ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240471

I-Adapt: Using IoU Adapter to Improve Pseudo Labels in Cross-Domain Object Detection

Qifeng Zhang[a], Changjian Chen[a,*], Zhizhong Liu[a] and Zhuo Tang[a,*]

[a]Hunan University

Abstract. Domain adaptation has been extensively explored in object detection. Through the utilization of self-training and the decoupling of adversarial feature learning from the training of the detector, current methods make detectors more transferable and ensure their discriminability. However, the presence of low-quality pseudo labels during self-training introduces noises to the training phase and thus degrades the model performance. To tackle this challenge, we introduce an I-adapt framework, whose IoU Adapter accurately predicts the Intersection over Union (IoU) between predicted boxes and their corresponding ground-truth boxes in both source and target domains. This enables an effective measure for the pseudo-label quality. Based on this measure, we propose a re-weighting strategy, which enforces the detector to focus on learning from high-quality pseudo labels. We achieve state-of-the-art (SOTA) performance in several cross-domain object detection tasks, proving the effectiveness of I-adapt.

1 Introduction

With the advancement of deep learning and the release of numerous visual datasets [5, 8, 20], object detection [1, 11, 31, 33, 36] has made great progress over the past decade. It has been used in various real-world applications, such as autonomous driving [35], video processing [19], and remote sensing [25]. In these applications, it has been observed that when the feature distribution in the test data differs greatly from that of the training data, the model performance will degrade significantly.

Cross-domain object detection (CDOD) emerges as an effective way to tackle this challenge. It adapts an object detector trained on a labeled source domain to an unlabeled target domain [4, 7]. Many methods have been proposed for CDOD in recent years. Among them, the self-training-based methods, which treat detections of detector in the target domain as pseudo labels, have gained widespread attention because of their effectiveness [30]. A representative technique along this line is enhancing self-training with adversarial feature learning [18, 26]. For example, a SOTA method, D-adapt [18], employs adapters to decouple the adversarial feature learning from the detector training process. This preserves the transferability while ensuring the discriminability of detectors (*e.g.*, detections in both source and target domains are well separated in Fig. 1B). Despite the effectiveness of this technique, it treats all pseudo labels equally, resulting in low-quality ones harming the training process (*e.g.*, Fig. 1A). To mitigate this issue, recent methods, such as Harmonious Teacher [7], utilize the re-weighting strategy. This strategy effectively reduces the negative influences of low-quality pseudo labels

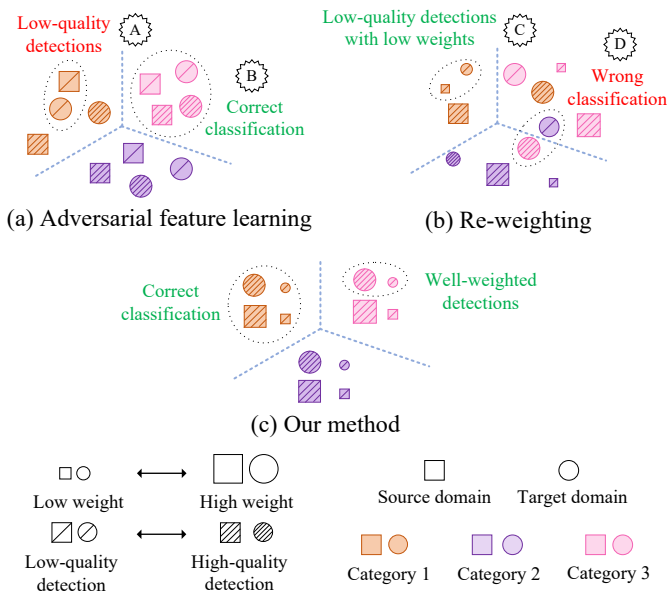

Figure 1: Comparison of existing techniques and our method.

(*e.g.*, Fig. 1C). However, as the re-weighting strategy pays less attention to transferability and discriminability, detections of different categories cannot be well separated when the domain gap is large (*e.g.*, Fig. 1D). The incompatibility of transferability/discriminability and high-quality pseudo labels in exiting self-training-based techniques inspires us to consider answering the following question in this study: **how to ensure the quality of pseudo labels while improving the transferability and discriminability of the detectors?**

To this end, we propose a self-training framework, I-adapt. I-adapt aims to get a reliable quality metric of pseudo labels in the target domain while keeping the transferability and discriminability of the detectors. Observing that treating the classification score as the quality of the pseudo label [6, 21, 26] is not suitable (*e.g.*, in Fig. 2a, the orange bounding box is of higher quality than the white one but has a lower classification score), we expect to find a more reliable metric. In [40], researchers found that by aligning the classification scores with the IoUs between the detections and the corresponding ground-truth boxes, object detectors can achieve a more accurate quality ranking of candidate detections. From such observation, we combine this IoU-classification consistency constraint with D-adapt, which uses decoupled adversarial feature learning to ensure the transferability/discriminability of the detectors. The IoU-classification consistency constraint ensures the classification score is better aligned with the pseudo-label quality in the source domain,

* Corresponding authors. Email: {changjianchen, ztang}@hnu.edu.cn

and the decoupled adversarial feature learning transfers this ability to the target domain while ensuring discriminability. With such a joint optimization, the classification score becomes a more reliable quality metric for pseudo labels in the target domain. For example, in Fig. 2b, with our method, the green box of high quality has a higher calibrated confidence score than the blue one of low quality. Based on this improved classification score (named quality score), we designed a re-weighting strategy that makes the detector treat pseudo labels of various qualities differently and focus on learning from high-quality ones. As the bounding box pseudo labels are calculated after the generation of category pseudo labels and quality scores, they may not be consistent. Therefore, we further developed a Mutual Improvement method to adjust the bounding boxes, quality scores, and category pseudo labels mutually for consistency. We conducted several experiments on four CDOD tasks to validate the effectiveness of our method compared with the SOTA CDOD methods.

In summary, the main contributions of our work are:

- **An improved self-training-based framework** that gets a reliable quality metric of pseudo labels in the target domain while keeping the transferability and discriminability of the detectors.
- **A re-weighting strategy** that makes the model treat pseudo labels of various qualities differently and focus on learning from high-quality ones.
- **Extensive experiments** that show that our method achieves state-of-the-art performance on several CDOD tasks.

2 Related Work

2.1 Object Detection

Based on the model architecture, current object detectors can be generally categorized into three types: CNN-based, transformer-based, and CNN-transformer hybrid detectors. Our work is most related to CNN-based object detectors. Accordingly, this section is dedicated to reviewing the relevant literature within this category.

