{if } Conf_t > \theta_1 \text{ and} \\ & Conf_{bbx_t} < \theta_3 \text{ and} \\ & Conf_t >= Conf_s, \\ 0, & \text{if } Conf_t < \theta_2, \\ L_u^1, & otherwise. \end{cases} \tag{10}$$

In such cases, we calculate three sub-losses. If the classification confidence falls below the low threshold θ_2, it is categorized as background, and no loss is calculated. When classification thresholds fall between the θ_2 and θ_1, they are classified as uncertain pseudo-labels. However, most of these uncertain pseudo-labels exhibit lower regression accuracy, resulting in many false positives. As shown in Equation 11, the obj_{conf_s} and cls_{conf_s} is the confidence from the student

model, the $obj_{conf_t}$and cls_{conf_t} is the teacher model:

$$L_{u\ (\theta_2< Conf <\theta_1)}^{1} = \begin{cases} L_u^{\text{reg}}, & \text{if } obj_{conf_t} > 0.90 \text{ and} \\ & Conf_{bbx_t} < \theta_3^1 <= Conf_{bbx_s} \text{ and} \\ & obj_{conf_t} >= obj_{conf_s}, \\ L_u^{cls}, & \text{if } cls_{conf_t} > 0.95 \text{ and} \\ & cls_{conf_t} >= cls_{conf_s}. \end{cases} \tag{11}$$

Among the uncertain ones, if the confidence from the teacher model is above 0.90 and more than the student model, the Gaussian regression confidence is below θ_3^1 and the student, it indicates that uncertain pseudo-labels have accurate regression but lack sufficient classification scores. In such cases, only the regression loss is computed. Conversely, when the class confidence from the teacher model exceeds 0.95 and exceeds the student model, it suggests that uncertain pseudo-labels have sufficient classification scores. In this scenario, calculating the classification loss only, the equation is above. We would like to emphasize that θ_3^1 is less than θ_3. Such operations help convert uncertain pseudo-labels into high-quality pseudo-labels, thereby increasing the number of usable pseudo-labels. Moreover, the PLF utilizes a soft labeling method that effectively mitigates false positive pseudo-labels resulting from inconsistencies between classification and regression tasks. By and large, this method ensures the accurate calculation of high-quality pseudo-label loss.

Why do we not use student model predictions as high-quality pseudo-labeling when the student model outperforms the teacher model? See Supplementary Material for details. It is noteworthy that, in this paper, the augmentation for the target domain with restoration and super-resolution is a complement to PLF.

4 Experiments

In this section, we conduct experiments on four representative public dataset pairs to assess the effectiveness of the proposed model, primarily focusing on scenarios characterized by foggy, rainy, and night. The training dataset comprises data from Cityscapes and BDD100K-daytime, representing clear weather conditions; the test datasets, including the synthetic foggy dataset, Foggy Cityscapes; the real-world foggy dataset, RTTS; and the synthetic rainy dataset, Rain-Cityscapes, the real-world night dataset: BDD100K-night.

4.1 Implementation details

During the two-stage experiments, we used YOLOv5 and YOLOv7 as the baseline detectors. The experiments are conducted on two A100 GPUs with a batch size of 16. The images will resize to 960×960 during both the testing and training phases. In the generation stage, we train Cityscapes to foggy Cityscapes, Cityscapes to rain Cityscapes for 50 epochs, the Cityscapes to RTTS task, and

BDD100K daytime to BDD100K-night for 20 epochs. In the composition phase, the training epoch is 30 for all experiments. We use RIDCP [43] for image haze removal, STB [6] for image rain removal, and SMG [48] for low-light image enhancement.

4.2 Experimental Results

Results on Foggy Cityscapes. We compare our model with nine representative domain adaptive object detection methods. These methods include DM [21], HTCN [4], MIGADA [63], SIGMA++ [22], TDD [13], CMT [2], Confmix [26], R-YOLO [39], and SSDA-YOLO [62]. These methods are based on Faster R-CNN, FCOS, and YOLO detectors. As depicted in Tab. 1, Oracle indicates that training and validation are both on Foggy Cityscapes. We can see that, our proposed algorithm achieves a significant advantage over other algorithms. Ours-YOLOv5L improves 2.5% mAP, 3.8% mAP, 9.3% mAP and 6.5% mAP over SSDA-YOLOv5L, R-YOLOv5L, Confmix, and CMT, respectively. Ours-YOLOv7 also exceeds R-YOLOv7 and SSDA-YOLOv7 by more than 3.8 % mAP and 2.6 % mAP,respectively. In addition, the large object categories: bus, train, and truck all increase significantly, and the small target motorcycle category shows a rise of nearly 10% mAP, which highlights the effectiveness of our proposed loss function for both large and small targets. Surprisingly, we observe that Ours-YOLOv5L even outperforms Oracle-YOLOv5L in motorcycle and truck categories, indicating these categories are more suitable for data alignment.