CNN-based detectors can be generally categorized into two types: one-stage and two-stage detectors. One-stage detectors directly regress the localizations and categories of objects from images. For example, YOLO series object detectors [11, 34] are known for their effective balance between real-time detection and accuracy. RetinaNet [27] applies the focal loss to handle the extreme foreground-background class imbalance encountered during the training. FCOS [33] applies a fully convolutional network in a per-pixel prediction fashion to simplify the detection procedure. Two-stage detectors, such as Faster-RCNN, introduce the region proposal network (RPN) [31] to generate class-agnostic proposals, and then refine these proposals to get the final localizations and classifications. Mask R-CNN [12] detects objects in an image while concurrently producing a segmentation mask for each instance.

The main goal of the aforementioned methods is to improve the detector performance in labeled datasets. However, when the test data differs visually from the training data, the detector performance will drop greatly. In this paper, we focus on improving the cross-domain generalization ability of both one-stage and two-stage detectors.

2.2 Cross-Domain Object Detection (CDOD)

CDOD adapts an object detector trained on a labeled source domain to an unlabeled target domain. Recent popular works can be generally grouped into four types: vanilla adversarial feature learning [13, 14, 15, 23, 32, 37, 39, 44], self-training strategy [7, 18, 26], graph reasoning [9, 24], and style transfer [6, 16]. In this paper, we focus on two lines: adversarial feature learning and self-training strategy. The main idea of adversarial feature learning is extracting domain-invariant features from the source and target domain to make features more transferable. The domain-invariant features can be extracted from different levels, such as image-level [13, 23, 32, 39], pixel-level [15, 44] and instance-level [44]. Though successful, directly applying adversarial feature learning in detectors could lead to the distortion of semantic features and the reduction of detector discriminability [10], leading to the sub-optimal performance. Self-training strategy aims to generate reliable pseudo labels in the target domain to retrain the detector in a supervised manner. Unbiased Mean Teacher (UMT) [6] utilizes Teacher-Student paradigm (TS) with pixel-level adaptation to improve the transferability of detectors towards the target domain. Adaptive Teacher (AT) [26] combines the self-training strategy with adversarial feature learning to narrow the domain gap between source and target domains. D-adapt [18] utilizes adapters to separate adversarial feature learning from the detector training process, thus ensuring transferability and discriminability. Despite the effectiveness of UMT, AT and D-adapt, their performance is limited by the low-quality pseudo labels. To address this issue, Harmonious Teacher [7] proposes a re-weight strategy to reduce the negative influences of low-quality pseudo labels. However, when the domain gap is large, the weights will be inconsistent with the quality of pseudo labels, and pseudo labels of different categories cannot be well separated, thus misguiding the model training.

In this paper, we propose an improved adapter-based self-training framework, which generates reliable quality weights for pseudo labels and ensures the transferability/discriminability of the detector.

(a) D-adapt (b) I-adapt (ours) (c) Ground truth

Figure 2: the visualizations of pseudo labels generated by D-adapt [18] (a) and I-adapt (b). From the visualizations, we can see that the scores of pseudo labels generated by I-adapt are more consistent with the pseudo-label quality than these generated by D-adapt, and thus are more suitable to be the weights of pseudo labels.

3 Preliminary and Background

3.1 Problem Setting

We assume there are n_s labeled samples $\mathcal{D}_s = \{(X_s^i, B_s^i, C_s^i)\}_{i=1}^{n_s}$ from the source domain, and n_t unlabeled samples $\mathcal{D}_t = \{(X_t^i,)\}_{i=1}^{n_t}$ from the target domain. Here X_s^i and X_t^i denote the i-th image from the source domain and the target domain, respectively. B_s^i and C_s^i denote the bounding boxes and category labels of objects in the i-th image, respectively.

3.2 D-adapt: A CDOD Method Without Re-weighting

As depicted on the left side of Fig. 3, D-adapt consists of three networks, the detector G^{det}, the Category Adapter (CA), and the Bbox Adapter (BA). CA and BA are used to generate pseudo labels for training G^{det}. Taking Faster-RCNN [31] as an example, the training process of D-adapt unfolds as follows.

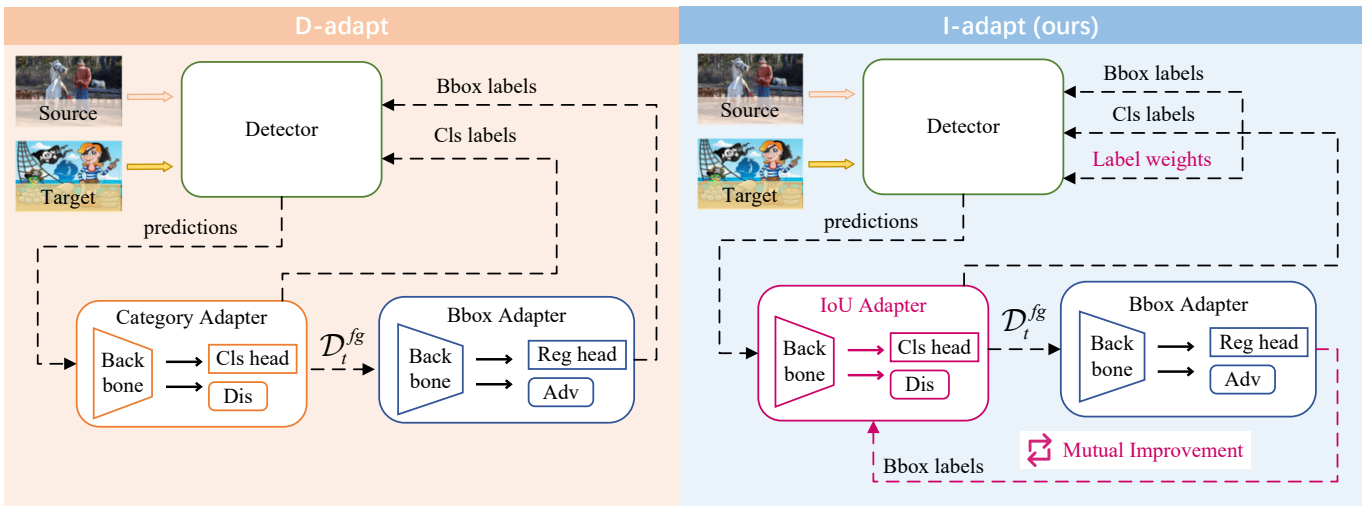

Figure 3: Comparison between D-adapt and I-adapt. D-adapt consists of three networks, the detector, the Category Adapter (CA), and the Bbox Adapter (BA). CA and BA can generate category and bounding box pseudo labels respectively in the target domain. In I-adapt, we replace CA with the proposed IoU Adapter, which can generate category pseudo labels and quality scores, and propose Mutual Improvement to handle the problem of inconsistency in pseudo labels. The details will be discussed in Sec. 4.

(1) Pre-train the detector. For an effective initialization, D-adapt first pre-trains G^{det} on the source domain $\mathcal{D}_s$ with the training loss of Faster-RCNN L_s^{det}, which consists of RPN loss L^{RPN}, RoI classification loss L_{cls}^{ROI}, and RoI localization loss L_{reg}^{ROI}.

(2) Generate new data distributions for training adapters. Then, the pre-trained G^{det} is used to generate two new data distributions, the source and the target detection distributions $\mathcal{D}_s^{pred}$ and $\mathcal{D}_t^{pred}$. Each detection in them consists of a region x^{det} cropped in image X, its bounding box b^{det}, and predicted category c^{det}. Since samples in $\mathcal{D}_s$ are labeled, each detection in $\mathcal{D}_s^{pred}$ is annotated with a ground-truth bounding box b_s and a ground-truth category label c_s, following the annotation strategy of RoI in Faster-RCNN.

(3) Train CA to generate category pseudo labels. By utilizing adversarial feature learning, CA is trained on $\mathcal{D}_s^{pred}$ and $\mathcal{D}_t^{pred}$ to generate category pseudo label $\hat{c}_t$ for each region $x_t^{det} \in \mathcal{D}_t^{pred}$.

(4) Train BA to generate bounding box pseudo labels. According to c_s and $\hat{c}_t$, D-adapt separates the foreground detections $\mathcal{D}_s^{fg}$ and $\mathcal{D}_t^{fg}$ from $\mathcal{D}_s^{pred}$ and $\mathcal{D}_t^{pred}$, respectively. Then, an IoU disparity discrepancy method is applied to train BA on $\mathcal{D}_s^{fg}$ and $\mathcal{D}_t^{fg}$. This method encourages the feature extractor of BA to output domain-invariant features. After the training, BA is utilized to generate a bounding box pseudo label $\hat{b}_t$ for each $x_t^{det} \in \mathcal{D}_t^{fg}$.

(5) Train the detector with pseudo labels. Based on the category and bounding box pseudo labels generated by the two adapters, G^{det} is trained with loss L_t^{det} in the target domain in a supervised manner.

Steps (2) to (5) are repeated for T iterations until the training process converges. Note that the two adapters (CA and BA) are only used in the training phase.

Through these steps, D-adapt decouples the adversarial adaption (steps (3) and (4)) from the training of the detector (step (5)). This ensures the transferability and discriminability of the detector, which makes D-adapt achieve competitive performance on various datasets [18]. However, when analyzing the results of D-adapt, we identify one issue that limits its performance: **many pseudo labels generated by CA and BA are of low-quality, which harms the training of the detector in step (5).** To tackle this problem, we propose a new self-training-based CDOD framework, I-adapt, with a re-weighting strategy, enabling the detector to focus on learning from high-quality pseudo labels and reducing the effect of low-quality ones [29, 38]. As its simplicity, this method can be easily integrated with existing CDOD methods to improve detection performance.

Algorithm 1: I-adapt Training Pipeline

Input: Source domain $\mathcal{D}_s$, target domain $\mathcal{D}_t$, number of iterations T
Output: Cross-domain object detector G^{det}

initialize the object detector G^{det} by optimizing with L_s^{det};
for $t = 1 \rightarrow T$ **do**
  generate detections $\mathcal{D}_s^{pred}$ and $\mathcal{D}_t^{pred}$ for each sample in $\mathcal{D}_s$ and $\mathcal{D}_t$ by G^{det};
  for *each mini-batch in* $\mathcal{D}_s^{pred}$ *and* $\mathcal{D}_t^{pred}$ **do**
    train the IoU Adapter G^{IA};
  end
  generate category pseudo label $\hat{c}_t$ and quality score $\hat{q}_t$ for each detection in $\mathcal{D}_t^{pred}$;
  separate foreground detections $\mathcal{D}_s^{fg}$ and $\mathcal{D}_t^{fg}$ from $\mathcal{D}_s^{pred}$ and $\mathcal{D}_t^{pred}$;
  for *each mini-batch in* $\mathcal{D}_s^{fg}$ *and* $\mathcal{D}_t^{fg}$ **do**
    train the Bbox Adapter G^{BA};
  end
  generate bbox pseudo label for each detection in $\mathcal{D}_t^{fg}$;
  generate consistent detections $\mathcal{D}_t^{MI}$ from $\mathcal{D}_t^{fg}$;
  train the object detector G^{det} by optimizing with L_t^{det};
end

4 Proposed Method

In this section, we use D-adapt as the base model to demonstrate the basic idea of the proposed I-adapt. As depicted on the right side of Fig. 3, I-adapt comprises two significant enhancements. First, it replaces the Category Adapter with the IoU Adapter (Sec. 4.1), which can generate category pseudo label and quality score for each detection in the target domain. As the bounding box (referred to as bbox) pseudo labels may not be consistent with the category pseudo labels

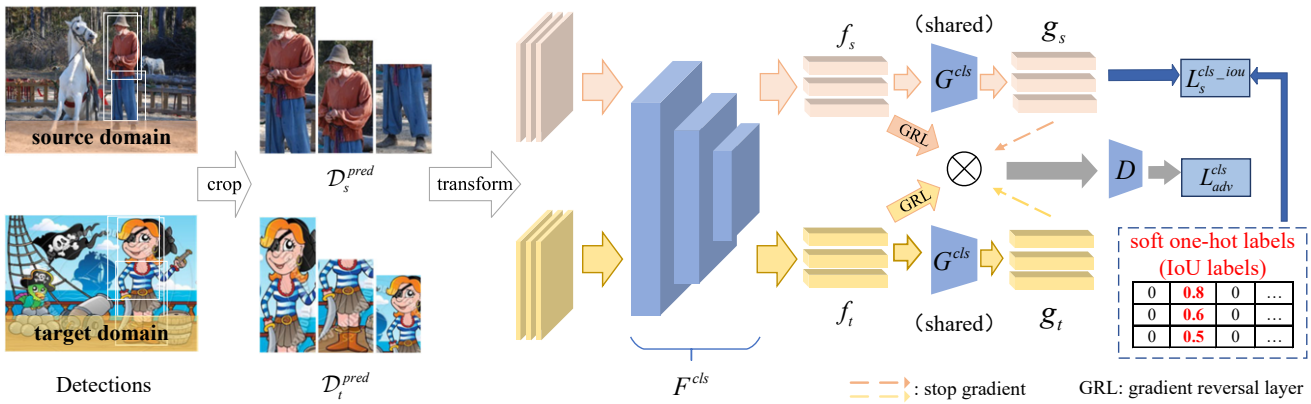

Figure 4: **An overview of our IoU Adapter.** The source domain and the target domain data share a common network, which is composed of three parts: the feature extractor F^{cls}, the classification head G^{cls}, and the domain discriminator D.

and quality scores during the training process, a Mutual Improvement method (Sec. 4.2) generating consistent detections $\mathcal{D}_t^{MI}$ is added after Step (4) in D-adapt. In line with I-adapt, we define the training losses of the detector (Sec. 4.3) with some modifications compared to D-adapt. A brief process of I-adapt is summarized in Algorithm 1.