Results on Rain Cityscapes. We compare our model with nine representative methods as shown in Tab. 1. In Tab. 2, we observe that our algorithm has shown a significant increase compared to other algorithms. Specifically, Ours-YOLOv5L improves 3.0% mAP, 4.9 % mAP, and 8.3 % mAP over R-YOLOv5L , SSDA-YOLOv5L, and ConfMix, respectively. Ours-YOLOv7 also over 4.7% mAP, 3.4 % mAP than R-YOLOv7 and SSDA-YOLOv7. In addition, the large target train and truck categories show a rise of more than 19% mAP in Ours-YOLOv5L, which highlights the effectiveness of our proposed loss function for large targets, even if the numbers are relatively sparse.

Results on RTTS. We further compare our model with the above ten representative methods on the real-world foggy dataset RTTS. As shown in Tab. 3, our algorithm exhibits a significant advantage over the Faster-RCNN-based, Fcos-based, and other YOLO-based algorithms. Furthermore, Ours-YOLOv5L shows an improvement of 5.9% mAP, 6.4% mAP, and 6.6% mAP over SSDA-YOLOv5L, R-YOLOv5L and ConfMix, respectively. Ours-YOLOv7 achieves a 5.7 % higher mAP than R-YOLOv7 and a 5.6 % mAP over SSDA-YOLOv7. We also observe a significant improvement in lean categories, such as bicycle and motorcycle, demonstrating the effectiveness of our algorithms in the scarcity category.

Table 1: Quantitative comparison results on the Foggy Cityscapes. "FRCNN" indicates Faster R-CNN.

Method	Detector	bus	bicycle	car	mcycle	person	rider	train	truck	mAP
source-YOLOv5	YOLOv5L	40.7	52.2	68.3	38.9	55.0	59.5	27.0	30.7	46.5
source-YOLOv7	YOLOv7	41.9	53.7	69.8	40.9	54.3	59.3	28.1	31.6	47.5
DM [21]	FRCNN	38.4	32.2	44.3	28.4	30.8	40.5	34.5	27.2	34.6
HTCN [4]	FRCNN	47.4	37.1	47.9	32.3	33.2	47.5	40.9	31.6	39.8
TDD [13]	FRCNN	53.0	49.1	68.2	38.9	50.7	53.7	45.1	35.1	49.2
CMT [2]	FRCNN	63.2	53.1	64.5	40.3	47.0	55.7	51.9	39.4	51.9
MIGADA [63]	FCOS	53.2	36.9	61.5	27.9	43.1	47.3	50.3	30.2	43.8
SIGMA++ [22]	FCOS	52.2	39.9	61.0	34.8	46.4	45.1	44.6	32.1	44.5
R-YOLO [39]	YOLOv5M	57.4	42.1	66.6	37.8	47.8	49.6	53.1	37.0	48.9
Confmix [26]	YOLOv5L	43.8	49.3	66.6	44.8	54.4	57.7	42.4	34.2	49.1
R-YOLO [39]	YOLOv5L	57.2	51.4	73.6	48.4	57.3	59.3	48.8	41.0	54.6
SSDA-YOLO [62]	YOLOV5L	63.0	53.6	74.3	47.4	60.6	62.1	48.0	37.8	55.9
Ours	YOLOv5L	62.9	54.3	75.8	48.8	**60.6**	62.0	57.8	44.7	58.4(+11.5)
R-YOLO [39]	YOLOv7	60.3	53.5	74.4	47.6	58.3	59.5	52.6	39.2	55.7
SSDA-YOLO [62]	YOLOv7	64.9	59.8	71.8	48.2	58.9	64.7	49.8	37.0	56.9
Ours	YOLOv7	**64.7**	**55.8**	**77.5**	**48.9**	60.1	**63.1**	**58.9**	**47.1**	**59.5(+12.0)**
Oracle	YOLOv5L	66.4	54.4	78.5	46.7	62.8	62.6	58.5	41.9	59.0
Oracle	YOLOv7	70.7	57.8	79.3	54.9	65.4	62.8	58.4	43.6	61.6

Table 2: Quantitative comparison results on the Rain Cityscapes.