4.1 IoU Adapter

As we discussed in Sec. 1, by aligning the classification score with gt_IoU (*i.e.*, the IoU between a detection and its corresponding ground truth box), detectors can achieve a more accurate quality ranking of candidate detections. However, as the ground-truth box in the target domain is not available, gt_IoU can not be calculated in the target domain. To tackle this issue, we propose the IoU Adapter (IA) to replace CA in D-adapt. It generates category pseudo labels while predicting gt_IoU accurately in the source and target domain.

Fig. 4 shows an overview of IA. Motivated by CDAN [28], it consists of three parts: a feature extractor F^{cls}, a domain discriminator D, and a classification head G^{cls}. Specifically, F^{cls} is used to extract features from images in both source and target domains. G^{cls} outputs the classification scores for input images. D serves to differentiate whether the features originate from the source or target domain. IA is trained using the following loss:

$$L^{IA} = L_s^{cls\_iou} + \lambda L_{adv}^{cls}. \tag{1}$$

Here L_{adv}^{cls} is the Adversarial Domain Adaption loss to make the learned features transferable and discriminative. $L_s^{cls\_iou}$ is the IoU-Classification Consistency loss to ensure that the classification scores are better aligned with the quality of pseudo labels. λ is a weight to balance G^{cls} and D.

Adversarial Domain Adaptation loss. This loss encourages F^{cls} to extract transferable and discriminative features f. To make features f more *transferable*, we encourage D to distinct the features from the source and target domains while encouraging F^{cls} to fool D, thereby enabling F^{cls} to extract domain-invariant features. In addition, to make features f more *discriminative*, classification information g is also input to D, making f align according to their respective categories instead of the dominant ones. Accordingly, the Adversarial Domain Adaptation loss can be written as:

$$\begin{aligned} L_{adv}^{cls} = -(&\mathbb{E}_{x_s^{det}\sim\mathcal{D}_s^{pred}} \log[D(f_s, g_s)] + \\ &\mathbb{E}_{x_t^{det}\sim\mathcal{D}_t^{pred}} \log[1 - D(f_t, g_t)]). \end{aligned} \tag{2}$$

IoU-Classification Consistency loss. This loss makes the classification head G^{cls} output classification score consistent with gt_IoU. To achieve this, we utilize the varifocal loss [40] during training. Specifically, in the training phase, we select a region x_s^{det} from $\mathcal{D}_s^{pred}$ and get its class prediction distribution $P = \{p_1, p_2, ..., p_n\}$ by feeding x_s^{det} into IA. Here p_i is the classification score for the i-th class, and n is the number of classes, including the background class. If using the standard one-hot ground-truth label encoding, its corresponding label is $Y = \{0, 0, \ldots, y_{c_s}, \ldots, 0\}$, where c_s is the ground-truth class, and $y_{c_s} = 1$. Instead, we change y_{c_s} in IA as follows,

$$y_{c_s} = \begin{cases} gt\_IoU_s & x_s^{det} \text{ is forground,} \\ 1 - gt\_IoU_s & x_s^{det} \text{ is background.} \end{cases} \tag{3}$$

Note that for the background region $x_s^{det} \in \mathcal{D}_s^{bg}$, we use the maximum IoU between it and all ground-truth boxes in the corresponding image as the value of gt_IoU, and use $1 - gt\_IoU$ as its y_{c_s}. The rationale behind this is that the smaller the IoU between the predicted bounding box and the ground-truth boxes, the more likely it is a background region.

Based on the modified ground-truth one-hot labels, the varifocal loss is expressed as follows:

$$L^{vfl}(y_i, p_i) = \begin{cases} -y_i(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) & y_i > 0, \\ -\alpha p_i^{\gamma} \log(1 - p_i) & y_i = 0. \end{cases} \tag{4}$$

Here p_i and y_i are the elements in P and Y, respectively. α and γ are two hyper-parameters to balance the losses of positive and negative samples. This varifocal loss aligns the classification scores with gt_IoUs, therefore making the classification scores a better measure for the quality of pseudo labels. Overall, the IoU-Classification Consistency loss can be written as:

$$L_s^{cls\_iou} = \mathbb{E}_{x_s^{det}\sim\mathcal{D}_s^{pred}} \frac{1}{n} \sum_{i=1}^{n} L^{vfl}(y_i, p_i). \tag{5}$$

After training IA with L^{IA}, $arg\max\{P\}$ of a region x_t^{det} is regarded as its category pseudo label $\hat{c}_t$, and $max\{P\}$ is regarded as its quality score $\hat{q}_t$.

4.2 Mutual Improvement

Got the category pseudo labels and quality scores, we further utilize BA to generate the bounding box pseudo label $\hat{b}_t$ for each detection

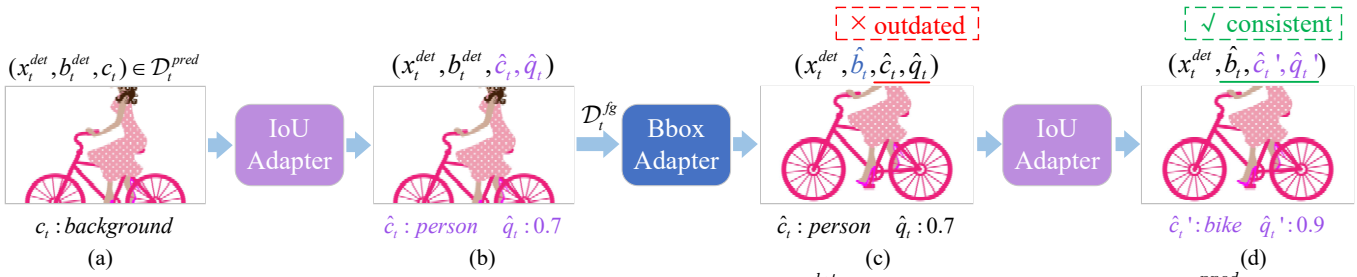

Figure 5: **An overview of pseudo label processing in I-adapt.** (a) Given a region x_t^{det} in the target detection distribution $\mathcal{D}_t^{pred}$, (b) we first use the IA to generate a relatively accurate category pseudo label $\hat{c}_t$ and a quality score of pseudo label $\hat{q}_t$ for it. (c) Then, BA is used to generate a bounding box pseudo label $\hat{b}_t$ for each foreground region $x_t^{det} \in \mathcal{D}_t^{fg}$. Since the bounding box is updated, the category pseudo label $\hat{c}_t$ and the quality score $\hat{q}_t$ may be outdated (*e.g.*, the object in the updated bounding box is more likely to be a "bike" instead of a "person"). (d) Thus, we utilize a Mutual Improvement step to update $\hat{b}_t$, $\hat{c}_t$, and $\hat{q}_t$ iteratively to make them consistent with each other.

in $\mathcal{D}_t^{fg}$. However, since the category pseudo label $\hat{c}_t$ and the quality score $\hat{q}_t$ are generated by IA based on the bounding box b_t^{det} (Fig. 5b), $\hat{c}_t$ and $\hat{q}_t$ will be inconsistent with $\hat{b}_t$ (Fig. 5c).