Method	Detector	bus	bicycle	car	mcycle	person	rider	train	truck	mAP
source-YOLOv5	YOLOv5L	48.4	37.1	74.1	31.8	51.7	53.9	26.8	26.2	43.8
source-YOLOv7	YOLOv7	47.3	46.8	62.1	40.0	51.6	55.4	30.7	30.7	45.6
DM [21]	FRCNN	32.1	36.1	57.9	27.8	50.1	29.3	31.6	40.6	38.2
HTCN [4]	FRCNN	32.4	37.2	58.9	32.3	53.2	31.8	35.1	40.9	40.2
TDD [13]	FRCNN	62.7	46.7	68.9	40.8	49.1	45.6	47.6	40.6	50.3
CMT [2]	FRCNN	63.5	48.1	68.1	47.9	48.5	46.7	48.7	45.4	52.1
MIGADA [63]	FCOS	33.2	36.9	61.5	37.9	55.1	37.3	40.3	41.4	43.0
SIGMA++ [22]	FCOS	50.7	39.1	61.1	38.8	56.4	45.1	42.6	41.1	46.9
R-YOLO [39]	YOLOv5M	59.3	28.3	67.2	26.5	41.8	40.5	44.4	35.5	42.9
ConfMix [26]	YOLOv5L	67.7	34.8	72.8	35.4	48.7	45.8	52.1	35.9	49.2
SSDA-YOLO [62]	YOLOv5L	62.7	43.7	65.9	43.8	58.9	49.3	50.8	45.9	52.6
R-YOLO [39]	YOLOv5L	67.2	42.5	79.1	37.0	55.7	57.9	56.6	40.2	54.5
Ours	YOLOv5L	66.4	48.7	**79.9**	38.7	60.3	60.2	**60.1**	**45.7**	57.5(+13.7)
R-YOLO [39]	YOLOv7	64.2	46.4	79.8	37.8	57.9	59.9	51.7	43.3	55.1
SSDA-YOLO [62]	YOLOv7	70.0	51.9	67.3	44.6	56.5	65.1	48.3	46.4	56.4
Ours	YOLOv7	**72.7**	**54.9**	79.1	**48.1**	**60.6**	**61.8**	55.5	45.6	**59.8(+14.2)**
Oracle	YOLOv5L	75.7	53.7	81.7	53.3	66.5	66.6	65.9	49.2	64.1
Oracle	YOLOv7	76.9	57.9	82.4	54.5	67.3	68.5	64.5	54.7	65.8

4.3 Ablation studies

We conducted ablation experiments on RTTS using YOLOv5L. These include global feature alignment, i.e., GFA, image-to-image translation, i.e., I2I, IAoU loss, mean-teacher, i.e., MT, and pseudo-labels filtering, i.e., PLF. As depicted in Tab. 4. We can see that the PLF method exhibits substantial improvements, achieving a 4.1% mAP. This demonstrates that our joint-filtering and student-aware strategy is more effective in filtering out low-quality pseudo-labels while providing high-quality pseudo-labels to steer the student model. In addition, the proposed IAoU loss, compared to the original CIoU loss, shows an increase of

Table 3: Quantitative comparison results on the RTTS.