To address this issue, a Mutual Improvement (MI) method is introduced. Specifically, the region x_t^{det} of each outdated detection is input into IA to generate new category pseudo label $\hat{c}_t{}'$ and quality score $\hat{q}_t{}'$, ensuring their consistency with $\hat{b}_t$ (Fig. 5d). We denote the new data distribution generated by MI as $\mathcal{D}_t^{MI}$. Each prediction in $\mathcal{D}_t^{MI}$ is denoted as $(x_t^{det}, \hat{b}_t, \hat{c}_t{}', \hat{q}_t{}')$. Although we can further utilize BA and IA in turn to improve the consistency, it will increase computational time. In practice, we find that an "IA-BA-IA" process is enough for downstream training.

4.3 Detector Training

Using Faster-RCNN as an example, we provide the details of the detector training losses used in I-adapt. As most of the losses are similar to that of D-adapt, we only describe the different parts: modification in RoI classification loss and training in the target domain.

Modification in RoI classification loss. In line with I-adapt, we use the varifocal loss (Eq. (4)) to replace the RoI classification loss L_{cls}^{ROI} in the model training. Specifically, when training the detector in the source domain, for each detection, the ground truth Y for varifocal loss is obtained in the same way as for IA. While in the target domain, Y is assigned to $\{0, 0, \ldots, y_{c_t^{psd}}, \ldots, 0\}$, where $y_{c_t^{psd}}$ is assigned to q_t^{psd}. c_t^{psd} is the category pseudo label in the target domain. q_t^{psd} is the quality score of each pseudo label in the target domain.

Training in the target domain. Similar to D-adapt, we also only train the RoI head in the target domain. In addition, the losses are only calculated on the regions where the pseudo labels are located. Based on the quality scores and pseudo labels generated by IA and BA, we train the detector with L_t^{det}:

$$\begin{aligned} L_t^{det} = \mathbb{E}_{\mathcal{D}_t^{bg}} e^{\hat{q}_t - 1} L_{cls}^{ROI}(X_t, b_t^{det}, \hat{c}_t, \hat{q}_t) + \\ \mathbb{E}_{\mathcal{D}_t^{MI}} e^{\hat{q}_t{}' - 1} L_{cls}^{ROI}(X_t, \hat{b}_t, \hat{c}_t{}', \hat{q}_t{}'). \end{aligned} \tag{6}$$

We assign a weight $e^{q_t^{psd}-1}$ for each pseudo label, enabling the detector to focus on learning from high-quality pseudo labels and reducing the effect of low-quality ones.

5 Experiments and Analysis

5.1 Datasets

We conducted the experiments on six datasets, including Cityscapes [5], Foggy Cityscapes [5], Sim10k [20], PASCAL VOC [8], Comic2k [17], and Clipart [17]. Cityscapes comprises a collection of outdoor street scenes captured under clear weather. Foggy Cityscapes is a synthetic foggy dataset generated from Cityscapes. Both of them contain 2,975 training images and 500 validation images of eight categories. Sim10k contains 10,000 images of driving scenes derived from the video game Grand Theft Auto V (GTA5). PASCAL VOC is a real-world dataset with 16,551 training images of 20 categories. Clipart consists of 1,000 cartoon images and shares 20 categories with PASCAL VOC. Comic2k has 1,000 comic images for training and another 1,000 for testing, sharing 6 categories with PASCAL VOC. Based on these datasets, we followed [18] to construct four CDOD tasks, including Cityscapes→Foggy Cityscapes, Sim10k→Cityscapes, PASCAL VOC→Clipart, and PASCAL VOC→Comic2k. To determine the degree of domain gaps of these tasks, we trained detectors on the source domains and tested them on both source and target domains. According to the performance gaps on these two domains, we found that the tasks Cityscapes→Foggy Cityscapes and Sim10k→Cityscapes have smaller domain gaps, and PASCAL VOC→Clipart/Comic2k have larger domain gaps.

5.2 Implementation Details

For a fair comparison, We compared I-adapt with the other methods based on Faster-RCNN detector with ResNet101 or VGG16 backbone, and the results are reported by mean Average Precision (mAP) with a threshold of 0.5. We conducted all experiments using an RTX 4090 GPU. The α and γ in Eq. 4 are set to 0.75 and 1.5 in small domain gap tasks (Cityscapes → Foggy Cityscapes, sim10k → Cityscapes), and 0.75 and 1.75 in large domain gap tasks (PASCAL VOC → Clipart, PASCAL VOC → Comic2k). The IoU Adapter (IA) is initialized with a ResNet101 pre-trained for ten epochs in the source domain. Then, we train IA in both domains for four epochs. The rest parameters were set as the same in D-adapt [18]. The models are evaluated on the target domain.

5.3 Result Analysis

As shown in Tables 1, 2, 3, and 4, we compare the proposed method with SOTA methods in four CDOD tasks. We also present the evaluation results of the detector trained on the source domain only in the row "source-only." The row "oracle" shows the results of the model that was trained and evaluated both on the target domain.

Real-to-Artistic Adaptation (PASCAL VOC→Clipart and PASCAL VOC→Comic2k). Tables 1 and 2 show the results of two CDOD tasks, PASCAL VOC→Clipart and PASCAL

Table 1: Results from PASCAL VOC to Clipart. ResNet101 is used as the backbone for all models.

Method	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP (%)
source-only	32.7	52.8	22.6	33.5	36.5	44.8	33.9	17.1	43.5	13.8	32.0	9.6	35.7	49.6	39.3	42.7	4.5	18.9	37.4	40.2	32.1
SWDA [32]	26.2	48.5	32.6	33.7	38.5	54.3	37.1	18.6	34.8	58.3	17.0	12.5	33.8	65.5	61.6	52.0	9.3	24.9	54.1	49.1	38.1
UMT [6]	39.6	59.1	32.4	35.0	45.1	61.9	48.4	7.5	46.0	**67.6**	21.4	29.5	48.2	75.9	70.5	56.7	25.9	28.9	39.4	43.6	44.1
MGA [44]	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_	44.8
TIA [42]	42.2	66.0	36.9	37.3	43.7	71.8	49.7	18.2	44.9	58.9	18.2	**29.1**	40.7	**87.8**	67.4	49.7	27.4	27.8	57.1	50.6	46.3
AT [26]	33.8	60.9	38.6	49.4	52.4	53.9	56.7	7.5	52.8	63.5	34.0	25.0	**62.2**	72.1	77.2	57.7	27.2	**52.0**	55.7	54.1	49.3
VSDSN [2]	53.6	64.3	**43.1**	41.0	44.7	76.6	53.6	29.3	45.8	62.5	28.4	25.0	44.8	84.0	69.0	44.6	25.7	37.0	59.3	**64.4**	49.8
MILA [22]	28.3	**80.0**	35.2	42.0	**56.7**	44.6	**61.5**	9.3	**59.0**	62.2	**44.0**	24.2	60.9	77.0	**79.0**	**62.5**	**29.3**	45.1	49.8	47.8	49.9
D-adapt [18]	56.4	63.2	42.3	40.9	45.3	77.0	48.7	25.4	44.3	58.4	31.4	24.5	47.1	75.3	69.3	43.5	27.9	34.1	60.7	64.0	49.0
I-Adapt(ours)	**57.2**	68.4	40.5	**47.2**	51.3	**79.8**	59.5	**30.6**	52.0	59.1	27.4	25.1	43.4	81.7	68.9	49.7	24.2	39.6	**64.8**	60.8	**51.6**

Table 2: Results from PASCAL VOC to Comic2k. ResNet101 is used as the backbone for all models.

Method	bike	bird	car	cat	dog	prsn	mAP (%)
source-only	35.6	12.9	24.5	15.4	17.2	37.3	23.8
MCAR [43]	47.9	20.5	37.4	20.6	24.5	50.2	33.5
VIDSN [2]	51.0	23.9	48.3	36.0	29.2	57.4	41.0
MILA [22]	**59.1**	28.5	**49.8**	28.3	35.7	**66.3**	44.6
D_adapt [18]	52.4	25.4	42.3	43.7	25.7	53.5	40.5
I-Adapt(ours)	52.1	**29.9**	47.6	**38.2**	**40.1**	65.6	**45.6**
oracle	41.5	29.4	37.6	48.1	35.4	70.1	43.7

Table 3: Results from Sim10k to Cityscapes. VGG16 is used as the backbone for all models.

Method	AP on Car (%)
source-only	36.6
DAF [4]	39.0
MGA [44]	49.8
PT [3]	55.1
VSDSN [2]	53.5
CSDA [9]	56.9
D-adapt [18]	50.3
I-Adapt(ours)	**57.1**
oracle	72.8

VOC→Comic2k, which have large domain gaps. In these two tasks, our method surpasses D-adapt by 2.6% and 5.1%, and improves the highest performance by 1.7% and 1.0%, respectively, achieving new SOTA performances in both tasks. In task PASCAL VOC → Comic2k, we notice that the performance of I-adapt even surpasses that of "oracle." A possible reason is that I-adapt generates abundant high-quality pseudo labels and annotates more instances compared to the annotations in Comic2k.

Synthetic-to-Real Adaptation (Sim10k→Cityscapes). As shown in Table 3, we evaluate I-adapt by adapting the detector trained in a synthetic dataset Sim10k to the real-world dataset Cityscapes. The result shows that the proposed method outperforms D-adapt by a large margin of 6.8%, and obtains 0.2% improvements compared to the SOTA method CSDA [9].

Clear-to-Foggy Weather Adaptation (Cityscapes→Foggy Cityscapes). In Table 4, we evaluate the effectiveness of the proposed model on an adverse weather adaptation task, from Cityscapes to Foggy Cityscapes. The proposed I-adapt exceeds the state-of-the-art method CSDA [9] by 0.7% while achieving 5.2% improvement compared with D-adapt. Note that we select the most challenging foggy condition (*i.e.*, 0.02) from the Foggy Cityscapes as the target domain.

In addition to these methods, we also compared with a recently proposed method, Harmonious Teacher (HT) [7]. The results are reported in the supplemental material [41]. It shows that our method is comparable with HT and surpasses HT on tasks with large domain gaps (*e.g.*, PASCAL VOC→Clipart and PASCAL VOC→Comic2k).

5.4 Ablation Studies

In this part, we will discuss how the proposed methods contribute to the performance of the detector. Denote $\mathcal{ACC}$ be the prediction accuracy, and $|\Delta si|$ be the average of $\mathcal{L}_1$ distances between classification scores of true-positive foreground predictions and their IoU with

Table 4: Results from Cityscapes to Foggy Cityscapes. VGG16 is used as the backbone for all models. "*" means that CycleGAN [45] is used to perform source-to-target translation.

Method	prsn	rider	car	truck	bus	trn	mot	bike	mAP (%)
source-only	31.2	38.4	38.3	11.5	18.3	4.6	21.0	30.8	24.3
DAF [4]	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
SWDA [32]	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3	34.3
MCAR [43]	32.0	42.1	43.9	31.3	44.1	43.4	37.4	36.6	38.8
MGA [44]	45.7	47.5	60.6	31.0	52.9	44.5	29.0	38.0	43.6
PT [3]	40.2	48.8	59.7	30.7	51.8	30.6	35.4	44.5	42.7
VSDSN [2]	42.3	51.1	56.0	25.6	41.6	33.1	32.8	40.9	40.4
VSDSN* [2]	45.0	**55.2**	61.7	29.3	44.0	29.0	36.2	**46.9**	43.4
CSDA [9]	**46.6**	46.3	**63.1**	28.1	**56.3**	53.7	33.1	39.1	45.8
D-adapt [18]	43.1	51.8	58.1	26.3	36.8	14.6	32.2	42.0	38.1
D-adapt* [18]	44.9	54.2	61.7	25.6	36.3	24.7	37.3	46.1	41.3
I-Adapt(ours)	44.4	53.3	60.0	27.4	38.9	34.3	34.4	45.0	42.2
I-Adapt*(ours)	46.2	42.6	62.2	**40.9**	46.2	**55.6**	**47.3**	30.8	**46.5**
oracle	49.4	55.8	69.5	33.6	43.8	32.8	38.3	46.5	46.7

Table 5: Ablation study in Cityscapes→Foggy Cityscapes. CA, IA, BA, MI represent Category Adapter, IoU Adapter, Bbox Adapter, and Mutual Improvement, respectively. The "Baseline" method means only training the detector in the source domain. Both D-adapt and I-adapt use CycleGAN [45] to perform source-to-target translation.

Method	CA	IA	BA	MI	mAP(%)
Baseline	_	_	_	_	23.5
D-adapt [18]	✓		✓		41.3
modules		✓	✓		45.9
		✓	✓	✓	**46.5**

corresponding ground-truth bounding boxes. We measure the inconsistency between classification scores and the pseudo-label qualities with $|\Delta si|$. All experiments have been conducted on two tasks: PASCAL VOC→Clipart (large domain gap) and Cityscapes→Foggy Cityscapes (small domain gap). T (number of iterations in Algorithm 1) is set to 1 for a fair comparison.