Method	Detector	person	car	bus	mcycle	bicycle	mAP
Baseline	YOLOv5	62.9	60.9	21.1	37.8	49.7	46.5
Baseline	YOLOv7	64.2	73.2	22.8	39.1	46.2	49.1
PDA [35]	FRCNN	37.4	54.7	17.2	22.5	38.5	34.1
DM [21]	FRCNN	37.1	56.1	27.9	26.8	40.1	37.7
HTCN [4]	FRCNN	32.4	57.2	28.9	32.3	43.2	38.8
TDD [13]	FRCNN	59.2	58.1	39.3	48.3	42.3	49.4
CMT [2]	FRCNN	67.3	65.9	30.4	37.9	57.6	51.8
MIGADA [63]	FCOS	56.2	56.9	41.5	45.9	48.1	49.7
SIGMA++ [22]	FCOS	61.1	54.8	43.6	47.2	50.3	51.4
R-YOLO [39]	YOLOv5M	60.9	72.5	32.8	52.8	37.6	51.3
ConfMix [26]	YOLOv5L	64.3	78.7	25.6	43.0	50.0	52.3
R-YOLO [39]	YOLOv5L	64.7	71.1	22.1	50.0	54.7	52.5
SSDA-YOLO [62]	YOLOv5L	62.1	73.1	34.6	54.9	40.2	53.0
Ours	YOLOv5L	70.8	78.0	21.5	**59.1**	**65.0**	58.9(+12.4)
R-YOLO [39]	YOLOv7	67.3	74.6	24.7	48.0	59.2	54.8
SSDA-YOLO [62]	YOLOv7	70.1	75.7	30.5	42.4	55.6	54.9
Ours	YOLOv7	**75.3**	**82.9**	**36.9**	52.0	55.5	**60.5(+11.4)**

Table 4: The ablation results of Cityscapes→RTTS. ✓:with, x:without.

GFA	I2I	IAoU	MT	PLF	mAP
x	x	x	x	x	46.5
✓	x	x	x	x	48.2
✓	✓	x	x	x	49.7
✓	✓	✓	x	x	51.8
✓	✓	✓	✓	x	54.8
✓	✓	✓	✓	✓	58.9

2.1% mAP, demonstrating that our proposed loss can enhance the quality of bounding box regression and is more effective in promoting feature alignment. Overall, the methods presented in this paper are effective, and when combined, they lead to a significant enhancement in model performance.

5 Conclusion

This paper introduces a generation-composition paradigm training framework to enhance object detection performance in challenging weather conditions. Specifically, in the generation phase, an optimized pre-training model generates high-quality pseudo-labels for the composition phase. With the assistance of restoration-enabled, IAoU loss, we establish global feature alignment to minimize disparities between the two domains. In the composition phase, we introduce the mean-teacher learning model with a joint-filtering and student-aware strategy to identify and select high-quality samples. The experimental results demonstrate our proposed methodology over alternative approaches in synthetic and real adverse weather conditions.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Project number 62101336); and in part by the Guangdong Basic and Applied Basic Research Foundation (Project number 2022A1515011301).

References

1. Balakrishna, S., Mustapha, A.A.: Progress in multi-object detection models: a comprehensive survey. Multim. Tools Appl. **82**(15), 22405–22439 (2023)
2. Cao, S., Joshi, D., Gui, L., Wang, Y.: Contrastive mean teacher for domain adaptive object detectors. In: CVPR. pp. 23839–23848 (2023)
3. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: CVPR. pp. 8869–8878 (2020)
4. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: CVPR. pp. 8866–8875 (2020)
5. Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., Pu, S.: Learning domain adaptive object detection with probabilistic teacher. In: ICML. vol. 162, pp. 3040–3055 (2022)
6. Chen, X., Li, H., Li, M., Pan, J.: Learning A sparse transformer network for effective image deraining. In: CVPR. pp. 5896–5905 (2023)
7. Chen, Y., Jhong, S., Hsia, C.: Roadside unit-based unknown object detection in adverse weather conditions for smart internet of vehicles. ACM Trans. Manag. Inf. Syst. **13**(4), 47:1–47:21 (2022)
8. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. pp. 3339–3348 (2018)
9. Choi, J., Chun, D., Kim, H., Lee, H.: Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In: ICCV. pp. 502–511 (2019)
10. Deng, J., Li, W., Chen, Y., Duan, L.: Unbiased mean teacher for cross-domain object detection. In: CVPR. pp. 4091–4101 (2021)
11. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.W.: Removing rain from single images via a deep detail network. In: CVPR. pp. 1715–1723 (2017)
12. Hahner, M., Dai, D., Sakaridis, C., Zaech, J., Gool, L.V.: Semantic understanding of foggy scenes with purely synthetic data. In: ITSC. pp. 3675–3681 (2019)
13. He, M., Wang, Y., Wu, J., Wang, Y., Li, H., Li, B., Gan, W., Wu, W., Qiao, Y.: Cross domain object detection by target-perceived dual branch distillation. In: CVPR. pp. 9560–9570 (2022)
14. He, Z., Zhang, L.: Domain adaptive object detection via asymmetric tri-way faster-rcnn. In: ECCV. vol. 12369, pp. 309–324 (2020)
15. Hnewa, M., Radha, H.: Multiscale domain adaptive yolo for cross-domain object detection. In: ICIP. pp. 3323–3327 (2021)
16. Hsu, H.K., Hung, W.C., Tseng, H.Y., Yao, C.H., Tsai, Y.H., Singh, M.K., Yang, M.H.: Progressive domain adaptation for object detection. WACV pp. 738–746 (2019), `https://api.semanticscholar.org/CorpusID:198167281`
17. Huang, S.C., Le, T.H., Jaw, D.W.: Dsnet: Joint semantic learning for object detection in inclement weather conditions. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(8) (2021)

18. Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: CVPR. pp. 5001–5009. Computer Vision Foundation / IEEE Computer Society (2018)
19. Kalwar, S., Patel, D., Aanegola, A., Konda, K.R., Garg, S., Krishna, K.M.: Gdip: Gated differentiable image processing for object detection in adverse conditions. In: ICRA. pp. 7083–7089 (2023)
20. Khodabandeh, M., Vahdat, A., Ranjbar, M., Macready, W.G.: A robust learning approach to domain adaptive object detection. CoRR **abs/1904.02361** (2019)
21. Kim, T., Jeong, M., Kim, S., Choi, S., Kim, C.: Diversify and match: A domain adaptive representation learning paradigm for object detection. In: CVPR. pp. 12456–12465. Computer Vision Foundation / IEEE (2019)
22. Li, W., Liu, X., Yuan, Y.: SIGMA++: improved semantic-complete graph matching for domain adaptive object detection. TPAMI **45**(7), 9022–9040 (2023)
23. Li, Y., Dai, X., Ma, C., Liu, Y., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: CVPR. pp. 7571–7580 (2022)
24. Lin, W., Chu, J., Leng, L., Miao, J., Wang, L.: Feature disentanglement in one-stage object detection. Pattern Recognit. **145**, 109878 (2024)
25. Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., Zhang, L.: Image-adaptive yolo for object detection in adverse weather conditions. In: AAAI. vol. 36, pp. 1792–1800 (2022)
26. Mattolin, G., Zanella, L., Ricci, E., Wang, Y.: Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In: WACV. pp. 423–433 (2023)
27. Mattolin, G., Zanella, L., Ricci, E., Wang, Y.: Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In: WACV. pp. 423–433 (2023)
28. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: CVPR. pp. 4500–4509 (2018)
29. Panagiotakopoulos, T., Dovesi, P.L., Härenstam-Nielsen, L., Poggi, M.: Online domain adaptation for semantic segmentation in ever-changing conditions. In: ECCV. vol. 13694, pp. 128–146 (2022)
30. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV. pp. 319–345 (2020)
31. Qiu, Z., Zheng, P., Tao, X., Wu, X.: Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems **30**(11), 3212–3232 (2019)
32. Rodriguez, A.L., Mikolajczyk, K.: Domain adaptation for object detection via style consistency. arXiv preprint arXiv:1911.10033 (2019)
33. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: CVPR. pp. 6956–6965 (2019)
34. Siliang, M., Yong, X.: Mpdiou: A loss for efficient and accurate bounding box regression. arXiv preprint arXiv:2307.07662 (2023)
35. Sindagi, V.A., Oza, P., Yasarla, R., Patel, V.M.: Prior-based domain adaptive object detection for hazy and rainy conditions. In: ECCV. vol. 12359, pp. 763–780 (2020)
36. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: ICLR (2017)
37. Vs, V., Gupta, V., Oza, P., Sindagi, V.A., Patel, V.M.: Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection. In: CVPR. pp. 4516–4526 (2021)