Ablation study on I-adapt. We conducted ablation studies on four tasks. Due to the limited space, we only show the results in the task Cityscapes→Foggy Cityscapes, and the rest of the results can be found in the supplemental material [41]. As shown in Table 5, by adding IoU Adapter and Mutual Improvement step by step, the model performance is enhanced gradually. It demonstrates that : (1) the quality scores accurately reflect the quality of pseudo labels, and the corresponding weights encourage the model to focus on learning from high-quality ones; (2) MI makes category pseudo labels and quality scores more consistent with bounding box pseudo labels.

To further explain why the IA and MI modules are effective, we conducted two more experiments.

Effectiveness of the IoU Adapter. Table 6a shows the effectiveness of IA mentioned in Sec 4.1. When the detector is only trained on the source domain, the $\mathcal{ACC}$ of target detections $\mathcal{D}_t^{pred}$ is low, resulting in poor performance of the detector in the target domain. By applying the proposed IoU Adapter, the $\mathcal{ACC}$ in $\mathcal{D}_t^{pred}$ increase by 34.1% and 28.0% respectively. And $|\Delta si|$ also drops by 0.20 and 0.04. These results show that IA learns transferable and discriminative features, and makes the classification scores more consistent with pseudo-label qualities in both small and large domain gap tasks.

Effectiveness of the Mutual Improvement. Table 6b illustrates the

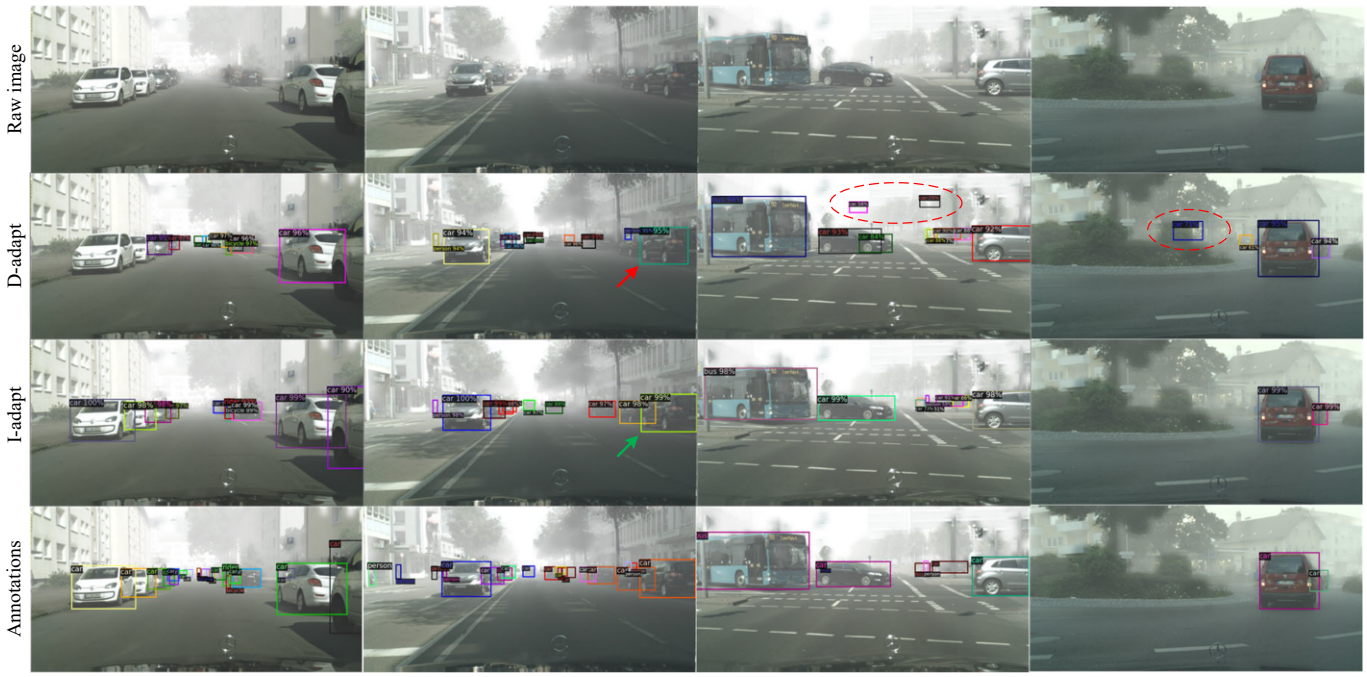

Figure 6: Detection results in the task Cityscapes→Foggy Cityscapes. We show the results of D-adapt [18] and I-adapt (ours) above.

Table 6: Ablation on IoU Adapter and Mutual Improvement. The bold font indicates better. V→CP and C→F represent task PASCAL VOC→Clipart, Cityscapes→Foggy Cityscapes, respectively.

(a) Ablation on IoU Adapter.

task	source only		IoU Adapter	
	$\mathcal{ACC}$ in $\mathcal{D}_t^{pred}$	$\lvert\Delta si\rvert$	$\mathcal{ACC}$ in $\mathcal{D}_t^{pred}$	$\lvert\Delta si\rvert$
V->CP	57.0%	0.32	**91.1% (↑34.1%)**	**0.12 (↓0.20)**
C->F	58.4%	0.14	**86.4% (↑28.0%)**	**0.10 (↓0.04)**

(b) Ablation on Mutual Improvement.

task	$\lvert\Delta si\rvert$			$\mathcal{ACC}$ in $\mathcal{D}_t^{fg}$	
	IA	IA+BA	IA+BA+MI	BA	BA+MI
V->CP	0.12	0.16	0.11	60.3%	**65.1% (↑4.8%)**
C->F	0.10	0.11	0.10	51.6%	**58.8% (↑7.1%)**

changes of $|\Delta si|$ and $\mathcal{ACC}$ after inputting foreground detections $\mathcal{D}_t^{fg}$ to different modules. Since D-adapt [18] has proved that BA can improve the localization performance of detection, we can infer that the quality of pseudo labels in $\mathcal{D}_t^{fg}$ will be better after applying BA to generate bounding box pseudo labels for each region $x_t^{det} \in \mathcal{D}_t^{fg}$, which means that $|\Delta si|$ will reduce. However, from columns 2 and 3, $|\Delta si|$ increases after "BA" is introduced. From this contradiction, we can infer that inconsistency happens between bounding box pseudo labels, category pseudo labels, and quality scores. By introducing Mutual Improvement (MI), $|\Delta si|$ drops by 0.05 and 0.01 in two tasks (columns 3 and 4), showing that MI improves the consistency between bounding box pseudo labels and quality scores. From the increment of $\mathcal{ACC}$ in $\mathcal{D}_t^{fg}$ (columns 5 and 6), the consistency between bounding box pseudo labels and category pseudo labels is also improved by MI. These results show that MI successfully improves the consistency between bounding box pseudo labels, category pseudo labels, and quality scores.