38. Wang, L., Wang, J., Zhu, L., Fu, H., Li, P., Cheng, G., Feng, Z., Li, S., Heng, P.: Dual multiscale mean teacher network for semi-supervised infection segmentation in chest CT volume for COVID-19. IEEE Trans. Cybern. **53**(10), 6363–6375 (2023)
39. Wang, L., Qin, H., Zhou, X., Lu, X., Zhang, F.: R-yolo: A robust object detector in adverse weather. IEEE Transactions on Instrumentation and Measurement **72**, 1–11 (2022)
40. Wang, Y., Song, Y., Ma, C., Zeng, B.: Rethinking image deraining via rain streaks and vapors. In: ECCV. vol. 12362, pp. 367–382 (2020)
41. Wang, Y., Yan, X., Zhang, K., Gong, L., Xie, H., Wang, F.L., Wei, M.: TogetherNet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning. Computer Graphics Forum (2022)
42. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: CVPR. pp. 17662–17672 (2022)
43. Wu, R., Duan, Z., Guo, C., Chai, Z., Li, C.: RIDCP: revitalizing real image dehazing via high-quality codebook priors. In: CVPR. pp. 22282–22291 (2023)
44. Xia, Z., Yan, C., Wei, S., Hong, G., Ping, Y.: Object detection in 20 years: A survey. Proceedings of the IEEE **111**(3), 257–276 (2023)
45. Xiong, F., Tian, J., Hao, Z., He, Y., Ren, X.: SCMT: self-correction mean teacher for semi-supervised object detection. In: IJCAI. pp. 1488–1494 (2022)
46. Xu, B., Chen, M., Guan, W., Hu, L.: Efficient teacher: Semi-supervised object detection for yolov5. arXiv preprint arXiv **abs/2302.07577** (2023)
47. Xu, C., Zhao, X., Jin, X., Wei, X.: Exploring categorical regularization for domain adaptive object detection. In: CVPR. pp. 11721–11730 (2020)
48. Xu, X., Wang, R., Lu, J.: Low-light image enhancement via structure modeling and guidance. In: CVPR. pp. 9893–9903 (2023)
49. Yang, Q., Wei, X., Wang, B., Hua, X., Zhang, L.: Interactive self-training with mean teachers for semi-supervised object detection. In: CVPR. pp. 5941–5950 (2021)
50. Yasarla, R., Priebe, C.E., Patel, V.M.: ART-SS: an adaptive rejection technique for semi-supervised restoration for adverse weather-affected images. In: ECCV. vol. 13678, pp. 699–718 (2022)
51. Ye, T., Zhang, Y., Jiang, M., Chen, L., Liu, Y., Chen, S., Chen, E.: Perceiving and modeling density for image dehazing. In: ECCV. vol. 13679, pp. 130–145 (2022)
52. Yoo, J., Chung, I., Kwak, N.: Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In: ECCV. vol. 13693, pp. 691–708 (2022)
53. Yu, H., Zheng, N., Zhou, M., Huang, J., Xiao, Z., Zhao, F.: Frequency and spatial dual guidance for image dehazing. In: ECCV. vol. 13679, pp. 181–198 (2022)
54. Yu, J., Liu, J., Wei, X., Zhou, H., Nakata, Y., Gudovskiy, D., Okuno, T., Li, J., Keutzer, K., Zhang, S.: Mttrans: Cross-domain object detection with mean teacher transformer. In: ECCV. pp. 629–645 (2022)
55. Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B.: A survey of modern deep learning based object detection models. Digital Signal Processing **126**, 103514 (2022)
56. Zhang, K., Luo, W., Ren, W., Wang, J., Zhao, F., Ma, L., Li, H.: Beyond monocular deraining: Stereo image deraining via semantic understanding. In: ECCV. vol. 12372, pp. 71–89 (2020)
57. Zhang, S., Tuo, H., Hu, J., Jing, Z.: Domain adaptive yolo for one-stage cross-domain detection. In: ACML. pp. 785–797 (2021)
58. Zhang, Y., Shi, Z., Zhang, Y.: Adioc loss: An auxiliary descent ioc loss function. Engineering Applications of Artificial Intelligence **116**, 105453 (2022)

59. Zhao, G., Li, G., Xu, R., Lin, L.: Collaborative training between region proposal localization and classification for domain adaptive object detection. In: ECCV. vol. 12363, pp. 86–102 (2020)
60. Zhao, Z., Wei, S., Chen, Q., Li, D., Yang, Y., Peng, Y., Liu, Y.: Masked retraining teacher-student framework for domain adaptive object detection. In: ICCV. pp. 18993–19003 (2023)
61. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: AAAI. vol. 34, pp. 12993–13000 (2020)
62. Zhou, H., Jiang, F., Lu, H.: SSDA-YOLO: semi-supervised domain adaptive YOLO for cross-domain object detection. Comput. Vis. Image Underst. **229**, 103649 (2023)
63. Zhou, W., Du, D., Zhang, L., Luo, T., Wu, Y.: Multi-granularity alignment domain adaptation for object detection. In: CVPR. pp. 9571–9580 (2022)
64. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)