5.5 *Qualitative Evaluation*

Fig. 6 shows four detection results obtained by D-adapt and I-adapt in the task Citysapes→Foggy Cityscapes. In the first column, I-adapt can detect more true-positive objects than D-adapt, showing the effectiveness of I-adapt. In the second column, I-adapt predicts a more accurate bounding box (pointed by green arrow) than D-adapt (pointed by red arrow). Additionally, from the pictures in the third and fourth columns, the detection results of D-adapt contain false positive results (*e.g.* , three "car" detections in the red dashed circle), which means that the detector is misguided by the low-quality pseudo labels generated by D-adapt. Compared with D-adapt, the proposed I-adapt predicts the bounding box more accurately and detects fewer false-positive objects. From these observations, we can infer that I-adapt can guide the detector to focus on learning from high-quality pseudo labels while reducing the effect of low-quality ones.

6 Conclusion

In this work, we propose an effective adapter-based self-training framework named I-adapt. The proposed IoU Adapter can generate category pseudo labels and quality scores of pseudo labels in the target domain. Based on the quality scores, we propose a re-weighting strategy. It enables the detector to focus on learning from high-quality pseudo labels and reduces the effect of low-quality ones. Moreover, the proposed Mutual Improvement makes category pseudo labels and quality scores more consistent with bounding box pseudo labels. Our proposed I-adapt can be easily combined with the other detectors to adapt them to new domains that have different feature distributions. The results of various experiments show that the proposed I-adapt surpasses current state-of-the-art methods.

7 Acknowledgement

The work is supported by the National Natural Science Foundation of China (Grant Nos. 62225205, 92055213), the National Key Research and Development Program of China (2021ZD40303), the Science and Technology Program of Changsha (kh2301011), Shenzhen Basic Research Project (Natural Science Foundation) (JCYJ20210324140002006).

References

[1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020.

[2] J. Chen, W. Deng, B. Peng, T. Liu, Y. Wei, and L. Liu. Variational information bottleneck for cross domain object detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2231–2236. IEEE, 2023.

[3] M. Chen, W. Chen, S. Yang, J. Song, X. Wang, L. Zhang, Y. Yan, D. Qi, Y. Zhuang, D. Xie, et al. Learning domain adaptive object detection with probabilistic teacher. In *ICML*, pages 3040–3055, 2022.

[4] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[6] J. Deng, W. Li, Y. Chen, and L. Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, pages 4091–4101, 2021.

[7] J. Deng, D. Xu, W. Li, and L. Duan. Harmonious teacher for cross-domain object detection. In *CVPR*, pages 23829–23838, 2023.

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88: 303–338, 2010.

[9] C. Gao, C. Liu, Y. Dun, and X. Qian. Csda: Learning category-scale joint feature for domain adaptive object detection. In *ICCV*, pages 11421–11430, 2023.

[10] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang. Domain adaptation via prompt learning. *TNNLS*, 2023.

[11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[13] L. He, W. Wang, A. Chen, M. Sun, C.-H. Kuo, and S. Todorovic. Bidirectional alignment for domain adaptive detection with transformers. In *ICCV*, pages 18775–18785, 2023.

[14] Z. He and L. Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, pages 6668–6677, 2019.

[15] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *ECCV*, pages 733–748. Springer, 2020.

[16] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 749–757, 2020.

[17] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, pages 5001–5009, 2018.

[18] J. Jiang, B. Chen, J. Wang, and M. Long. Decoupled adaptation for cross-domain object detection. In *ICLR*, 2022.

[19] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang. New generation deep learning for video object detection: A survey. *TNNLS*, 33(8):3195–3215, 2021.

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

[21] S. Kim, J. Choi, T. Kim, and C. Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *ICCV*, pages 6092–6101, 2019.

[22] O. Krishna, H. Ohashi, and S. Sinha. Mila: memory-based instance-level adaptation for cross-domain object detection. In *BMVC*, 2023.

[23] Q. Lang, L. Zhang, W. Shi, W. Chen, and S. Pu. Exploring implicit domain-invariant features for domain adaptive object detection. *TCSVT*, 33(4):1816–1826, 2022.

[24] W. Li, X. Liu, and Y. Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, pages 5291–5300, 2022.

[25] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li. Large selective kernel network for remote sensing object detection. In *ICCV*, pages 16794–16805, 2023.

[26] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, and P. Vajda. Cross-domain adaptive teacher for object detection. In *CVPR*, pages 7581–7590, 2022.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[28] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018.

[29] F. Lyu, C. Chen, J. Zhang, X. Feng, and Z. Tang. Visualization for supercomputer system: A survey. *Journal of Computer-Aided Design & Computer Graphics*, 2024.

[30] P. Oza, V. A. Sindagi, V. V. Sharmini, and V. M. Patel. Unsupervised domain adaptation of object detectors: A survey. *TPAMI*, 2023.

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[32] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019.

[33] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In *CVPR*, pages 9627–9636, 2019.

[34] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024.

[35] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *CVPR workshops*, pages 129–137, 2017.

[36] X. Xiao, M. Duan, Y. Song, Z. Tang, and W. Yang. Fake node-based perception poisoning attacks against federated object detection learning in mobile computing networks. In *DAC*, 2024.

[37] S. Xu, H. Zhang, X. Xu, X. Hu, Y. Xu, L. Dai, K.-S. Choi, and P.-A. Heng. Representative feature alignment for adaptive object detection. *TCSVT*, 33(2):689–700, 2022.

[38] W. Yang, Y. Guo, J. Wu, Z. Wang, L.-Z. Guo, Y.-F. Li, and S. Liu. Interactive reweighting for mitigating label quality issues. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1837–1852, 2024.

[39] J. Yoo, I. Chung, and N. Kwak. Unsupervised domain adaptation for one-stage object detector using offsets to bounding box. In *ECCV*, pages 691–708. Springer, 2022.

[40] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *CVPR*, pages 8514–8523, 2021.

[41] Q. Zhang, C. Chen, Z. Liu, and Z. Tang. Supplemental material of "I-adapt: Using IoU Adapter to improve pseudo labels in cross-domain object detection". *Zenodo*, 2024. Available at: https://zenodo.org/doi/10.5281/zenodo.12753207.

[42] L. Zhao and L. Wang. Task-specific inconsistency alignment for domain adaptive object detection. In *CVPR*, pages 14217–14226, 2022.

[43] Z. Zhao, Y. Guo, H. Shen, and J. Ye. Adaptive object detection with dual multi-label prediction. In *ECCV*, pages 54–69. Springer, 2020.

[44] W. Zhou, D. Du, L. Zhang, T. Luo, and Y. Wu. Multi-granularity alignment domain adaptation for object detection. In *CVPR*, pages 9581–9590, 2022.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.